Capacity Analysis of Finite State Channels

CAPACITY ANALYSIS OF FINITE STATE CHANNELS

BY

RUI XU, M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2017) McMaster University
(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Capacity Analysis of Finite State Channels
AUTHOR:	Rui Xu M.Sc., (Electromagnetic Field & Microwave Engineering) Southeast University, Nanjing, China
SUPERVISOR:	Dr. Jun Chen and Dr. Jian-Kang Zhang
NUMBER OF PAGES:	xiv, 110

To all the people who helped me during the past four years.

Lay Abstract

It is well known that with the knowledge of channel state, it is possible to increase the channel capacity. In this sense, knowing channel state never hurts. However, whether it is always beneficial to actively acquire channel state is another story. If we take into account the cost of measuring the channel state against the potential gain on the capacity, sometimes it may not appear very economic to do so. This thesis studies the effect of the quality of observed channel states on the channel capacity. It has been found out in some circumstances the channel capacity is very sensitive to the noise on the state information. On the other hand, it appears that the maximum capacity can be achieved with the knowledge of a small portion of the total channel state information under a slightly different setting. This thesis proves the generality of such phenomena in binary-input channels and provides the necessary and sufficient conditions for the occurrence of such phenomena for an arbitrary channel. This paper also introduces the idea of intrinsic capacity which can be used to measure the ultimate capacity potential of a channel by exploring the channel state. By viewing an arbitrary channel as a deterministic channel with state, the greatest possible and smallest possible capacities have been either derived or bounded in the thesis.

Abstract

Channels with state model communication settings where the channel statistics are not fully known or vary over transmissions. It is important for communication system to obtain the channel state information in terms of increasing channel capacity. This thesis addresses the effect of the quality of state information on channel capacity. Extreme scenarios are studied to reveal the limit in increasing channel capacity with the knowledge of state information.

We consider the channel with the perfect state information at the decoder, while the encoder is only available to a noisy state observation. The effect of the noisy state at the encoder to the channel capacity is studied. We show that for any binary-input channel, if the mutual information between the noisy state observation at the encoder and the true channel state is below a positive threshold determined solely by the state distribution, then the capacity is the same as that with no encoder side information. A complementary phenomenon is also revealed for the generalized probing capacity. Extensions beyond binary-input channels are developed.

We further investigate the channel capacity, when the causal channel state information (available at the encoder or the decoder or both) makes it deterministic. Every such a capacity is called an intrinsic capacity of the channel. Among them, the smallest and the largest, called the lower and the upper intrinsic capacities, are particularly studied. Their exact values are determined in most cases when the input or the output is binary. General lower and upper bounds are also provided for the lower and the upper intrinsic capacities with causal state information available at both sides. Byproducts of this work are a generalization of the Birkhoff-von Neumann theorem and a result on the uselessness of causal state information at the encoder.

Acknowledgements

I would like to express the deepest gratitude to Dr. Jun Chen. He not only guided me and encouraged me as a caring supervisor, but also influenced me and inspired me as a dedicated researcher with his appreciation to the beauty of knowledge, enthusiasm to scientific exploration, and insight into his research field. I would also like to thank Dr. Chen for his kindness and consideration which will never be forgotten. It is truly my fortune to be his student.

Also, I would like to give my sincere gratitude to my co-supervisor Dr. Jian-Kang Zhang for providing me the opportunity of Ph.D. study, and for his kindness, advice and support during the past four years. I am also deeply indebted to Dr. Shengtian Yang who gave me plenty of valuable advices and helped me a lot on my research. I am also very grateful to my committee members Dr. Sorina Dumitrescu and Dr. Tim Davidson for their insightful comments and inspiring advices. In addition, I want to thank the staff in the ECE department and in particular Mrs. Cheryl Gies who gave me a lot of help. My sincere appreciation also goes to all my friends in McMaster University, who have made this period a great time in my life.

Finally, I am deeply grateful to my parents for their unconditional love and always being very supportive in my life. I would also like to express my appreciation to my amazing wife. This work would not happen without her love and support.

Notation and Abbreviations

CSI	Channel state information
CSIT	Channel state information available at the transmitter
CSIR	Channel state information available at the receiver
DMC	Discrete memoryless channel
e.g., X	Random variable
e.g., $\mathcal{X}$	Alphabet of X
e.g., $\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
e.g., p_X	Distribution of X, and also a $1 \times \lvert\mathcal{X}\rvert$ row vector
e.g., $p_{Y\lvert X}$	Conditional distribution of Y given X, and also a $\lvert\mathcal{X}\rvert \times \lvert\mathcal{Y}\rvert$ stochastic matrix
e.g., $p_{Y\lvert X}(\cdot\lvert x)$	Conditional distribution of Y given $X = x$, and also a $1 \times \lvert\mathcal{Y}\rvert$ row vector
e.g., $p_{Y\lvert X}(y\lvert\cdot)$	A column vector equal to the y-column of $p_{Y\lvert X}$
$C(\cdot)$	Channel capacity
$H(\cdot)$	Entropy
$I(\cdot;\cdot)$	Mutual information
$D(\cdot\lvert\lvert\cdot)$	Divergence
$\mathbb{I}(\cdot)$	Indicator function

BES	Binary erasure channel
BSS	Binary symmetric channel
$\mathbb{P}(\cdot)$	Probability
$\mathcal{X}\backslash\mathcal{Y}$	Exclude set $\mathcal{Y}$ from set $\mathcal{X}$
$*$	Erased output
$\Vert \cdot \Vert_\infty$	Infinity norm
$\Vert \cdot \Vert_2$	2-norm
$\Vert \cdot \Vert_F$	Frobenius norm
A$\Rightarrow$B	A is sufficient to B
A$\Leftrightarrow$B	A is equivalent to B
$\mathbb{R}$	Real space
$\mathcal{P}$	Channel space
$\hat{\mathcal{P}}$	Deterministic channel space
$\mathrm{dec}(\cdot)$	The set of all possible convex combination
$\mathrm{IC}(\cdot)$	Intrinsic capacity
$\underline{\mathrm{IC}}(\cdot)$	Lower intrinsic capacity
$\overline{\mathrm{IC}}(\cdot)$	Upper intrinsic capacity
e.g. $\mathrm{supp}(p_X)$	The support of p_X
e.g. $\mathbf{a}$	Row vector
e.g. $\mathbf{M}$	Matrix
$\mathbf{1}$	All-one row vector
e.g. $(a)^+$	The value equals the greater between a and 0
$\Gamma(\cdot,\cdot)$	Rank probability

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Channels with state refer to channels whose conditional output probability distribution depends on a state process, and where the channel state information (CSI) signal is available at the transmitter (CSIT) or at the receiver (CSIR) or at both ends. Many studies have devoted over the years to a wide range of scenarios. Depending on the assumptions on the channel state and on the availability and quality (clean or noisy) of the state information at the transmitter and/or the receiver, a variety of problems arise to the interest of related physical situations.

Note that the CSI signal can be observed either causally or noncausally. In the causal case, the transmitter and/or the receiver at time n know only the CSI sequence from time 1 to n, whereas in the non-causal case, the realization of the state sequence from the start to the end of transmission is known before the transmission of any symbol begins. The causal CSIT channel model was introduced in [Shannon 1958] where the state is generated by an independent and identically distributed (i.i.d.)

process and noiselessly available at the transmitter. Shannon found that the capacity of this channel is equal to the capacity of an ordinary discrete memoryless channel (DMC) with an extended input alphabet. [Salehi 1992] generalized Shannon's result to the case where both the transmitter and receiver observe (possibly different) noisy versions of the state information. Later, Caire and Shamai showed this result to be a special case of Shannon's model and also determined that optimal codes can be constructed directly on the input alphabet when CSIT is a deterministic function of CSIR (Caire and Shamai, 1999). The non-causal CSIT model was introduced in [Kuznetsov and Tsybakov 1974]. Gel'fand and Pinsker found the capacity and suggested the optimal coding for perfect CIST and no CSIR (Gel'fand and Pinsker, 1980). An extension of the result has been developed in [Cover and Chiang 2002] for the case where the transmitter and the receiver have the knowledge of different nonperfect CSI sequences, which are correlated to the state sequence. Regarding CSIR, causal and non-causal CSI are not distinguished, since the receiver waits until the end of the transmission anyway, before decoding. Channels with CSIR are studied in [Caire and Shamai 1999; Heegard and Gamal 1983; Salehi 1992].

Driven by the rapid development of wireless communications systems over the last decade, numerous works have been devoted to more realistic analytic models and to the exploration of the fundamental limits on reliable information transmission over these systems. The presence of CSI has been shown to yield significant performance gains, for example, in improving a predetermined space-time code (Jöngren *et al.*, 2002), in outage probability (Bhashyam *et al.*, 2002) or in capacity (Sabharwal *et al.*, 2000). However, in practice, CSI is communicated over way-side channels, for which only limited resources of the system are located, and moreover, the existence of noise or distortions should also be taken into account. Some recent works have been devoted to the more realistic case of partial or nonperfect CSI (Rosenzweig *et al.*, 2005; Asnani *et al.*, 2011; Song and Chen, 2011). Motivated by the tradeoff between the improvement in the channel capacity due to the availability of CSIT and the potential cost of maintaining a high quality of the CSIT, in this thesis, we study the effect of the quality of the CSIT on the channel capacity. On the other hand, knowing that the more effort we put to analyze a channel, the better knowledge of the channel statistics we will obtain, we want to discover the ultimate potential gain in capacity due to the CSI. To characterize the potential capacity gain, we introduce the intrinsic capacity which is the channel capacity as if the channel is fully known as deterministic channels with state.

1.2 Background

Consider the channel with state depicted in Fig. 1.1, whose input, state and output, at time t, are $X(t) \in \mathcal{X}$, $S(t) \in \mathcal{S}$ and $Y(t) \in \mathcal{Y}$, respectively, where $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$ are the corresponding finite alphabets. The transmitter wishes to communicate a message M over the channel with state to the receiver with possible state information $S(t)$ available at the encoder and/or the decoder or neither. Unless otherwise specified, we assume throughout a DMC with state model, where the channel is memoryless in the sense that

$$p(y^n|x^n, s^n, m) = \prod_{i=1}^{n} p_{Y|X,S}(y_i|x_i, s_i).$$

In this model, the channel state is randomly chosen by nature where the state

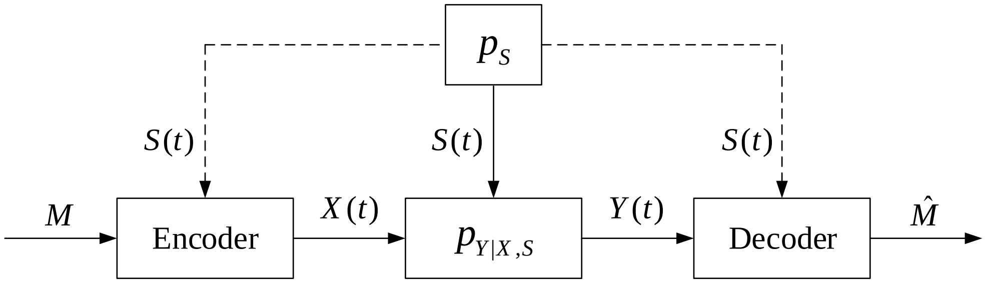

Figure 1.1: Channels with state model.

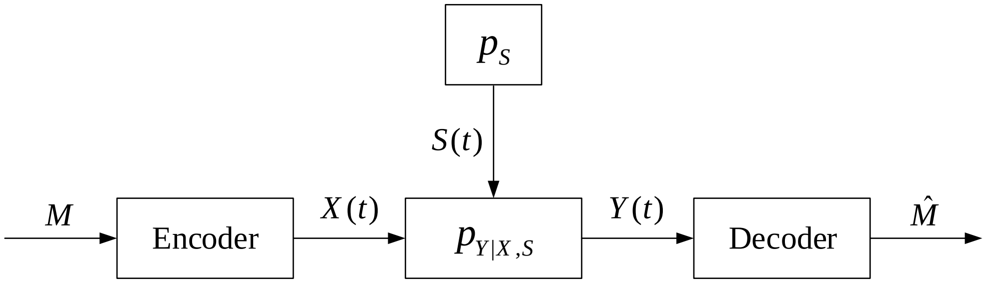

Figure 1.2: State information available at neither the encoder nor the decoder.

sequence $(S_1, S_2, \cdots)$ is i.i.d. with $S_i \sim p_s(s_i)$. The fact that the state changes over transmissions provides a temporal dimension to the availability of the channel state. The state information may be observed causally or non-causally. The capacity of this channel under various scenarios of state information availability is what we are interested in. When the state is available at the decoder, the capacity under the different temporal settings is the same. However, this is not the case when the state is available at the encoder. In the following subsections, we will introduce special cases of this general setup.

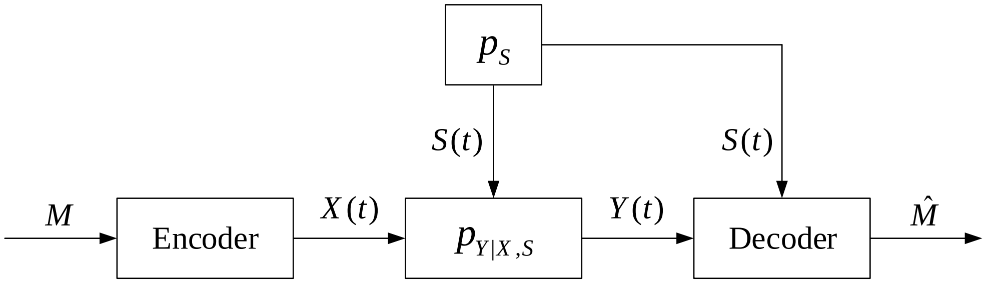

Figure 1.3: State information available only at the decoder.

1.2.1 State Information Available at Neither the Encoder nor the Decoder

When the state information is not available at either the encoder or the decoder, the model is shown in Fig. 1.2. Since neither the encoder nor the decoder have the knowledge of the state, the channel can be treated as a DMC without state by averaging the DMCs $p(y|x,s)$ over the state, i.e. let $p(y|x) = \sum_s p(s)p(y|x,s)$ be the DMC. Then it is easy to see that the capacity when the state information is not available at the encoder or the decoder is exactly the same as the capacity of an ordinary DMC, i.e.

$$C(p_{Y|X}) = \max_{p(x)} I(X;Y)$$

1.2.2 State Information Available only at the Decoder

When the state information is available only at the decoder, the model is shown in Fig. 1.3. In this model, the state information can be treated as part of the channel

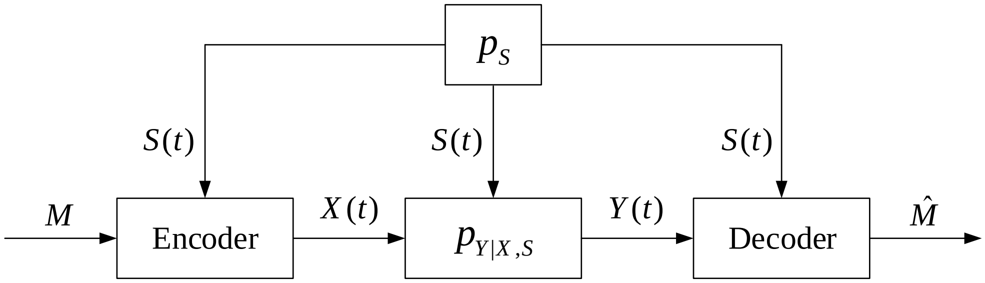

Figure 1.4: State information available at both the encoder and the decoder.

output. The channel capacity is

$$C_{\mathrm{D}}(p_{Y|X}) = \max_{p(x)} I(X;Y,S).$$

Note that the channel state is i.i.d. generated which means S is independent with both X and Y. Therefore, the capacity can also be written as

$$C_{\mathrm{D}}(p_{Y|X}) = \max_{p(x)} I(X;Y|S).$$

1.2.3 State Information Available at Both the Encoder and the Decoder

When the state information is available causally and/or non-causally at both the encoder and the decoder, the model is depicted in Fig. 1.4. In this model, the channel state S^n can be treated as a time-sharing sequence. Then channel capacity is the same for all four combinations and is given by

$$C_{\mathrm{ED}}(p_{Y|X}) = \max_{p(x|s)} I(X;Y|S).$$

1.2.4 State Information Available only at the Encoder

When the state information is available only at the encoder as depicted in Fig. 1.5, the capacity distinguishes between channels where the state information is observed causally and channels where the state information is observed non-causally.

For the causal case, the encoder knows only S^i before transmission i. In this scenario, an auxiliary random variable, which is independent of S, is introduced as U with $|\mathcal{U}| \leq |\mathcal{X}|^{|\mathcal{S}|}$. Then, the capacity of the DMC with DM state $p(y|x,s)p(s)$ when the state information is available causally only at the encoder is

$$C_{\mathrm{E}}(p_{Y|X}) = \max_{p(u),x(u,s)} I(U;Y)$$

The coding scheme corresponds to attaching a "mapping device" $x(u,s)$ with two inputs U and S and one output X in front of the actual channel input. In this way, a new DMC is induced as $p(y|u) = \sum_s p(y|x(u,s),s)p(s)$ with input U, output Y and capacity C_{E}. Note that we can view the encoding as being performed over the set of all functions $\{x_u(s) : \mathcal{S} \to \mathcal{X}\}$ indexed by u as the input alphabet. This technique of coding over functions onto $\mathcal{X}$ instead of actual symbols in $\mathcal{X}$ is referred to as the *Shannon strategy*. It can reduce the cardinality bound of $\mathcal{U}$ to $\min\{(|\mathcal{X}|-1)|\mathcal{S}|+1, |\mathcal{Y}|\}$.

When the state information is available noncausally only at the encoder, Gelfand-Pinsker Theorem gives the channel capacity as follows.

$$C_{\mathrm{GP}}(p_{Y|X}) = \max_{p(u|s),x(u,s)} (I(U;Y) - I(U;S)),$$

where $\mathcal{U} \leq \min\{|\mathcal{X}| \cdot |\mathcal{S}|, |\mathcal{Y}| + |\mathcal{S}| - 1\}$.

Recall that when the state information is causally available at the encoder, U is

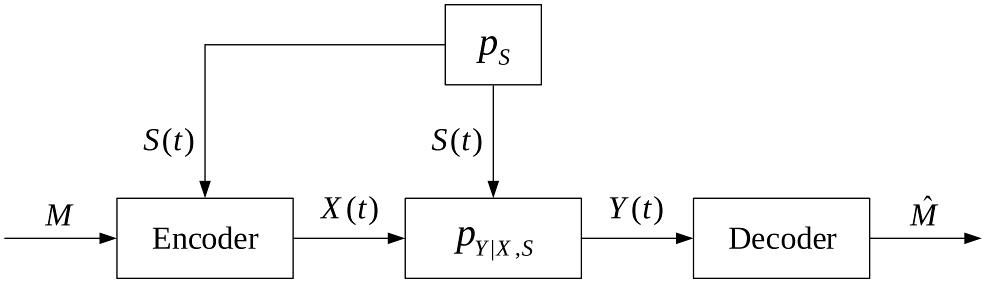

Figure 1.5: State information available only at the encoder.

independent of S which leads to $I(U;S) = 0$. The capacity of the causal case can also be written as

$$C_{\mathrm{E}}(p_{Y|X}) = \max_{p(u),x(u,s)} (I(U;Y) - I(U;S)).$$

As we can see, these two expressions of causal and non-causal cases have the same form, except that in the causal case the maximization is over $p(u)$ instead of $p(u|s)$.

1.3 Thesis Contributions and Outline

This thesis focuses on the theoretical analysis of the capacity of channels with state. In Chapter 2, we study the channel model when the perfect state information is available at the decoder and the encoder is only accessible to a noisy version of the state information. We find that for a binary-input channel, when the quality of the state information at the encoder is below a certain threshold, the channel capacity is as low as if there was no state information at the encoder at all. On the other hand, a generalized probing capacity is as high as if there was perfect state information at the encoder when the quality of the state information is above a certain threshold at the encoder end. We claim that this surprising phenomena can in fact be observed for all binary-input channels. The main results of this chapter are summarized in Theorems 2.1 and 2.2 in which the thresholds of the phenomena have also been given. The rest of Chapter 2 is organized as follows. We present the proofs of Theorems 2.1 and 2.2 in Sections 2.2 and 2.3, respectively. The validity of these two results under various modified conditions is discussed in Section 2.4. Section 2.5 contains some concluding remarks.

In Chapter 3, we introduce the idea of intrinsic capacity. Based on the idea that any channel can be seen as deterministic channels with state which we call intrinsic state, the intrinsic capacity is the channel capacity after all the uncertainty of the channel being eliminated. We show that each intrinsic state distribution corresponds to some ordinary channel, and the intrinsic capacity solely depends on the distribution of the intrinsic state. However, the mapping from an ordinary channel to its intrinsic state is not 1-to-1 or unique. We suspect that the true mapping should depend on the physical nature or many other factors of the channel. Despite of that, it is still possible to determine the lower and upper bound of the intrinsic capacity. On the other hand, according to the availability of the intrinsic state at the encoder and/or the decoder, the analysis of the intrinsic capacity should also be scenario-specific which could complicate the problem. The main contributions of Chapter 3 are as follows. 1) We study the structure of the convex polytope consisting of all convex combinations of deterministic channels for a generic channel; 2) We prove a generalization of the Birkhoff-von Neumann theorem for a family of channel matrices with integer-valued column-sum vector constraints from below and above, respectively; 3) When the intrinsic state is available at both the encoder and the decoder, the lower and upper intrinsic capacity is determined for binary-input or binary-output channels, and general upper and lower bounds are also provided for the non-binary cases; 4) For a binary-output channel, the lower and upper intrinsic capacity are determined when the intrinsic state is only available at the encoder; 5) For a binary-input channel, the lower and upper intrinsic capacity are determined when the intrinsic state is available at the decoder only. Chapter 3 is organized as follows. Section 3.2 formulates the problem of intrinsic capacities. The simplest case, the binary-input binary-output channel, is first studied in Section 3.3. The main results of this chapter are then presented in Section 3.4.

The rest of this thesis is organized as follows. In Chapter 2, we first state our main results Theorems 2.1 and 2.2 in Section 2.1; we then present the proofs of those two theorems in Sections 2.2 and 2.3, respectively; the validity of these two results under various modified conditions is discussed in Section 2.4; section 2.5 contains some concluding remarks. In Chapter 3, we first introduce the idea of intrinsic capacity in Section 3.1; section 3.2 formulates the problem of intrinsic capacities; The simplest case, the binary-input binary-output channel, is first studied in Section 3.3; the main results of this chapter are then presented in Section 3.4. Finaly, Chapter 4 concludes the thesis.

Chapter 2

When is Noisy State Information at the Encoder as Useless as No Information or as Good as Noise-Free State?

2.1 Introduction

Consider a memoryless channel $p_{Y|X,S}$ with input X, output Y, and state S. We assume that the channel state S, distributed according to p_S, is provided to the decoder, and a noisy state observation $\tilde{S}$, generated by S through side channel $p_{\tilde{S}|S}$, is available causally at the encoder. Here X, Y, S, and $\tilde{S}$ are defined over finite alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{S}$, and $\tilde{\mathcal{S}}$, respectively. In this setting (see Fig. 2.1), Shannon's remarkable result (Shannon, 1958) (see also (Caire and Shamai, 1999, Eq. (3)) and

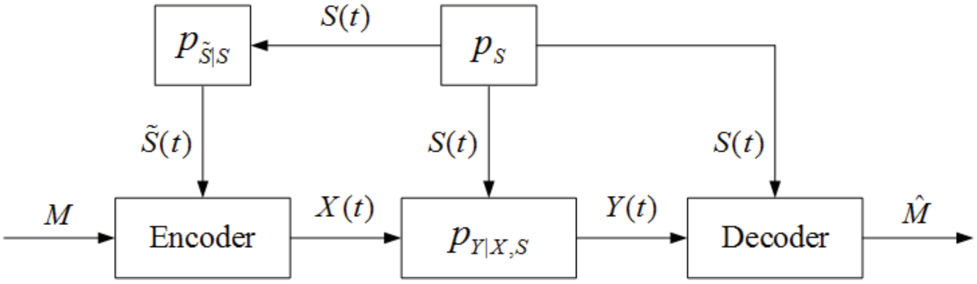

Figure 2.1: Channel model.

(Gamal and Kim, 2011, Th. 7.2)) implies that the channel capacity is given by

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \triangleq \max_{p_U} I(U; Y|S). \tag{2.1}$$

The auxiliary random variable U is defined over alphabet $\mathcal{U}$ with $|\mathcal{U}| = |\mathcal{X}|^{|\tilde{S}|}$, whose joint distribution with $(X, Y, S, \tilde{S})$ factors as

$$\begin{aligned} p_{U,X,Y,S,\tilde{S}}(u, x, y, s, \tilde{s}) = p_U(u) p_S(s) p_{\tilde{S}|S}(\tilde{s}|s) \mathbb{I}(x = \psi(u, \tilde{s})) p_{Y|X,S}(y|x, s), \\ u \in \mathcal{U}, x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}, \tilde{s} \in \tilde{S}, \end{aligned} \tag{2.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\psi(u, \cdot)$, $u \in \mathcal{U}$, are $|\mathcal{X}|^{|\tilde{S}|}$ different mappings from $\tilde{\mathcal{S}}$ to $\mathcal{X}$. Without loss of generality, we set $\mathcal{X} = \{0, 1, \cdots, |\mathcal{X}| - 1\}$, $\mathcal{S} = \{0, 1 \cdots, |\mathcal{S}| - 1\}$, $\mathcal{U} = \{0, 1, \cdots, |\mathcal{X}|^{|\tilde{S}|} - 1\}$, and order the mappings $\psi(u, \cdot)$, $u \in \mathcal{U}$, in such a way that the first $|\mathcal{X}|$ mappings[1] are

$$\psi(u, \cdot) \equiv u, \quad u \in \mathcal{X}; \tag{2.3}$$

[1] These are the mappings that ignore the encoder side information.

moreover, we assume that $\rho \triangleq \min_{s\in\mathcal{S}} p_S(s) > 0$. The capacity formula (2.1) can be simplified in the following two special cases. Specifically, when there is no encoder side information, the channel capacity reduces to (Gamal and Kim, 2011, Eq. (7.2))

$$\underline{C}(p_{Y|X,S}, p_S) \triangleq \max_{p_X} I(X;Y|S), \tag{2.4}$$

where $p_{X,Y,S}(x,y,s) = p_X(x)p_S(s)p_{Y|X,S}(y|x,s)$; on the other hand, when perfect state information is available at the encoder (as well as the decoder), the channel capacity becomes (Gamal and Kim, 2011, Eq. (7.3))

$$\overline{C}(p_{Y|X,S}, p_S) \triangleq \max_{p_{X|S}} I(X;Y|S), \tag{2.5}$$

where $p_{X,Y,S}(x,y,s) = p_S(s)p_{X|S}(x|s)p_{Y|X,S}(y|x,s)$.

For comparison, consider the following similarly defined quantity

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \triangleq \max_{p_U} I(X;Y|S),$$

where the joint distribution of $(U,X,Y,S,\tilde{S})$ is also given by (2.2). We shall refer to $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$ as the generalized probing capacity. By the functional representation lemma (Gamal and Kim, 2011, p. 626) (see also (Wang *et al.*, 2011, Lemma 1)), $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$ can be defined equivalently as

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \triangleq \max_{p_{X|\tilde{S}}} I(X;Y|S),$$

where

$$p_{X,Y,S,\tilde{S}}(x,y,s,\tilde{s}) = p_S(s)p_{\tilde{S}|S}(\tilde{s}|s)p_{X|\tilde{S}}(x|\tilde{s})p_{Y|X,S}(y|x,s), \quad x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}, \tilde{s} \in \tilde{\mathcal{S}}.$$

Clearly,

$$\begin{aligned} \underline{C}(p_{Y|X,S}, p_S) &\leq C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \\ &\leq C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \\ &\leq \overline{C}(p_{Y|X,S}, p_S). \end{aligned} \tag{2.6}$$

Moreover, we have

$$\begin{aligned} C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) &= C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \\ &= \underline{C}(p_{Y|X,S}, p_S) \end{aligned} \tag{2.7}$$

if S and $\tilde{S}$ are independent (i.e., $I(S;\tilde{S}) = 0$), and

$$\begin{aligned} C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) &= C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \\ &= \overline{C}(p_{Y|X,S}, p_S) \end{aligned} \tag{2.8}$$

if S is a deterministic function of $\tilde{S}$ (i.e., $H(S|\tilde{S}) = 0$).

To elucidate the operational meaning of $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$ and its connection with $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$, it is instructive to consider the special case where $p_{\tilde{S}|S}$ is a binary erasure channel with erasure probability ϵ (denoted by BEC(ϵ)), which corresponds to the probing channel setup studied in [Asnani *et al.* 2011]. The probing channel model

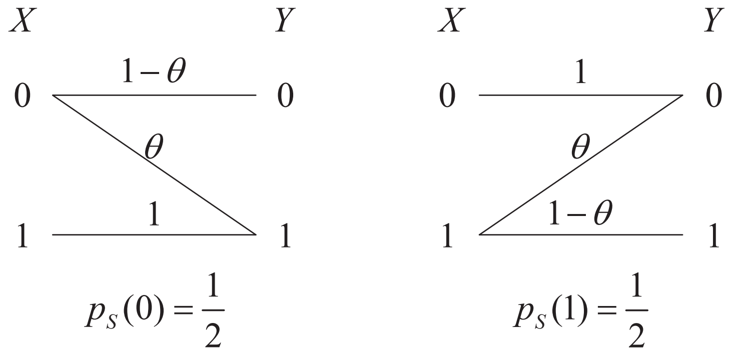

Figure 2.2: Illustration of $p_{Y|X,S}$ and p_S given by (2.9) and (2.10), respectively.

is essentially the same as the one in Fig. 2.1 except that, in Fig. 2.1, the encoder (which, with high probability, observes approximately $n\epsilon$ state symbols out of the whole state sequence of length n when n is large enough) has no control of the exact positions of these $n(1-\epsilon)$ symbols whereas, in the probing channel model, the encoder has the freedom to specify the positions of these $n(1-\epsilon)$ symbols according to the message to be sent. It is shown by [Asnani *et al.* 2011] that this additional freedom increases the achievable rate from $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ to $C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$. Now consider an example (see also Fig. 2.2) where

$$p_{Y|X,S}(y|x,s) = \begin{cases} 1-\theta, & (x,y,s)=(0,0,0) \text{ or } (1,1,1), \\ \theta, & (x,y,s)=(0,1,0) \text{ or } (1,0,1), \\ 0, & (x,y,s)=(1,0,0) \text{ or } (0,1,1), \\ 1, & (x,y,s)=(1,1,0) \text{ or } (0,0,1), \end{cases} \tag{2.9}$$

$$p_S(0) = p_S(1) = \frac{1}{2}. \tag{2.10}$$

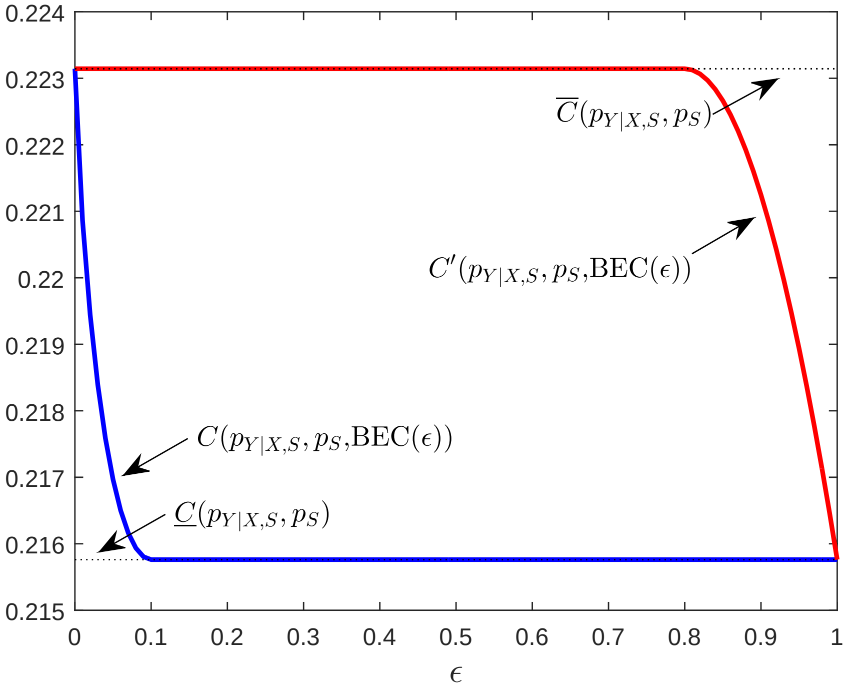

Figure 2.3: Plots of $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ and $C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0, 1]$, where $p_{Y|X,S}$ and p_S are given by (2.9) with $\theta = \frac{1}{2}$ and (2.10), respectively.

For this example, it can be verified that

$$\underline{C}(p_{Y|X,S}, p_S) = \begin{cases} \log 2, & \theta = 0, \\ \frac{1}{2}\left((1-\theta)\log 2 + \log\frac{2}{1+\theta} + \theta\log\frac{2\theta}{1+\theta}\right), & \theta \in (0,1), \\ 0, & \theta = 1, \end{cases}$$

$$\overline{C}(p_{Y|X,S}, p_S) = \begin{cases} \log 2, & \theta = 0, \\ \log\left(1 + (1-\theta)\theta^{\frac{\theta}{1-\theta}}\right), & \theta \in (0,1), \\ 0, & \theta = 1. \end{cases}$$

Note that $\overline{C}(p_{Y|X,S}, p_S)$ is strictly greater than $\underline{C}(p_{Y|X,S}, p_S)$ unless $\theta = 0$ or $\theta = 1$.

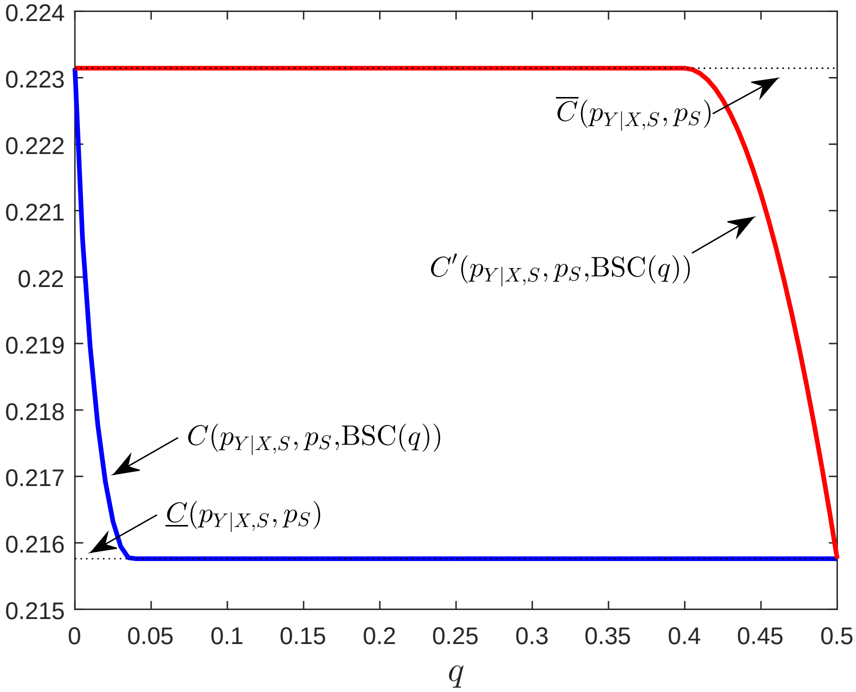

Figure 2.4: Plots of $C(p_{Y|X,S}, p_S, \mathrm{BSC}(q))$ and $C'(p_{Y|X,S}, p_S, \mathrm{BSC}(q))$ against q for $q \in [0, \frac{1}{2}]$, where $p_{Y|X,S}$ and p_S are given by (2.9) with $\theta = \frac{1}{2}$ and (2.10), respectively.

It follows by (2.7) and (2.8) that

$$\begin{aligned} C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))\big|_{\epsilon=1} &= C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))\big|_{\epsilon=1} \\ &= \underline{C}(p_{Y|X,S}, p_S), \\ C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))\big|_{\epsilon=0} &= C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))\big|_{\epsilon=0} \\ &= \overline{C}(p_{Y|X,S}, p_S). \end{aligned}$$

To gain a better understanding, we plot $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ and $C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0, 1]$ in Fig. 2.3. It turns out that, somewhat counterintuitively, $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ coincides with $\underline{C}(p_{Y|X,S}, p_S)$ way before ϵ reaches 1. That is to say, when ϵ is above a certain threshold strictly less than 1, the noisy state observation $\tilde{S}$ is useless and can be ignored (as far as the channel capacity is concerned). On the other hand, it can be seen that $C'(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ is equal to $\overline{C}(p_{Y|X,S}, p_S)$ for a large range of ϵ strictly greater than 0. Hence, in terms of the probing capacity, the noisy state observation can be as good as the perfect one. As shown in Fig. 2.4, the same phenomena arise if we choose $p_{\tilde{S}|S}$ to be a binary symmetric channel with crossover probability q (denoted by BSC(q)).

The contributions of the present work are summarized in the following theorems, which indicate that the aforedescribed surprising phenomena can in fact be observed for all binary-input channels.

Theorem 2.1 *For any binary-input channel $p_{Y|X,S}$, state distribution p_S, and side channel $p_{\tilde{S}|S}$,*

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \underline{C}(p_{Y|X,S}, p_S)$$

if $I(S;\tilde{S}) \leq \frac{\rho^2}{2e^2}$, where $\rho \triangleq \min_{s\in\mathcal{S}} p_S(s)$.

Theorem 2.2 *For any binary-input channel $p_{Y|X,S}$, state distribution p_S, and side channel $p_{\tilde{S}|S}$,*

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \overline{C}(p_{Y|X,S}, p_S)$$

if $H(S|\tilde{S}) \leq \frac{2\rho\log 2}{(|\mathcal{S}|-1)(e-1)}$, where $\rho \triangleq \min_{s\in\mathcal{S}} p_S(s)$.

On the surface these two results may look rather similar. One might even suspect the existence of a certain duality between them. However, it will be seen that the underlying reasons are actually quite different. The proof of Theorem 2.1 hinges upon, among other things, a perturbation analysis. In contrast, Theorem 2.2 is essentially a manifestation of an induced Markov structure.

The conditions in Theorem 2.1 and Theorem 2.2 are stated in terms of bounds on $I(S;\tilde{S})$ and $H(S|\tilde{S})$; as a consequence, they depend inevitably on p_S. As shown by Theorem 2.3 in Section 2.2 and Theorem 2.4 in Section 2.3, it is in fact possible to establish these two results under more general conditions on $p_{\tilde{S}|S}$ that are universal for all binary-input channels and state distributions.

Throughout this chapter, all logarithms are base-e.

2.2 Proof of Theorem 2.1

First consider the special case where $p_{\tilde{S}|S}$ is a generalized erasure channel (with erasure probability $\epsilon \in [0,1]$) defined as

$$p_{\tilde{S}^{(\epsilon)}_{\mathrm{GE}}|S}(\tilde{s}|s) = \begin{cases} 1-\epsilon, & \tilde{s} = s, \\ \epsilon, & \tilde{s} = *, \\ 0, & \text{otherwise}, \end{cases} \qquad s \in \mathcal{S}, \tilde{s} \in \mathcal{S} \cup \{*\}.$$

Lemma 2.1 *Given any binary-input channel $p_{Y|X,S}$ and state distribution p_S,*

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}^{(\epsilon)}_{\mathrm{GE}}|S}) = \underline{C}(p_{Y|X,S}, p_S)$$

for $\epsilon \in [1-e^{-1}, 1]$.

Remark: Lemma 2.1 provides a universal upper bound[2] on the erasure probability threshold above which the encoder side information is useless. The actual threshold, however, depends on $p_{Y|X,S}$ and p_S (see Section 2.4.1 for a detailed analysis).

Proof: As indicated by (2.1), the capacity of the channel model in Fig. 2.1 (i.e., $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$) is equal to that of channel $p_{Y,S|U}$, where

$$p_{Y,S|U}(y, s|u) = \sum_{\tilde{s}\in\tilde{\mathcal{S}}} p_S(s)p_{\tilde{S}|S}(\tilde{s}|s)p_{Y|X,S}(y|\psi(u,\tilde{s}), s), \quad u \in \mathcal{U}, y \in \mathcal{Y}, s \in \mathcal{S}.$$

According to [Gallager 1968, Th. 4.5.1], p_U is a capacity-achieving input distribution of channel $p_{Y,S|U}$ (i.e., p_U is a maximizer of the optimization problem in (2.1)) if and only if there exists some number C such that

$$D(p_{Y,S|U}(\cdot,\cdot|u)\|p_{Y,S}) = C, \quad u \in \mathcal{U} \text{ with } p_U(u) > 0,$$
$$D(p_{Y,S|U}(\cdot,\cdot|u)\|p_{Y,S}) \leq C, \quad u \in \mathcal{U} \text{ with } p_U(u) = 0;$$

furthermore, the number C is equal to $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$. In view of (2.3), we have

$$p_{Y,S|U}(y, s|u) = p_{Y,S|X}(y, s|u), \quad u \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}.$$

Let $p_{\hat{X}}$ be a capacity-achieving input distribution of channel $p_{Y,S|X}$ (i.e, $p_{\hat{X}}$ is a maximizer of the optimization problem in (2.4)). Define

$$p_{\hat{U}}(u) = \begin{cases} p_{\hat{X}}(u), & u \in \mathcal{X}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

[2] Numerical simulations suggest that this universal upper bound is not tight. Determining the exact universal erasure probability threshold remains an open problem.

It is clear that $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \underline{C}(p_{Y|X,S}, p_S)$ if and only if $p_{\hat{U}}$ is a capacity-achieving input distribution of channel $p_{Y,S|U}$.

Now consider the special case where $p_{\tilde{S}|S}$ is a generalized erasure channel with erasure probability ϵ, and define

$$D_{\mathrm{GE}}(p_U, \epsilon, u) = D(p_{Y,S|U}(\cdot, \cdot|u)\|p_{Y,S}) \tag{2.12}$$

to stress the dependence of $D(p_{Y,S|U}(\cdot, \cdot|u)\|p_{Y,S})$ on p_U, ϵ, and u. It can be verified that

$$\begin{aligned}
&p_{Y,S|U}(y, s|u) \\
&= \sum_{\tilde{s} \in \mathcal{S} \cup \{*\}} p_S(s) p_{\tilde{S}^{(\epsilon)}|S}(\tilde{s}|s) p_{Y|X,S}(y|\psi(u, \tilde{s}), s) \\
&= p_S(s)\epsilon p_{Y|X,S}(y|\psi(u, *), s) + p_S(s)(1-\epsilon) p_{Y|X,S}(y|\psi(u, s), s) \\
&= p_S(s)(p_{Y|X,S}(y|\psi(u, s), s) + \epsilon\delta(u, y, s)),
\end{aligned} \tag{2.13}$$

where

$$\delta(u, y, s) = p_{Y|X,S}(y|\psi(u, *), s) - p_{Y|X,S}(y|\psi(u, s), s), \quad u \in \mathcal{U}, y \in \mathcal{Y}, s \in \mathcal{S}. \tag{2.14}$$

Since $|\mathcal{X}| = 2$, there is no loss of generality in assuming that (Shulman and Feder, 2004, Th. 2)

$$p_{\hat{X}}(x) > e^{-1}, \quad x \in \mathcal{X}. \tag{2.15}$$

To the end of proving Lemma 2.1, it suffices to show that, for $\epsilon \in [1-e^{-1}, 1]$,

$$\begin{aligned} D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) &= \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{X}, \\ D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) &\leq \underline{C}(p_{Y|X,S}, p_S), \quad \text{otherwise.} \end{aligned}$$

Clearly, $p_{\hat{U}}$ is a capacity-achieving input distribution of channel $p_{Y,S|U}$ when $\epsilon = 1$. Therefore, we have[3]

$$\begin{aligned} D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) &= \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{X}, && (2.16) \\ D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) &\leq \underline{C}(p_{Y|X,S}, p_S), \quad \text{otherwise.} && (2.17) \end{aligned}$$

Note that

$$\begin{aligned} & D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) \\ &= \sum_{y\in\mathcal{Y}, s\in\mathcal{S}} p_{Y,S|U}(y,s|u) \log \frac{p_{Y,S|U}(y,s|u)}{\sum_{u'\in\mathcal{U}} p_{\hat{U}}(u') p_{Y,S|U}(y,s|u')} \\ &= \sum_{y\in\mathcal{Y}, s\in\mathcal{S}} p_S(s)(p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)) \\ &\quad \times \log \frac{p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)}{\sum_{u'\in\mathcal{U}} p_{\hat{U}}(u')(p_{Y|X,S}(y|\psi(u',s),s) + \epsilon\delta(u',y,s))} && (2.18) \\ &= \sum_{y\in\mathcal{Y}, s\in\mathcal{S}} p_S(s)(p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)) \\ &\qquad \times \log \frac{p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)}{\sum_{x\in\mathcal{X}} p_{\hat{X}}(x) p_{Y|X,S}(y|x,s)}, \quad \epsilon \in [0,1], u \in \mathcal{U}, && (2.19) \end{aligned}$$

[3] The inequality in (2.17) is in fact an equality.

where (2.18) is due to (2.13), and (2.19) is due to (2.3) and (2.11). Moreover,

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon}D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}p_S(s)\delta(u,y,s)\times\left(\log\frac{p_{Y|X,S}(y|\psi(u,s),s)+\epsilon\delta(u,y,s)}{\sum_{x\in\mathcal{X}}p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)}+1\right)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}p_S(s)\delta(u,y,s)\times\log\frac{p_{Y|X,S}(y|\psi(u,s),s)+\epsilon\delta(u,y,s)}{\sum_{x\in\mathcal{X}}p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)}\\
&\quad+\sum_{s\in\mathcal{S}}p_S(s)\sum_{y\in\mathcal{Y}}\delta(u,y,s)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}p_S(s)\delta(u,y,s)\times\log\frac{p_{Y|X,S}(y|\psi(u,s),s)+\epsilon\delta(u,y,s)}{\sum_{x\in\mathcal{X}}p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\epsilon\in[0,1],u\in\mathcal{U}.
\end{aligned}\tag{2.20}$$

Define

$$\mathcal{G}_\delta=\{u\in\mathcal{U}:\delta(u,y,s)=0\text{ for all }y\in\mathcal{Y}\text{ and }s\in\mathcal{S}\}.\tag{2.21}$$

In light of (2.19),

$$D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u)=D_{GE}(p_{\hat{U}},1,u),\quad\epsilon\in[0,1],u\in\mathcal{G}_\delta.\tag{2.22}$$

For any $u\in\mathcal{U}\backslash\mathcal{G}_\delta$, there must exist some $y\in\mathcal{Y}$ and $s\in\mathcal{S}$ such that $\delta(u,y,s)\neq 0$; furthermore, since $|\mathcal{X}|=2$, we have

$$\begin{aligned}
\delta(u,y,s)>0\Longrightarrow p_{Y|X,S}&(y|\psi(u,s),s)+\epsilon\delta(u,y,s)\\
&=b(y,s)+\epsilon(a(y,s)-b(y,s)),
\end{aligned}\tag{2.23}$$

$$\delta(u,y,s)<0\Longrightarrow p_{Y|X,S}(y|\psi(u,s),s)+\epsilon\delta(u,y,s)$$

$$= a(y,s) + \epsilon(b(y,s) - a(y,s)), \tag{2.24}$$

where

$$\begin{aligned} a(y,s) &= \max_{x\in\mathcal{X}} p_{Y|X,S}(y|x,s), \\ b(y,s) &= \min_{x\in\mathcal{X}} p_{Y|X,S}(y|x,s). \end{aligned}$$

Continuing from (2.20),

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon} D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) \\
&= \sum_{y\in\mathcal{Y}, s\in\mathcal{S}} p_S(s)\delta(u,y,s) \times \log \frac{p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)}{\sum_{x\in\mathcal{X}} p_{\hat{X}}(x) p_{Y|X,S}(y|x,s)} \\
&\geq \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)>0} \delta(u,y,s) \times \log \frac{p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)}{(1-e^{-1})a(y,s) + e^{-1}b(y,s)} \\
&\quad + \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)<0} \delta(u,y,s) \times \log \frac{p_{Y|X,S}(y|\psi(u,s),s) + \epsilon\delta(u,y,s)}{e^{-1}a(y,s) + (1-e^{-1})b(y,s)} \qquad (2.25) \\
&= \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)>0} \delta(u,y,s) \times \log \frac{b(y,s) + \epsilon(a(y,s) - b(y,s))}{(1-e^{-1})a(y,s) + e^{-1}b(y,s)} \\
&\quad + \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)<0} \delta(u,y,s) \times \log \frac{a(y,s) + \epsilon(b(y,s) - a(y,s))}{e^{-1}a(y,s) + (1-e^{-1})b(y,s)} \qquad (2.26) \\
&\geq \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)>0} \delta(u,y,s) \times \log \frac{(1-e^{-1})a(y,s) + e^{-1}b(y,s)}{(1-e^{-1})a(y,s) + e^{-1}b(y,s)} \\
&\quad + \sum_{s\in\mathcal{S}} p_S(s) \sum_{y\in\mathcal{Y}:\delta(u,y,s)<0} \delta(u,y,s) \times \log \frac{e^{-1}a(y,s) + (1-e^{-1})b(y,s)}{e^{-1}a(y,s) + (1-e^{-1})b(y,s)} \\
&= 0, \quad \epsilon \in [1-e^{-1}, 1], u \in \mathcal{U}, \qquad (2.27)
\end{aligned}$$

where (2.25) is due to (2.15), and (2.26) is due to (2.23) and (2.24). Combining (2.16), (2.17), (2.22), (2.27), and the fact $\mathcal{X} \subseteq \mathcal{G}_\delta$ yields the desired result. □

Recall (Gamal and Kim, 2011, p. 112) that $p_{\tilde{S}_1|S}$ (with input alphabet $\mathcal{S}$ and output alphabet $\tilde{\mathcal{S}}_1$) is said to be a stochastically degraded version of $p_{\tilde{S}_2|S}$ (with input alphabet $\mathcal{S}$ and output alphabet $\tilde{\mathcal{S}}_2$) if there exists $p_{\tilde{S}_1|\tilde{S}_2}$ satisfying

$$p_{\tilde{S}_1|S}(\tilde{s}_1|s) = \sum_{\tilde{s}_2 \in \tilde{\mathcal{S}}_2} p_{\tilde{S}_2|S}(\tilde{s}_2|s) p_{\tilde{S}_1|\tilde{S}_2}(\tilde{s}_1|\tilde{s}_2), \quad s \in \mathcal{S}, \tilde{s}_1 \in \tilde{\mathcal{S}}_1. \tag{2.28}$$

We can write (2.28) equivalently as

$$p_{\tilde{S}_1|S} = p_{\tilde{S}_2|S} p_{\tilde{S}_1|\tilde{S}_2}$$

by viewing $p_{\tilde{S}_1|S}$, $p_{\tilde{S}_2|S}$, and $p_{\tilde{S}_1|\tilde{S}_2}$ as probability transition matrices.

The following result is obvious and its proof is omitted.

Lemma 2.2 *If $p_{\tilde{S}_1|S}$ is a stochastically degraded version of $p_{\tilde{S}_2|S}$, then*

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}_1|S}) \leq C(p_{Y|X,S}, p_S, p_{\tilde{S}_2|S}).$$

Next we extend Lemma 2.1 to the general case by characterizing the condition under which $p_{\tilde{S}|S}$ is a stochastically degraded version of $p_{\tilde{S}^{(\epsilon)}_{\mathrm{GE}}|S}$.

Lemma 2.3 *$p_{\tilde{S}|S}$ is a stochastically degraded version of $p_{\tilde{S}^{(\epsilon)}_{GE}|S}$ if and only if*

$$\sum_{\tilde{s} \in \tilde{S}} \min_{s \in \mathcal{S}} p_{\tilde{S}|S}(\tilde{s}|s) \geq \epsilon. \tag{2.29}$$

Proof: The problem boils down to finding a necessary and sufficient condition for the existence of $p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}$ such that

$$p_{\tilde{S}|S}(\tilde{s}|s) = \sum_{\tilde{s}' \in \mathcal{S} \cup \{*\}} p_{\tilde{S}_{\mathrm{GE}}^{(\epsilon)}|S}(\tilde{s}'|s) p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|\tilde{s}'), \quad s \in \mathcal{S}, \tilde{s} \in \tilde{\mathcal{S}}. \tag{2.30}$$

It suffices to consider the case $\epsilon \in [0, 1)$ since Lemma 2.3 is trivially true when $\epsilon = 1$. Note that

$$\begin{aligned} &\sum_{\tilde{s}' \in \mathcal{S} \cup \{*\}} p_{\tilde{S}_{\mathrm{GE}}^{(\epsilon)}|S}(\tilde{s}'|s) p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|\tilde{s}') \\ &= (1-\epsilon) p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|s) + \epsilon p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|*), \quad s \in \mathcal{S}, \tilde{s} \in \tilde{\mathcal{S}}. \end{aligned} \tag{2.31}$$

Combining (2.30) and (2.31) gives

$$p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|s) = \frac{p_{\tilde{S}|S}(\tilde{s}|s) - \epsilon p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|*)}{1-\epsilon}, \quad s \in \mathcal{S}, \tilde{s} \in \tilde{\mathcal{S}}. \tag{2.32}$$

In light of (2.32),

$$\begin{aligned} &\sum_{\tilde{s} \in \tilde{\mathcal{S}}} p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|s) = 1, \quad s \in \mathcal{S}, \\ &\Longleftrightarrow \sum_{\tilde{s} \in \tilde{\mathcal{S}}} p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|*) = 1, \\ &p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|s) \geq 0, \quad s \in \mathcal{S}, \tilde{s} \in \tilde{\mathcal{S}}, \\ &\Longleftrightarrow \min_{s \in \mathcal{S}} p_{\tilde{S}|S}(\tilde{s}|s) \geq \epsilon p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}(\tilde{s}|*), \quad \tilde{s} \in \tilde{\mathcal{S}}. \end{aligned} \tag{2.33}$$

It can be readily seen that the existence of conditional distribution $p_{\tilde{S}|\tilde{S}_{\mathrm{GE}}^{(\epsilon)}}$ satisfying

(2.30) is equivalent to the existence of probability vector $(p_{\tilde{S}|\tilde{S}^{(\epsilon)}_{\mathrm{GE}}}(\tilde{s}|*))_{\tilde{s}\in\tilde{\mathcal{S}}}$ satisfying (2.33). Clearly, (2.29) is a necessary and sufficient condition for the existence of such $(p_{\tilde{S}|\tilde{S}^{(\epsilon)}_{\mathrm{GE}}}(\tilde{s}|*))_{\tilde{s}\in\tilde{\mathcal{S}}}$. □

Theorem 2.3 *For any binary-input channel $p_{Y|X,S}$, state distribution p_S, and side channel $p_{\tilde{S}|S}$,*

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \underline{C}(p_{Y|X,S}, p_S)$$

if

$$\sum_{\tilde{s}\in\tilde{S}} \min_{s\in\mathcal{S}} p_{\tilde{S}|S}(\tilde{s}|s) \geq 1 - e^{-1}. \tag{2.34}$$

Proof: In view of Lemmas 2.1, 2.2, and 2.3, we have

$$C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \leq \underline{C}(p_{Y|X,S}, p_S) \tag{2.35}$$

if (2.34) is satisfied.

Combining (2.6) and (2.35) completes the proof of Theorem 2.3. □

Now we proceed to prove Theorem 2.1 by translating (2.34) (which is a condition on $p_{\tilde{S}|S}$ that is universal for all binary input channels and state distributions) to an upper bound on $I(S;\tilde{S})$. This upper bound, however, depends inevitably on the state distribution.

For any $p_{\tilde{S}|S}$ violating (2.34) (i.e, $\sum_{\tilde{s}\in\tilde{S}} \min_{s\in\mathcal{S}} p_{\tilde{S}|S}(\tilde{s}|s) < 1 - e^{-1}$), we have

$$\begin{aligned} I(S;\tilde{S}) &\geq \frac{1}{2}\left(\sum_{s\in\mathcal{S},\tilde{s}\in\tilde{\mathcal{S}}} p_S(s) \left| p_{\tilde{S}}(\tilde{s}) - p_{\tilde{S}|S}(\tilde{s}|s) \right| \right)^2 \qquad (2.36)\\ &\geq \frac{1}{2}\left(\sum_{\tilde{s}\in\tilde{\mathcal{S}}} p_S(s(\tilde{s})) \left| p_{\tilde{S}}(\tilde{s}) - p_{\tilde{S}|S}(\tilde{s}|s(\tilde{s})) \right| \right)^2\\ &\geq \frac{1}{2}\left(\rho \sum_{\tilde{s}\in\tilde{\mathcal{S}}} \left| p_{\tilde{S}}(\tilde{s}) - p_{\tilde{S}|S}(\tilde{s}|s(\tilde{s})) \right| \right)^2\\ &\geq \frac{1}{2}\left(\rho \left| \sum_{\tilde{s}\in\tilde{\mathcal{S}}} p_{\tilde{S}}(\tilde{s}) - \sum_{\tilde{s}\in\tilde{\mathcal{S}}} p_{\tilde{S}|S}(\tilde{s}|s(\tilde{s})) \right| \right)^2\\ &> \frac{\rho^2}{2e^2}, \end{aligned}$$

where (2.36) is due to Pinsker's inequality (Csiszár and Körner, 2011, p. 44), and $s(\tilde{s})$ is a minimizer of $\min_{s\in\mathcal{S}} p_{\tilde{S}|S}(\tilde{s}|s)$, $\tilde{s} \in \tilde{\mathcal{S}}$. As a consequence, (2.34) must hold if $I(S;\tilde{S}) \leq \frac{\rho^2}{2e^2}$. This completes the proof of Theorem 2.1.

2.3 Proof of Theorem 2.2

First consider the special case where $p_{\tilde{S}|S}$ is a generalized symmetric channel (with crossover probability $q \in [0, \frac{1}{|\mathcal{S}|}]$) defined as

$$p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}(\tilde{s}|s) = \begin{cases} 1 - (|\mathcal{S}| - 1)q, & \tilde{s} = s, \\ q, & \text{otherwise}, \end{cases} \qquad s \in \mathcal{S}, \tilde{s} \in \mathcal{S}.$$

Lemma 2.4 *$C'(p_{Y|X,S}, p_S, p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}) = \overline{C}(p_{Y|X,S}, p_S)$ if and only if*

$$\min_{x\in\mathcal{X}_+, s\in\mathcal{S}} \frac{p_{\hat{X}|S}(x|s)}{\sum_{s'\in\mathcal{S}} p_{\hat{X}|S}(x|s')} \geq q \tag{2.37}$$

for some $p_{\hat{X}|S} \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of maximizers of the optimization problem in (2.5), and $\mathcal{X}_+ = \{x \in \mathcal{X} : \sum_{s\in\mathcal{S}} p_{\hat{X}|S}(x|s) > 0\}$.

Proof: Clearly, $C'(p_{Y|X,S}, p_S, p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}) = \overline{C}(p_{Y|X,S}, p_S)$ if and only if there exists $p_{\hat{X}|S} \in \mathcal{P}$ that is a stochastically degraded version of $p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}$. When $q = \frac{1}{|\mathcal{S}|}$, (2.37) is equivalent to the desired condition that $\hat{X}$ needs to be independent of S. When $q \in [0, \frac{1}{|\mathcal{S}|})$, $p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}$ is invertible and

$$p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} = \begin{pmatrix} \frac{q-1}{|\mathcal{S}|q-1} & \frac{q}{|\mathcal{S}|q-1} & \cdots & \frac{q}{|\mathcal{S}|q-1} \\ \frac{q}{|\mathcal{S}|q-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{q}{|\mathcal{S}|q-1} \\ \frac{q}{|\mathcal{S}|q-1} & \cdots & \frac{q}{|\mathcal{S}|q-1} & \frac{q-1}{|\mathcal{S}|q-1} \end{pmatrix}. \tag{2.38}$$

The problem boils down to finding a necessary and sufficient condition under which $p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} p_{\hat{X}|S}$ is a valid probability transition matrix (i.e., all entries are non-negative and the sum of each row vector is equal to 1). Note that

$$p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} p_{\hat{X}|S} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.39}$$

Moreover, all entries of $p^{-1}_{\tilde{S}^{(q)}_{GS}|S}p_{\hat{X}|S}$ are non-negative if and only if

$$\frac{-p_{\hat{X}|S}(x|s)+q\sum_{s'\in\mathcal{S}}p_{\hat{X}|S}(x|s')}{|\mathcal{S}|q-1}\geq 0,\quad x\in\mathcal{S},s\in\mathcal{S},$$

which is equivalent to (2.37). □

The following result is obvious and its proof is omitted.

Lemma 2.5 *If* $p_{\tilde{S}_1|S}$ *is a stochastically degraded version of* $p_{\tilde{S}_2|S}$*, then*

$$C'(p_{Y|X,S},p_S,p_{\tilde{S}_1|S})\leq C'(p_{Y|X,S},p_S,p_{\tilde{S}_2|S}).$$

Lemma 2.6 $p_{\tilde{S}^{(q)}_{GS}|S}$ *is a stochastically degraded version of* $p_{\tilde{S}|S}$ *if*

$$\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_+:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{\sum_{s'\in\mathcal{S}}p_{\hat{S}|S}(\hat{s}|s')}\leq q, \tag{2.40}$$

where $\hat{S}$ *is the maximum likelihood estimate of* S *based on* $\tilde{S}$*, and* $\mathcal{S}_+=\{\hat{s}\in\mathcal{S}:\sum_{s\in\mathcal{S}}p_{\hat{S}|S}(\hat{s}|s)>0\}$.

Proof: The case $q=\frac{1}{|\mathcal{S}|}$ is trivial. When $q\in[0,\frac{1}{|\mathcal{S}|})$, $p_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}$ is invertible and $p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}$ is given by (2.38). It can be shown (see the derivation of (2.39)) that the sum of each row of $p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}p_{\hat{S}|S}$ is equal to 1; moreover, the off-diagonal entries of $p^{-1}_{\tilde{S}^{(q)}_{\mathrm{GS}}|S}p_{\hat{S}|S}$ are non-positive if and only if

$$\frac{-p_{\hat{S}|S}(\hat{s}|s)+q\sum_{s'\in\mathcal{S}}p_{\hat{S}|S}(\hat{s}|s')}{|\mathcal{S}|q-1}\leq 0,\quad s\in\mathcal{S},\hat{s}\in\mathcal{S}_+:s\neq\hat{s},$$

which is equivalent to (2.40). Therefore, (2.40) ensures that $p_{\tilde{S}_{\mathrm{GS}}^{(q)}|S}^{-1} p_{\hat{S}|S}$ is a non-singular M-matrix, which in turn ensures that $p_{\hat{S}|S}^{-1} p_{\tilde{S}_{\mathrm{GS}}^{(q)}|S}$ exists and is a non-negative matrix (Plemmons, 1977). Hence, if (2.40) is satisfied, then $p_{\hat{S}|S}^{-1} p_{\tilde{S}_{\mathrm{GS}}^{(q)}|S}$ is a valid probability transition matrix (the requirement that the entries in each row of $p_{\hat{S}|S}^{-1} p_{\tilde{S}_{\mathrm{GS}}^{(q)}|S}$ add up to 1 is automatically satisfied), which implies that $p_{\tilde{S}_{\mathrm{GS}}^{(q)}|S}$ is a stochastically degraded version of $p_{\hat{S}|S}$ (and consequently a stochastically degraded version of $p_{\tilde{S}|S}$). □

Theorem 2.4 *For any binary-input channel $p_{Y|X,S}$, state distribution p_S, and side channel $p_{\tilde{S}|S}$,*

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \overline{C}(p_{Y|X,S}, p_S)$$

if

$$\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_+ : s\neq\hat{s}} \frac{p_{\hat{S}|S}(\hat{s}|s)}{\sum_{s'\in\mathcal{S}} p_{\hat{S}|S}(\hat{s}|s')} \le \frac{1}{(|\mathcal{S}|-1)e - |\mathcal{S}| + 2}, \tag{2.41}$$

where $\hat{S}$ is the maximum likelihood estimate of S based on $\tilde{S}$.

Proof: Since $|\mathcal{X}| = 2$, it follows from [Shulman and Feder 2004, Th. 2] that there exists $p_{\hat{X}|S} \in \mathcal{P}$ satisfying

$$p_{\hat{X}|S}(x|s) > e^{-1}, \quad x \in \mathcal{X}, s \in \mathcal{S}.$$

For such $p_{\hat{X}|S}$,

$$\min_{x\in\mathcal{X}_+, s\in\mathcal{S}} \frac{p_{\hat{X}|S}(x|s)}{\sum_{s'\in\mathcal{S}} p_{\hat{X}|S}(x|s')} \ge \frac{e^{-1}}{e^{-1} + (|\mathcal{S}|-1)(1-e^{-1})}$$

$$= \frac{1}{(|\mathcal{S}|-1)e-|\mathcal{S}|+2}.$$

In view of of Lemmas 2.4, 2.5, and 2.6, we have

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \geq \overline{C}(p_{Y|X,S}, p_S) \tag{2.42}$$

if (2.41) is satisfied. Combining (2.6) and (2.42) completes the proof of Theorem 2.4. □

Now we are in a position to prove Theorem 2.2. Let $\hat{S}$ and $\hat{S}'$ denote respectively the maximum likelihood estimate and the maximum *a posteriori* estimate of S based on $\tilde{S}$. According to [Ho and Verdú 2010, Th. 11],

$$\mathbb{P}(S \neq \hat{S}') \leq \frac{H(S|\tilde{S})}{2\log 2}. \tag{2.43}$$

It can be verified that

$$\begin{aligned}\sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_{\hat{S}|S}(\hat{s}|s) &\leq \sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_{\hat{S}'|S}(\hat{s}|s)\\ &\leq \frac{1}{\rho}\sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_S(s)p_{\hat{S}'|S}(\hat{s}|s)\\ &= \frac{\mathbb{P}(S\neq\hat{S}')}{\rho}. \end{aligned}\tag{2.44}$$

Substituting (2.43) into (2.44) yields

$$\sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_{\hat{S}|S}(\hat{s}|s) \leq \hbar \triangleq \frac{H(S|\tilde{S})}{2\rho\log 2}. \tag{2.45}$$

Note that

$$\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_{+}:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{\sum_{s'\in\mathcal{S}}p_{\hat{S}|S}(\hat{s}|s')}\leq\frac{\hbar}{\hbar+\mathbb{I}(\hbar\leq 1)}. \tag{2.46}$$

Indeed, (2.46) is trivially true when $\hbar > 1$; moreover, when $\hbar \leq 1$,

$$\begin{aligned}
&\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_{+}:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{\sum_{s'\in\mathcal{S}}p_{\hat{S}|S}(\hat{s}|s')}\\
&\leq\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_{+}:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{p_{\hat{S}|S}(\hat{s}|s)+p_{\hat{S}|S}(\hat{s}|\hat{s})}\\
&=\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_{+}:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{p_{\hat{S}|S}(\hat{s}|s)+1-\sum_{\hat{s}'\in\mathcal{S}:\hat{s}'\neq\hat{s}}p_{\hat{S}|S}(\hat{s}'|\hat{s})}\\
&\leq\max_{s\in\mathcal{S},\hat{s}\in\mathcal{S}_{+}:s\neq\hat{s}}\frac{p_{\hat{S}|S}(\hat{s}|s)}{2p_{\hat{S}|S}(\hat{s}|s)+1-\hbar} &(2.47)\\
&\leq\frac{\hbar}{\hbar+1}, &(2.48)
\end{aligned}$$

where (2.47) and (2.48) are due to (2.45). In view of Theorem 2.4, It suffices to have

$$\frac{\hbar}{\hbar+\mathbb{I}(\hbar\leq 1)}\leq\frac{1}{(|\mathcal{S}|-1)e-|\mathcal{S}|+2}. \tag{2.49}$$

Note that (2.49) is not satisfied when $\hbar > 1$ since its left-hand side is equal to 1 whereas its right-hand side is strictly less than 1 ($\hbar > 1$ implies $|\mathcal{S}| \geq 2$). When $\hbar \leq 1$, we can rewrite (2.49) as[4]

$$\hbar\leq\frac{1}{(|\mathcal{S}|-1)(e-1)},$$

[4] Note that $\hbar \leq \frac{1}{(|\mathcal{S}|-1)(e-1)}$ implies $\hbar \leq 1$ when $|\mathcal{S}| \geq 2$. The case $|\mathcal{S}| = 1$ is trivial since $\hbar$ can only take the value 0.

which is exactly the desired result. This completes the proof of Theorem 2.2.

In Appendix A.1, we give an alternative proof of Theorem 2.2 with a different threshold on $H(S|\tilde{S})$.

2.4 Extension and Discussion

2.4.1 Extension of Theorem 2.1

It is interesting to know to what extent Theorem 2.1 can be extended beyond the binary-input case. This subsection is largely devoted to answering this question. For any $p_{Y|X,S}$ and p_S, define

$$\underline{\epsilon}(p_{Y|X,S}, p_S) = \min\left\{\epsilon \in [0,1] : C(p_{Y|X,S}, p_S, p_{\tilde{S}^{(\epsilon)}_{GE}|S}) = \underline{C}(p_{Y|X,S}, p_S)\right\},$$

$$\underline{q}(p_{Y|X,S}, p_S) = \min\left\{q \in [0, \frac{1}{|\mathcal{S}|}] : C(p_{Y|X,S}, p_S, p_{\tilde{S}^{(q)}_{GS}|S}) = \underline{C}(p_{Y|X,S}, p_S)\right\}.$$

Proposition 2.1 *1. There exists $\alpha(p_{Y|X,S}, p_S) > 0$ such that $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \underline{C}(p_{Y|X,S}, p_S)$ for all $p_{\tilde{S}|S}$ satisfying $I(S;\tilde{S}) \leq \alpha(p_{Y|X,S}, p_S)$ if and only if $\underline{\epsilon}(p_{Y|X,S}, p_S) < 1$.*

2. $\underline{\epsilon}(p_{Y|X,S}, p_S) < 1$ if and only if

$$\sum_{y\in\mathcal{Y}, s\in\mathcal{S}} p_S(s)\delta(u,y,s) \times \log\frac{p_{Y|X,S}(y|\psi(u,*),s)}{\sum_{x\in\mathcal{X}} p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)} > 0, \; u \in \mathcal{U}_{+}\backslash\mathcal{G}_{\delta}, \quad (2.50)$$

where $\delta(u,y,s)$ and $\mathcal{G}_{\delta}$ are defined in (2.14) and (2.21), respectively, $p_{\hat{X}}$ is an

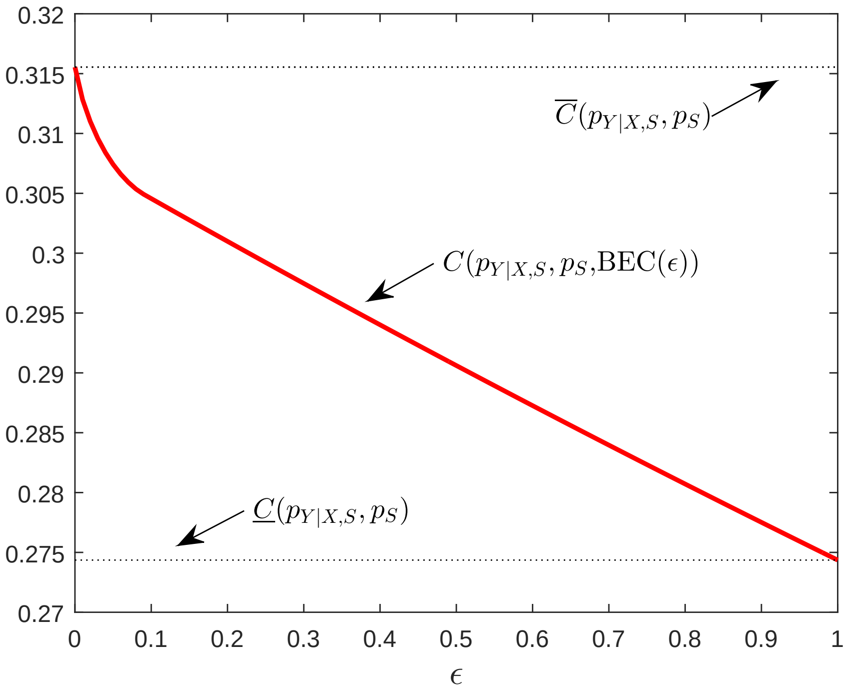

Figure 2.5: Plot of $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0, 1]$, where $p_{Y|X,S}$ and p_S are given by (2.62) and (2.63), respectively.

arbitrary maximizer of the optimization problem in (2.4), and

$$\mathcal{U}_+ = \Bigg\{ u \in \mathcal{U} : \sum_{y \in \mathcal{Y}, s \in \mathcal{S}} p_S(s) p_{Y|X,S}(y|\psi(u,*), s) \\ \times \log \frac{p_{Y|X,S}(y|\psi(u,*), s)}{\sum_{x \in \mathcal{X}} p_{\hat{X}}(x) p_{Y|X,S}(y|x, s)} = \underline{C}(p_{Y|X,S}, p_S) \Bigg\}.$$

Remark: All maximizers of the optimization problem in (2.4) give rise to the same $\sum_{x \in \mathcal{X}} p_{\hat{X}}(x) p_{Y|X,S}(y|x, s)$, $y \in \mathcal{Y}$, $s \in \mathcal{S}$ (Gallager, 1968, p. 96, Cor. 2).

Proof: The first statement can be easily extracted from the proof of Theorem 2.1.

Now we proceed to prove the second statement. First recall the definitions of $D_{\mathrm{GE}}(p_U, \epsilon, u)$ and $p_{\hat{U}}$ in (2.12) and (2.11), respectively. Since $p_{\hat{U}}$ is a capacity-achieving input distribution of channel $p_{Y,S|U}$ when $\epsilon = 1$, we must have

$$\begin{aligned} D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) &= \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{U} \text{ with } p_{\hat{U}}(u) > 0, \\ D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) &\leq \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{U} \text{ with } p_{\hat{U}}(u) = 0, \end{aligned}$$

which, together with the fact $\mathcal{U}_+ = \{u \in \mathcal{U} : D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) = \underline{C}(p_{Y|X,S}, p_S)\}$, implies

$$\{u \in \mathcal{U} : p_{\hat{U}}(u) > 0\} \subseteq \mathcal{U}_+, \tag{2.51}$$

$$D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) = \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{U}_+, \tag{2.52}$$

$$D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) < \underline{C}(p_{Y|X,S}, p_S), \quad \text{otherwise.} \tag{2.53}$$

It can be verified that

$$D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) = D_{\mathrm{GE}}(p_{\hat{U}}, 1, u), \quad \epsilon \in [0, 1], u \in \mathcal{G}_\delta. \tag{2.54}$$

Moreover, in view of (2.20), we can write (2.50) equivalently as

$$\left.\frac{\partial}{\partial \epsilon} D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u)\right|_{\epsilon=1} > 0, \quad u \in \mathcal{U}_+ \backslash \mathcal{G}_\delta. \tag{2.55}$$

According to (2.52)–(2.55), there exists $\epsilon(p_{Y|X,S}, p_S) \in [0, 1)$ such that

$$D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) = \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{U}_+ \cap \mathcal{G}_\delta, \tag{2.56}$$

$$D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) \leq \underline{C}(p_{Y|X,S}, p_S), \quad \text{otherwise} \tag{2.57}$$

for $\epsilon \geq \epsilon(p_{Y|X,S}, p_S)$. In light of (2.51) and the fact $\{u \in \mathcal{U} : p_{\hat{U}}(u) > 0\} \subseteq \mathcal{X} \subseteq \mathcal{G}_\delta$, we have

$$\{u \in \mathcal{U} : p_{\hat{U}}(u) > 0\} \subseteq \mathcal{U}_+ \cap \mathcal{G}_\delta. \tag{2.58}$$

Combining (2.56), (2.57), and (2.58) proves the "if" part of the second statement. Next we turn to the "only if" part of the second statement. Assuming the existence of $\epsilon(p_{Y|X,S}, p_S) \in [0, 1)$ such that $C(p_{Y|X,S}, p_S, p_{\tilde{S}^{(\epsilon)}|S}) = \underline{C}(p_{Y|X,S}, p_S)$ for $\epsilon \geq \epsilon(p_{Y|X,S}, p_S)$ (or equivalently $p_{\hat{U}}$ is a capacity-achieving input distribution of channel $p_{Y,S|U}$ for $\epsilon \geq \epsilon(p_{Y|X,S}, p_S)$), we must have

$$D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) \leq \underline{C}(p_{Y|X,S}, p_S), \quad \epsilon \geq \epsilon(p_{Y|X,S}, p_S), u \in \mathcal{U}. \tag{2.59}$$

It can be verified that

$$\begin{aligned}
&\frac{\partial^2}{\partial \epsilon^2} D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) \\
&= \sum_{y \in \mathcal{Y}, s \in \mathcal{S}} \frac{p_S(s)\delta^2(u, y, s)}{p_{Y|X,S}(y|\psi(u, s), s) + \epsilon\delta(u, y, s)} \\
&> 0, \quad \epsilon \in [0, 1], u \in \mathcal{U} \backslash \mathcal{G}_\delta.
\end{aligned} \tag{2.60}$$

Moreover, by the definition of $\mathcal{U}_+$,

$$D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) = \underline{C}(p_{Y|X,S}, p_S), \quad u \in \mathcal{U}_+. \tag{2.61}$$

Note that (2.59), (2.60), and (2.61) hold simultaneously for $u \in \mathcal{U}_+ \backslash \mathcal{G}_\delta$, from which (2.50) (or equivalently (2.55)) can be readily deduced. This completes the proof of

Proposition 2.1. □

As shown by the following example, the necessary and sufficient condition (2.50) is not always satisfied when $|\mathcal{X}| > 2$. Let

$$p_{Y|X,S}(y|x,s) = \begin{cases} 1, & (x,y,s) = (0,0,0) \text{ or } (1,1,1), \\ 0, & (x,y,s) = (0,1,0) \text{ or } (1,0,1), \\ \frac{2}{5}, & (x,y,s) = (1,0,0) \text{ or } (0,1,1), \\ \frac{3}{5}, & (x,y,s) = (1,1,0) \text{ or } (0,0,1), \\ \frac{3}{10}, & (x,y,s) = (2,0,0), \\ \frac{1}{5}, & (x,y,s) = (2,0,1), \\ \frac{7}{10}, & (x,y,s) = (2,1,0), \\ \frac{4}{5}, & (x,y,s) = (2,1,1), \end{cases} \tag{2.62}$$

$$p_S(0) = p_S(1) = \frac{1}{2}. \tag{2.63}$$

For this example, it can be verified that $\hat{u} \in \mathcal{U}_+ \backslash \mathcal{G}_\delta$ and

$$\sum_{y\in\mathcal{Y},s\in\mathcal{S}} p_S(s)\delta(\hat{u},y,s) \log \frac{p_{Y|X,S}(y|\psi(\hat{u},*),s)}{\sum_{x\in\mathcal{X}} p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)} < 0,$$

where $\psi(\hat{u},\cdot)$ is given by $\psi(\hat{u},0) = 2$, $\psi(\hat{u},1) = 1$, and $\psi(\hat{u},*) = 1$; indeed, Fig. 2.5 shows that $C(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon)) > \underline{C}(p_{Y|X,S}, p_S)$ for $\epsilon \in [0,1)$.

The proof of Proposition 2.1 in fact suggests a strategy for computing $\underline{\epsilon}(p_{Y|X,S}, p_S)$. Let $p_{\hat{X}}$ be an arbitrary maximizer of the optimization problem in (2.4) and define $p_{\hat{U}}$ according to (2.11). Note that

- $D_{\mathrm{GE}}(p_{\hat{U}}, 1, u) \leq \underline{C}(p_{Y|X,S}, p_S)$ for $u \in \mathcal{U}$ (see (2.52) and (2.53)),

- $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u)$ does not depend on ϵ for $u \in \mathcal{G}_\delta$ (see (2.54)),
- $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u)$ is a strictly convex function of ϵ for $u \in \mathcal{U}\backslash\mathcal{G}_\delta$ (see (2.60)).

Hence, for each $u \in \mathcal{U}$, there are three mutually exclusive cases.

1. $D_{\mathrm{GE}}(p_{\hat{U}},0,u) \leq \underline{C}(p_{Y|X,S},p_S)$: We have $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) \leq \underline{C}(p_{Y|X,S},p_S)$ for $\epsilon \in [\epsilon(u),1]$, where $\epsilon(u)=0$.
2. $D_{\mathrm{GE}}(p_{\hat{U}},0,u) > D_{\mathrm{GE}}(p_{\hat{U}},1,u) = \underline{C}(p_{Y|X,S},p_S)$ and $\frac{\partial}{\partial\epsilon}D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u)\Big|_{\epsilon=1} \leq 0$ (this case can arise only when $|\mathcal{X}|>2$): We have $D_{\mathrm{GE}}(p_{\hat{U}},0,u) > \underline{C}(p_{Y|X,S},p_S)$ for $\epsilon \in [0,\epsilon(u))$, where $\epsilon(u)=1$.
3. Otherwise: We have $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) > \underline{C}(p_{Y|X,S},p_S)$ for $\epsilon \in [0,\epsilon(u))$ and $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) \leq \underline{C}(p_{Y|X,S},p_S)$ for $\epsilon \in [\epsilon(u),1]$, where $\epsilon(u)$ is the unique solution of $D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) = \underline{C}(p_{Y|X,S},p_S)$ for $\epsilon \in (0,1)$.

It can be readily shown that

$$\underline{\epsilon}(p_{Y|X,S},p_S) = \max_{u\in\mathcal{U}} \epsilon(u). \tag{2.64}$$

We can compute $\underline{q}(p_{Y|X,S},p_S)$ in a similar way. Define

$$D_{\mathrm{GS}}(p_U,q,u) = D(p_{Y,S|U}(\cdot,\cdot|u)\|p_{Y,S}),$$

where

$$p_{Y,S|U}(y,s|u) = p_S(s)(p_{Y|X,S}(y|\psi(u,s),s) + q\omega(u,y,s))$$

with

$$\omega(u,y,s)=\sum_{\tilde{s}\in\mathcal{S}:\tilde{s}\neq s}p_{Y|X,S}(y|\psi(u,\tilde{s}),s)-(|\mathcal{S}|-1)p_{Y|X,S}(y|\psi(u,s),s),$$

$$u\in\mathcal{U},y\in\mathcal{Y},s\in\mathcal{S}.$$

Again, let $p_{\hat{U}}$ be defined[5] according to (2.11). It can be verified that

$$\begin{aligned}
&D_{\mathrm{GS}}(p_{\hat{U}},q,u)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}p_S(s)(p_{Y|X,S}(y|\psi(u,s),s)+q\omega(u,y,s))\\
&\qquad\times\log\frac{p_{Y|X,S}(y|\psi(u,s),s)+q\omega(u,y,s)}{\sum_{x\in\mathcal{X}}p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)},\quad q\in[0,\frac{1}{|\mathcal{S}|}],u\in\mathcal{U},\\
&\frac{\partial}{\partial q}D_{\mathrm{GS}}(p_{\hat{U}},q,u)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}p_S(s)\delta(u,y,s)\\
&\qquad\times\log\frac{p_{Y|X,S}(y|\psi(u,s),s)+q\omega(u,y,s)}{\sum_{x\in\mathcal{X}}p_{\hat{X}}(x)p_{Y|X,S}(y|x,s)},\quad q\in[0,\frac{1}{|\mathcal{S}|}],u\in\mathcal{U},\\
&\frac{\partial^2}{\partial q^2}D_{\mathrm{GS}}(p_{\hat{U}},q,u)\\
&=\sum_{y\in\mathcal{Y},s\in\mathcal{S}}\frac{p_S(s)\delta^2(u,y,s)}{p_{Y|X,S}(y|\psi(u,s),s)+q\omega(u,y,s)}>0,\quad q\in[0,\frac{1}{|\mathcal{S}|}],u\in\mathcal{U}\backslash\mathcal{G}_\omega,
\end{aligned}$$

where

$$\mathcal{G}_\omega=\{u\in\mathcal{U}:\omega(u,y,s)=0\text{ for all }y\in\mathcal{Y}\text{ and }s\in\mathcal{S}\}.$$

[5] Note that the underlying $\mathcal{U}$ depends on $\tilde{\mathcal{S}}$. In particular, $|\mathcal{U}|=|\mathcal{X}|^{|\mathcal{S}|}$ when $p_{\tilde{S}|S}$ is a generalized symmetric channel whereas $|\mathcal{U}|=|\mathcal{X}|^{|\mathcal{S}|+1}$ when $p_{\tilde{S}|S}$ is a generalized erasure channel.

Clearly,

- $D_{\mathrm{GS}}(p_{\hat{U}}, \frac{1}{|\mathcal{S}|}, u) \leq \underline{C}(p_{Y|X,S}, p_S)$ for $u \in \mathcal{U}$,
- $D_{\mathrm{GS}}(p_{\hat{U}}, q, u)$ does not depend on q for $u \in \mathcal{G}_\omega$,
- $D_{\mathrm{GS}}(p_{\hat{U}}, q, u)$ is a strictly convex function of q for $u \in \mathcal{U} \backslash \mathcal{G}_\omega$.

Hence, for each $u \in \mathcal{U}$, there are also three mutually exclusive cases.

1. $D_{\mathrm{GS}}(p_{\hat{U}}, 0, u) \leq \underline{C}(p_{Y|X,S}, p_S)$: We have $D_{\mathrm{GS}}(p_{\hat{U}}, q, u) \leq \underline{C}(p_{Y|X,S}, p_S)$ for $q \in [q(u), 1]$, where $q(u) = 0$.
2. $D_{\mathrm{GS}}(p_{\hat{U}}, 0, u) > D_{\mathrm{GS}}(p_{\hat{U}}, \frac{1}{|\mathcal{S}|}, u) = \underline{C}(p_{Y|X,S}, p_S)$ and $\frac{\partial}{\partial q} D_{\mathrm{GS}}(p_{\hat{U}}, q, u)\Big|_{q=\frac{1}{|\mathcal{S}|}} \leq 0$ (this case can arise only when $|\mathcal{X}| > 2$): We have $D_{\mathrm{GS}}(p_{\hat{U}}, 0, u) > \underline{C}(p_{Y|X,S}, p_S)$ for $q \in [0, q(u))$, where $q(u) = \frac{1}{|\mathcal{S}|}$.
3. Otherwise: We have $D_{\mathrm{GS}}(p_{\hat{U}}, q, u) > \underline{C}(p_{Y|X,S}, p_S)$ for $q \in [0, q(u))$ and $D_{\mathrm{GS}}(p_{\hat{U}}, q, u) \leq \underline{C}(p_{Y|X,S}, p_S)$ for $q \in [q(u), \frac{1}{|\mathcal{S}|}]$, where $q(u)$ is the unique solution of $D_{\mathrm{GS}}(p_{\hat{U}}, q, u) = \underline{C}(p_{Y|X,S}, p_S)$ for $q \in (0, \frac{1}{|\mathcal{S}|})$.

It can be readily shown that

$$\underline{q}(p_{Y|X,S}, p_S) = \max_{u \in \mathcal{U}} q(u). \tag{2.65}$$

For $p_{Y|X,S}$ and p_S illustrated in Fig. 2.2 (see also (2.9) and (2.10)), we show in Appendix A.2 that

$$\underline{\epsilon}(p_{Y|X,S}, p_S) = \begin{cases} \hat{\epsilon}(\theta), & \theta \in (0,1), \\ 0, & \text{otherwise}, \end{cases} \tag{2.66}$$

$$\underline{q}(p_{Y|X,S}, p_S) = \begin{cases} \hat{q}(\theta), & \theta \in (0,1), \\ 0, & \text{otherwise}, \end{cases} \tag{2.67}$$

where $\hat{\epsilon}(\theta)$ is the unique solution of

$$\begin{aligned} &\epsilon(1-\theta)\log 2\epsilon + (1-\epsilon(1-\theta))\log\frac{2(1-\epsilon(1-\theta))}{1+\theta} \\ &= (1-\theta)\log 2 + \theta\log\frac{2\theta}{1+\theta} \end{aligned}$$

for $\epsilon \in (0,1)$, and $\hat{q}(\theta)$ is the unique solution of

$$\begin{aligned} &q(1-\theta)\log 2q + (1-q(1-\theta))\log\frac{2(1-q(1-\theta))}{1+\theta} \\ &= \frac{1}{2}\left((1-\theta)\log 2 + \log\frac{2}{1+\theta} + \theta\log\frac{2\theta}{1+\theta}\right) \end{aligned}$$

for $q \in (0, \frac{1}{2})$. Setting $\theta = \frac{1}{2}$ gives $\underline{\epsilon}(p_{Y|X,S}, p_S) \approx 0.1$ (cf. Fig. 2.3) and $\underline{q}(p_{Y|X,S}, p_S) \approx 0.037$ (cf. Fig. 2.4).

2.4.2 Extension of Theorem 2.2

We shall extend Theorem 2.2 in a similar fashion. For any $p_{Y|X,S}$ and p_S, define

$$\begin{aligned} \overline{\epsilon}(p_{Y|X,S}, p_S) &= \max\left\{\epsilon \in [0,1] : C'(p_{Y|X,S}, p_S, p_{\tilde{S}^{(\epsilon)}_{\mathrm{GE}}|S}) = \overline{C}(p_{Y|X,S}, p_S)\right\}, \\ \overline{q}(p_{Y|X,S}, p_S) &= \max\left\{q \in [0, \frac{1}{|\mathcal{S}|}] : C'(p_{Y|X,S}, p_S, p_{\tilde{S}^{(q)}_{\mathrm{GE}}|S}) = \overline{C}(p_{Y|X,S}, p_S)\right\}. \end{aligned}$$

Proposition 2.2 *1. There exists $\beta(p_{Y|X,S}, p_S) > 0$ such that $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \overline{C}(p_{Y|X,S}, p_S)$ for all $p_{\tilde{S}|S}$ satisfying $H(S|\tilde{S}) \le \beta(p_{Y|X,S}, p_S)$ if and only if $\overline{q}(p_{Y|X,S}, p_S) >$*

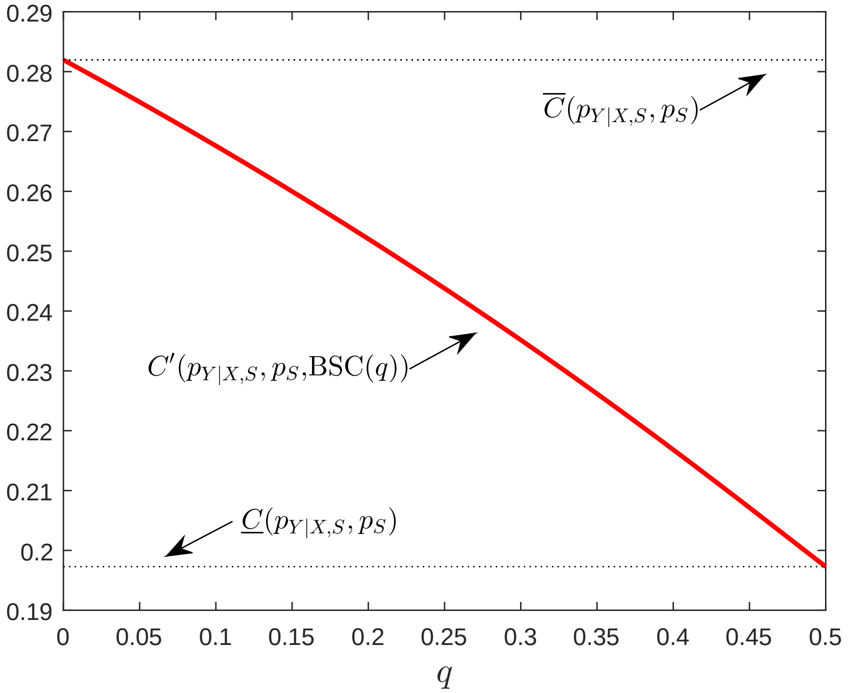

Figure 2.6: Plot of $C'(p_{Y|X,S}, p_S, \text{BSC}(q))$ against q for $q \in [0, \frac{1}{2}]$, where $p_{Y|X,S}$ and p_S are given by (2.69) and (2.70), respectively.

0.

2. *$\overline{q}(p_{Y|X,S}, p_S) > 0$ if and only if there exists $p_{\hat{X}|S} \in \mathcal{P}$ such that*

$$\{x \in \mathcal{X} : p_{\hat{X}|S}(x|s) > 0\} = \mathcal{X}_+, \quad s \in \mathcal{S}. \tag{2.68}$$

Proof: The first statement can be easily extracted from the proof of Theorem 2.2. The second statement is a consequence of Lemma 2.4. □

As shown by the following example, the necessary and sufficient condition (2.68) is not always satisfied when $|\mathcal{X}| > 2$. Let

$$p_{Y|X,S}(y|x,s) = \begin{cases} 1, & (x,y,s) = (0,0,0) \text{ or } (2,1,1), \\ 0, & (x,y,s) = (0,1,0) \text{ or } (2,0,1), \\ \frac{2}{5}, & (x,y,s) = (1,0,0) \text{ or } (0,1,1), \\ \frac{3}{5}, & (x,y,s) = (1,1,0) \text{ or } (0,0,1), \\ \frac{4}{5}, & (x,y,s) = (2,0,0) \text{ or } (1,1,1), \\ \frac{1}{5}, & (x,y,s) = (2,1,0) \text{ or } (1,0,1), \end{cases} \tag{2.69}$$

$$p_S(0) = p_S(1) = \frac{1}{2}. \tag{2.70}$$

For this example, it can be verified that the maximizer of the optimization problem in (2.5), denoted by $p_{\hat{X}|S}$, is unique and

$$\{x \in \mathcal{X} : p_{\hat{X}|S}(x|0) > 0\} = \{0,1\},$$
$$\{x \in \mathcal{X} : p_{\hat{X}|S}(x|1) > 0\} = \{0,2\};$$

indeed, Fig. 2.6 shows that $C'(p_{Y|X,S}, p_S, \mathrm{BSC}(q)) < \overline{C}(p_{Y|X,S}, p_S)$ for $q \in (0, \frac{1}{2}]$.

In view of Lemmas 2.3 and 2.4, we have

$$\overline{\epsilon}(p_{Y|X,S}, p_S) = \max_{p_{\hat{X}|S} \in \mathcal{P}} \sum_{x \in \mathcal{X}} \min_{s \in \mathcal{S}} p_{\hat{X}|S}(x|s), \tag{2.71}$$

$$\overline{q}(p_{Y|X,S}, p_S) = \max_{p_{\hat{X}|S} \in \mathcal{P}} \min_{x \in \mathcal{X}_+, s \in \mathcal{S}} \frac{p_{\hat{X}|S}(x|s)}{\sum_{s' \in \mathcal{S}} p_{\hat{X}|S}(x|s')}. \tag{2.72}$$

Note that $\mathcal{P}$ does not depend on p_S (under the assumption $\rho > 0$); as a consequence, $\overline{\epsilon}(p_{Y|X,S}, p_S)$ and $\overline{q}(p_{Y|X,S}, p_S)$ do not depend on p_S either. For $p_{Y|X,S}$ and p_S illustrated in Fig. 2.2 (see also (2.9) and (2.10)), we show in Appendix A.3 that

$$\overline{\epsilon}(p_{Y|X,S}, p_S) = \begin{cases} 2\left(1 + (1-\theta)\theta^{\frac{\theta}{1-\theta}}\right)^{-1} \theta^{\frac{\theta}{1-\theta}}, & \theta \in (0,1), \\ 1, & \text{otherwise}, \end{cases} \tag{2.73}$$

$$\overline{q}(p_{Y|X,S}, p_S) = \begin{cases} \left(1 + (1-\theta)\theta^{\frac{\theta}{1-\theta}}\right)^{-1} \theta^{\frac{\theta}{1-\theta}}, & \theta \in (0,1), \\ \frac{1}{2}, & \text{otherwise}. \end{cases} \tag{2.74}$$

Setting $\theta = \frac{1}{2}$ gives $\overline{\epsilon}(p_{Y|X,S}, p_S) = \frac{4}{5}$ (cf. Fig. 2.3) and $\overline{q}(p_{Y|X,S}, p_S) = \frac{2}{5}$ (cf. Fig. 2.4).

2.4.3 Two Implicit Conditions

In this subsection, we shall examine the following two implicit conditions in Theorem 2.1:

1. perfect state information at the decoder,

2. causal noisy state observation at the encoder.

If no state information is available at the decoder, then the channel capacity is given by

$$\tilde{C}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \triangleq \max_{p_U} I(U;Y),$$

where the joint distribution of $(U, X, Y, S, \tilde{S})$ is given by (2.2). Furthermore, if there is also no state information available at the encoder, then the channel capacity becomes

$$\underline{\tilde{C}}(p_{Y|X,S}, p_S) \triangleq \max_{p_X} I(X;Y), \tag{2.75}$$

where $p_{X,Y,S}(x,y,s) = p_X(x)p_S(s)p_{Y|X,S}(y|x,s)$. Define

$$\underline{\tilde{\epsilon}}(p_{Y|X,S}, p_S) = \min\left\{\epsilon \in [0,1] : \tilde{C}(p_{Y|X,S}, p_S, p_{\tilde{S}^{(\epsilon)}_{\mathrm{GE}}|S}) = \underline{\tilde{C}}(p_{Y|X,S}, p_S)\right\}.$$

The proof of the following result is similar to that of Proposition 2.1 and is omitted.

Proposition 2.3 *1. There exists $\tilde{\alpha}(p_{Y|X,S}, p_S) > 0$ such that $\tilde{C}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \underline{\tilde{C}}(p_{Y|X,S}, p_S)$ for all $p_{\tilde{S}|S}$ satisfying $I(S;\tilde{S}) \le \tilde{\alpha}(p_{Y|X,S}, p_S)$ if and only if $\underline{\tilde{\epsilon}}(p_{Y|X,S}, p_S)$ < 1.*

2. $\underline{\tilde{\epsilon}}(p_{Y|X,S}, p_S) < 1$ if and only if

$$\begin{aligned}&\sum_{y\in\mathcal{Y}}\left(\sum_{s\in\mathcal{S}} p_S(s)\delta(u,y,s)\right)\\ &\times \log\frac{\sum_{s\in\mathcal{S}} p_S(s)p_{Y|X,S}(y|\psi(u,*),s)}{\sum_{x\in\mathcal{X},s\in\mathcal{S}} p_{\hat{X}}(x)p_S(s)p_{Y|X,S}(y|x,s)} > 0,\\ &\qquad\qquad u \in \tilde{\mathcal{U}}_{+}\backslash\tilde{\mathcal{G}}_{\delta},\end{aligned} \tag{2.76}$$

where $\delta(u,y,s)$ is defined in (2.14), $p_{\hat{X}}$ is an arbitrary maximizer of the optimization problem in (2.75), and

$$\tilde{\mathcal{G}}_{\delta} = \left\{u \in \mathcal{U} : \sum_{s\in\mathcal{S}} p_S(s)\delta(u,y,s) = 0 \text{ for all } y \in \mathcal{Y}\right\},$$

$$\tilde{\mathcal{U}}_{+} = \Bigg\{ u \in \mathcal{U} : \sum_{y\in\mathcal{Y}} \Bigg(\sum_{s\in\mathcal{S}} p_S(s) p_{Y|X,S}(y|\psi(u,*),s) \Bigg) \\ \times \log \frac{\sum_{s\in\mathcal{S}} p_S(s) p_{Y|X,S}(y|\psi(u,*),s)}{\sum_{x\in\mathcal{X},s\in\mathcal{S}} p_{\hat{X}}(x) p_S(s) p_{Y|X,S}(y|x,s)} = \underline{\tilde{C}}(p_{Y|X,S}, p_S) \Bigg\}.$$

As shown by the following example, the necessary and sufficient condition (2.76) is not always satisfied even when $|\mathcal{X}| = 2$. Let

$$Y = X \oplus S, \quad \mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0,1\}, \tag{2.77}$$

$$p_S(1) = \mu \in (0, \frac{1}{2}), \tag{2.78}$$

where $\oplus$ is the modulo-2 addition. It can be verified that (2.76) is not satisfied for this example; indeed, Fig. 2.7 indicates that

$$\tilde{C}(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon)) > \underline{\tilde{C}}(p_{Y|X,S}, p_S), \quad \epsilon \in [0,1). \tag{2.79}$$

Here we give an alternative way to prove (2.79). Write $S = \tilde{S} \oplus \Delta$, where $\tilde{S}$ and Δ are two mutually independent Bernoulli random variables with

$$p_{\tilde{S}}(1) = \nu \in [0, \mu],$$
$$p_{\Delta}(1) = \frac{\mu - \nu}{1 - 2\nu}.$$

It is clear that

$$\tilde{C}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \log 2 - H(\Delta)$$
$$> \log 2 - H(S)$$

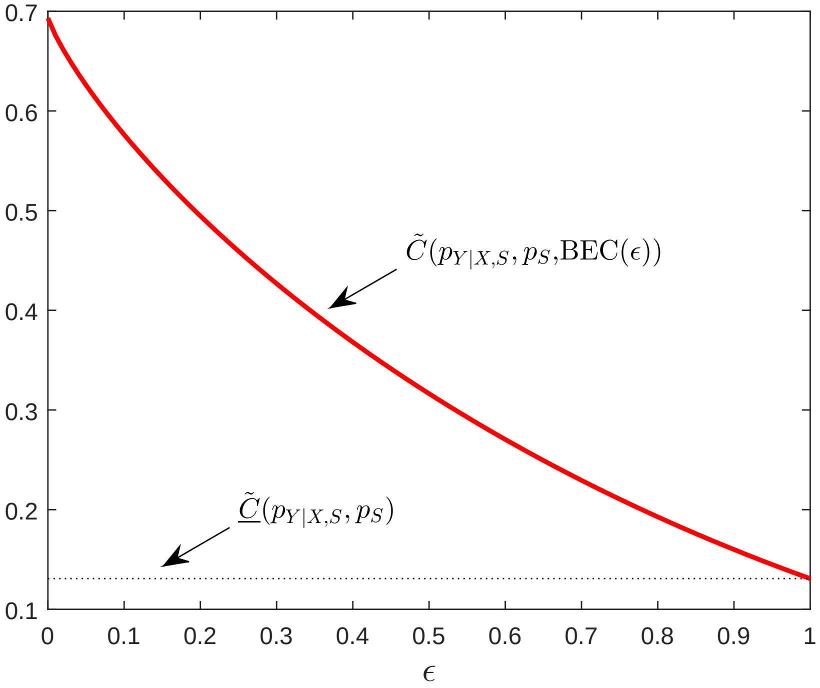

Figure 2.7: Plot of $\tilde{C}(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0, 1]$, where $p_{Y|X,S}$ and p_S are given by (2.77) with $\mu = \frac{1}{4}$ and (2.78), respectively.

$$= \underline{\tilde{C}}(p_{Y|X,S}, p_S), \quad \nu \in (0, \mu]. \tag{2.80}$$

In light of Lemma 2.3, $p_{\tilde{S}|S}$ is a stochastically degraded version of BEC(ϵ) and consequently

$$\tilde{C}(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon)) \geq \tilde{C}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \tag{2.81}$$

if $H(S) - H(\Delta) \leq \frac{\mu^2(1-\epsilon)^2}{2}$. Combining (2.80) and (2.81) proves (2.79).

Now we proceed to examine the second implicit condition. If the noisy state observation is available non-causally at the encoder, the Gelfand-Pinsker Theorem (Gel'fand and Pinsker, 1980) (see also (Gamal and Kim, 2011, Th. 7.3)) indicates

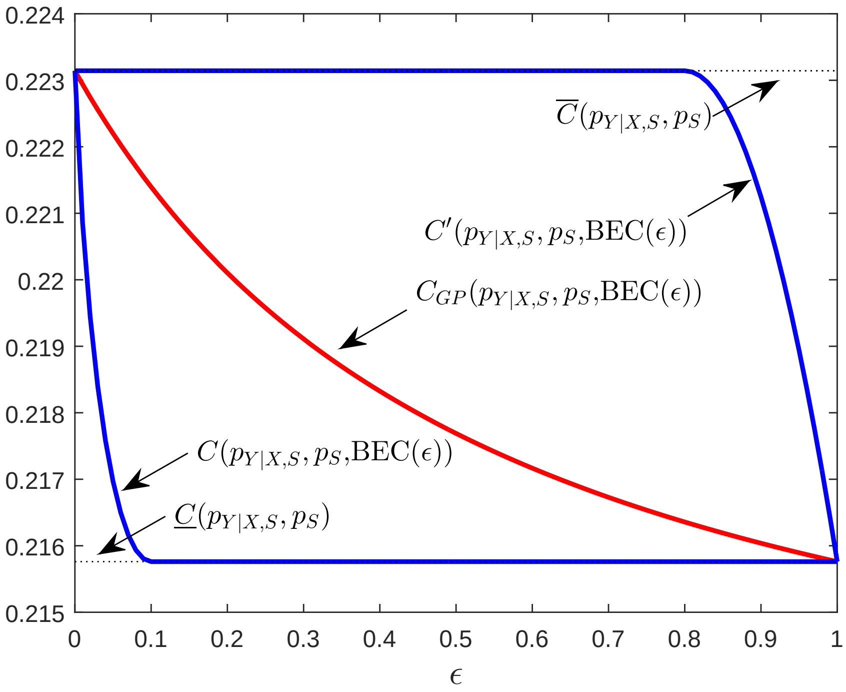

Figure 2.8: Plot of $C_{\text{GP}}(p_{Y|X,S}, p_S, \text{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0, 1]$, where $p_{Y|X,S}$ and p_S are given by (2.9) with $\theta = \frac{1}{2}$ and (2.10), respectively.

that the channel capacity is given by

$$C_{\text{GP}}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \triangleq \max_{p_{U|\tilde{S}}} I(U; Y, S) - I(U; \tilde{S}),$$

where the joint distribution of $(U, X, Y, S, \tilde{S})$ factors as

$$p_{U,X,Y,S,\tilde{S}}(u, x, y, s, \tilde{s}) = p_S(s)p_{\tilde{S}|S}(\tilde{s}|s)p_{U|\tilde{S}}(u|\tilde{s})\mathbb{I}(x = \psi(u, \tilde{s}))p_{Y|X,S}(y|x, s),$$
$$u \in \mathcal{U}, x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}, \tilde{s} \in \tilde{S}.$$

It turns out that $C_{\mathrm{GP}}(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$ is bounded between $C(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$ and $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S})$, i.e.,

$$\begin{aligned} C(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) &\leq C_{GP}(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) \\ &\leq C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}). \end{aligned}$$

Indeed, the first inequality is obvious, and the second one holds because

$$\begin{aligned} I(U;Y,S) - I(U;\tilde{S}) &\leq I(U;Y,S) - I(U;S) \\ &= I(U;Y|S) \\ &\leq I(X;Y|S). \end{aligned}$$

In Fig. 2.8 we plot $C_{\mathrm{GP}}(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ against ϵ for $\epsilon \in [0,1]$, where $p_{Y|X,S}$ and p_S are given by (2.9) with $\theta = \frac{1}{2}$ and (2.10), respectively; it can be seen that $C_{\mathrm{GP}}(p_{Y|X,S}, p_S, \mathrm{BEC}(\epsilon))$ is strictly greater than $\underline{C}(p_{Y|X,S}, p_S)$ except when $\epsilon = 1$. So the causality condition on the noisy state observation at the encoder is not superfluous for Theorem 2.1.

2.5 Conclusion

We have shown that the capacity of binary-input[6] channels is very "sensitive" to the quality of the encoder side information whereas the generalized probing capacity is very "robust". Here the words "sensitive" and "robust" should not be understood in a quantitative sense. Indeed, it is known (Shulman and Feder, 2004) that, when

[6] In fact, both numerical simulation and theoretical analysis suggest that similar results hold for many (but not all) non-binary input channels.

$|\mathcal{X}| = 2$, the ratio of $\underline{C}(p_{Y|X,S}, p_S)$ to $\overline{C}(p_{Y|X,S}, p_S)$ is at least 0.942 and the difference between these two quantities is at most $\sim$0.011 bit; in other words, the gain that can be obtained by exploiting the encoder side information (or the loss that can be incurred by ignoring the encoder side information) is very limited anyway.

Binary signalling is widely used, especially in wideband communications. So our work might have some practical relevance. However, great caution should be exercised in interpreting Theorems 2.1 and 2.2. Specifically, both results rely on the assumption that the channel state takes values from a finite set[7], which is not necessarily satisfied in reality; moreover, the freedom of power control in real communication systems is not captured by our results. Nevertheless, our work can be viewed as an initial step towards a better understanding of the fundamental performance limits of communication systems where the transmitter side information and the receiver side information are not deterministically related.

Finally, it is worth mentioning that our results might have their counterparts in source coding.

[7] In contrast, the assumption $|\mathcal{Y}| < \infty$ and $|\tilde{\mathcal{S}}| < \infty$ is not essential

Chapter 3

Intrinsic Capacity

3.1 Introduction

The capacity of a channel with state may be increased by utilizing the information about the state status available at the encoder and/or the decoder. In other words, the knowledge about the channel may be used to increase the channel capacity. If, in the extreme case, we have all the knowledge about the channel, then the channel reduces to a deterministic channel. In this case, can we surely have a higher capacity?

The answer is “no”, as is shown by the following example. Consider the binary symmetric channel with crossover probability 0.5:

$$p_{Y|X} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix},$$

where each entry $p_{Y|X}(y|x)$ denoting the conditional probability of output y given input x with $\mathcal{X}, \mathcal{Y} = \{0, 1\}$. The channel capacity is clearly zero.

Suppose that the channel has a binary state S with $p_S(0) = p_S(1) = 0.5$. Let us consider the following two models.

Model 1. For every realization of S, $p_{Y|X,S}$ is either $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Model 2. For every realization of S, $p_{Y|X,S}$ is either $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ or $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$.

It is easy to verify that $p_{Y|X} = \sum_{S \in \mathcal{S}} p_{Y|X,S} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$. If the actual model of the channel is Model 1, with the state information S available at the encoder and decoder, the channel becomes a deterministic perfect channel, so that the capacity increases to one. On the contrary, if the underlying mechanism of the channel is Model 2, then for every realization of S, the channel becomes a deterministic useless channel, and hence, even with S known at both sides, the channel capacity is still zero. This example shows that although the whole knowledge of a channel can be used to eliminate the uncertainty of the channel, it does not necessarily increase its capacity.

Now let us suppose that we can eliminate all the uncertainty of a DMC regardless of time and cost. Ultimately, the channel can be seen as a deterministic DMC with state, though the realization of this ultimate model is uncertain and not unique which may be determined by the physical nature of the channel. Under this premise, any DMC can be seen as certain realization of some deterministic channel with state. Let us call such state as *intrinsic state*. The capacity of the channel with intrinsic state known at the encoder and/or the decoder is called the channel's *intrinsic capacity*. The greatest possible intrinsic capacity of the channel is called the *upper intrinsic* *capacity*. The smallest possible intrinsic capacity of the channel is called the *lower intrinsic capacity*.

Whereas the intrinsic capacity of a channel is uncertain and not unique, the lower and the upper intrinsic capacity are determined and unique by definition. On the other hand, according to the availability of the causal state information at the encoder and/or the decoder, there are three kinds of the lower and the upper intrinsic capacity. Throughout this chapter, we denote the lower and upper intrinsic capacity by $\underline{\mathrm{IC}}$ and $\overline{\mathrm{IC}}$, respectively. And we use subscript $_\mathrm{E}$, $_\mathrm{D}$ and $_\mathrm{ED}$ to indicates the cases where the state information is available at the encoder only, at the decoder only and at the both ends, respectively.

3.2 Problem Formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. A channel $p_{Y|X}$ is a stochastic matrix with each entry $p_{Y|X}(x, y)$ denoting the probability of output $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$. A deterministic channel is a special channel whose stochastic matrix is a zero-one matrix, so that it uniquely identifies a map of $\mathcal{X}$ into $\mathcal{Y}$. In the sequel, deterministic channels and maps will be regarded as equivalent objects, so that their notations and conventions can be integrated with no ambiguities.

It is clear that the set of all channels forms a convex polytope in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. We denote this polytope by $\mathcal{P}_{Y|X}$, which consists of all the possible $|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrices. The vertices of $\mathcal{P}_{Y|X}$ are exactly all the deterministic channels and hence every channel can be expressed as a convex combination of some deterministic channels. Such a convex decomposition is not unique and each decomposition essentially gives a channel with intrinsic state. Since a channel with intrinsic state may have a larger capacity when the state information is available at the encoder, the decoder, or both, we are interested in the potential gain of such a channel with intrinsic state.

Let us denote the set of all deterministic channels with input X and output Y by $\hat{\mathcal{P}}_{Y|X}$ with $|\hat{\mathcal{P}}_{Y|X}| = |\mathcal{X}|^{|\mathcal{Y}|}$, and the intrinsic state by $\hat{S}$ with finite alphabet $\hat{\mathcal{S}}$. Also, by setting $|\hat{\mathcal{S}}| = |\hat{\mathcal{P}}_{Y|X}|$ and making a bijection between $\hat{\mathcal{S}}$ and $\hat{\mathcal{P}}_{Y|X}$, each value in $\hat{\mathcal{S}}$ uniquely corresponds to a possible deterministic channel $\hat{p}_{Y|X,\hat{S}}$. Then the set of all possible convex combinations of a channel $p_{Y|X}$ is

$$\mathrm{dec}(p_{Y|X}) \triangleq \left\{ p_{\hat{S}} \in \mathcal{P}_{\hat{S}} \colon p_{Y|X} = \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,S}(\cdot|\cdot,\hat{s}) \right\}.$$

where $\mathcal{P}_{\hat{S}}$ is the set of all probability distributions over $\hat{\mathcal{S}}$ and can be regarded as the set of all the possible $1 \times |\hat{\mathcal{P}}_{Y|X}|$ stochastic matrices or vectors.

Since $I(X;Y)$ can be seen as a function of the input distribution and the channel, let us define a function $\tilde{I}(\cdot,\cdot)$ with parameter p_X and $p_{Y|X}$ which is also equivalent to $I(X;Y)$ as follows.

$$\begin{aligned}\tilde{I}(p_X, p_{Y|X}) &\triangleq \sum_{x\in\mathcal{X}} p_X(x) D(p_{Y|X}||p_X p_{Y|X}) \\ &= I(X;Y)\end{aligned}$$

Note that a channel is determined by the distribution of its intrinsic state. And given the availability of the intrinsic state $\hat{S}$ at the encoder, or the decoder, or both, we have the following three kinds of intrinsic capacities (Gamal and Kim, 2011, Ch. 7).

1. $\hat{S}$ is available at the encoder only. We define the intrinsic capacity of channel

$p_{Y|X}$ as a function $J_{\mathrm{E}}(\cdot,\cdot)$ with parameters p_U and $p_{\hat{S}}$ by

$$\begin{aligned} C_{\mathrm{E}}(p_{Y|X}) &= \mathrm{IC}_{\mathrm{E}}(p_{\hat{S}}) \\ &\triangleq \max_{p_U} J_{\mathrm{E}}(p_U, p_{\hat{S}}) \\ &= \max_{p_U} \tilde{I}(p_U, p_{Y|U}) \end{aligned}$$

where the auxiliary random variable U is defined over alphabet $\mathcal{U}$ with $|\mathcal{U}| \leq \min\{(|\mathcal{X}|-1)|\hat{\mathcal{S}}|+1, |\mathcal{Y}|\}$, whose joint distribution with $(X, Y, \hat{S})$ factors as

$$\begin{aligned} p_{U,X,Y,\hat{S}}(u,x,y,\hat{s}) = p_U(u)p_{\hat{S}}(\hat{s})\mathbb{I}(x = \psi(u,\hat{s}))\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}), \\ u \in \mathcal{U}, x \in \mathcal{X}, y \in \mathcal{Y}, \hat{s} \in \hat{\mathcal{S}}, \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\psi(u,\cdot)$, $u \in \mathcal{U}$, are $|\mathcal{X}|^{|\hat{\mathcal{S}}|}$ different mappings from $\hat{\mathcal{S}}$ to $\mathcal{X}$. It is worth mentioning that $J_{\mathrm{E}}(p_U, p_{\hat{S}})$ is an equivalent transformation of $\tilde{I}(p_U, p_{Y|U})$ in which $p_{Y|U}$ is in fact a function of $p_{\hat{S}}$.

2. $\hat{S}$ is available at the decoder only. Similarly, we define the intrinsic capacity of channel $p_{Y|X}$ as a function $J_{\mathrm{D}}(\cdot,\cdot)$ with parameters p_U and $p_{\hat{S}}$ by

$$\begin{aligned} C_{\mathrm{D}}(p_{Y|X}) &= \mathrm{IC}_{\mathrm{D}}(p_{\hat{S}}) \\ &\triangleq \max_{p_X} J_{\mathrm{D}}(p_X, p_{\hat{S}}) \\ &= \max_{p_X} \sum_{\hat{s}} p_{\hat{S}}(\tilde{s})\tilde{I}(p_X, p_{Y|X}) \end{aligned}$$

where $p_{Y|X} = \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,S}(\cdot|\cdot,\hat{s})$. Also, $J_{\mathrm{D}}(p_X, p_{\hat{S}})$ is an equivalent transformation of $\sum_{\hat{s}} p_{\hat{S}}(\hat{s})\tilde{I}(p_X, p_{Y|X})$.

3. $\hat{S}$ is available at both the encoder and the decoder. We have the intrinsic capacity of channel $p_{Y|X}$ as follows.

$$\begin{aligned} C_{\mathrm{ED}}(p_{Y|X}) &= \mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}}) \\ &= \max_{p_{X|\hat{S}}} \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s}) \tilde{I}\left(p_{X|\hat{S}}(\cdot|\hat{s}), \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s})\right) \\ &= \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s}) \log \operatorname{rank}\left(\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s})\right). \end{aligned}$$

Then, given a channel $p_{Y|X}$, we can define its intrinsic capacity set by

$$\{\mathrm{IC}_f(p_{\hat{S}})\} \triangleq \left\{C_f(p_{Y|X})\colon p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})\right\}.$$

where the subscript $f \in \{\mathrm{E}, \mathrm{D}, \mathrm{ED}\}$ indicates the different cases of the availability of $\hat{S}$.

Furthermore, we define the lower and the upper intrinsic capacities of $p_{Y|X}$ by

$$\underline{\mathrm{IC}}_f(p_{Y|X}) \triangleq \inf_{p_{\hat{S}}\in\mathrm{dec}(p_{Y|X})} C_f(p_{Y|X}).$$

and

$$\overline{\mathrm{IC}}_f(p_{Y|X}) \triangleq \sup_{p_{\hat{S}}\in\mathrm{dec}(p_{Y|X})} C_f(p_{Y|X}).$$

respectively.

We close this section with some results on the analytic properties of J_f and IC_f.

For any $p_X, p'_X \in \mathcal{P}_X$,

$$\begin{aligned} \mathrm{d}(p_X, p'_X) &\triangleq \frac{1}{2}\left\|p_X - p'_X\right\|_1 \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} |p_X(x) - p'_X(x)| \end{aligned}$$

is called the *statistical distance* on $\mathcal{P}_X$. Given the product space $\mathcal{P}_X \times \mathcal{P}_Y$, we define its product metric by

$$\mathrm{d}((p_X, p_Y), (p'_X, p'_Y)) \triangleq \max\left\{\mathrm{d}(p_X, p'_X), \mathrm{d}(p_Y, p'_Y)\right\},$$

which induces the usual product topology. Thus for any channels $p_{Y|X}, p'_{Y|X} \in \mathcal{P}_{Y|X}$, we have the *channel distance*

$$\begin{aligned} \mathrm{d}(p_{Y|X}, p'_{Y|X}) &\triangleq \mathrm{d}\left(\left(p_{Y|X}(\cdot|x)\right)_{x\in\mathcal{X}}, \left(p'_{Y|X}(\cdot|x)\right)_{x\in\mathcal{X}}\right) \\ &= \max_{x\in\mathcal{X}} \mathrm{d}\left(p_{Y|X}(\cdot|x), p'_{Y|X}(\cdot|x)\right). \end{aligned}$$

Proposition 3.1 *(a) $J_E(p_U, p_{\hat{S}})$ is uniformly continuous, and it is convex in $p_{\hat{S}}$ for fixed p_U and is concave in p_U for fixed $p_{\hat{S}}$.*

(b) $J_D(p_X, p_{\hat{S}})$ is uniformly continuous, and it is linear in $p_{\hat{S}}$ for fixed p_X and is concave in p_X for fixed $p_{\hat{S}}$.

Proof: (a) The function $J_{\mathrm{E}}(p_U, p_{\hat{S}})$ can be rewritten as $\tilde{I}(p_U, g(p_{\hat{S}}))$ where

$$g(p_{\hat{S}}) = p_{Y|U}$$

in which each row of $p_{Y|U}$ can be expressed for a given u as follows

$$\begin{aligned} p_{Y|U}(\cdot|u) &= p_{\hat{S}} p_{Y|U,\hat{S}}(\cdot|u,\cdot) \\ &= p_{\hat{S}} \hat{p}_{Y|X,\hat{S}}(\cdot|\psi(u,\cdot),\cdot). \end{aligned}$$

Note that for a given u, $\hat{p}_{Y|X,\hat{S}}(\cdot|\psi(u,\cdot),\cdot)$ can be seen as a channel between $\hat{S}$ and Y. By Proposition B.2, for $p_{\hat{S}}, p'_{\hat{S}} \in \mathcal{P}_S$,

$$\begin{aligned} \mathrm{d}(g(p_{\hat{S}}), g(p'_{\hat{S}})) &= \max_{u\in\mathcal{U}} \mathrm{d}(p_{\hat{S}} \hat{p}_{Y|X,\hat{S}}(\cdot|\psi(u,\cdot),\cdot), p_{\hat{S}} \hat{p}'_{Y|X,\hat{S}}(\cdot|\psi(u,\cdot),\cdot)) \\ &\leq \mathrm{d}(p_{\hat{S}}, p'_{\hat{S}}), \end{aligned}$$

so that $g(\cdot)$ is uniformly continuous, and hence $J_{\mathrm{E}}(p_U, p_{\hat{S}})$ is uniformly continuous (Proposition B.5). It is also clear that $g(\cdot)$ is a linear function, so that $J_{\mathrm{E}}(p_U, p_{\hat{S}})$ is convex for fixed p_U and is concave for fixed $p_{\hat{S}}$ (Cover and Thomas, 2006, Th. 2.7.4).

(b) The function $J_{\mathrm{D}}(p_X, p_{\hat{S}})$ can be written as

$$J_{\mathrm{D}}(p_X, p_{\hat{S}}) = p_{\hat{S}} g(p_X),$$

where $g(p_X)$ is defined as

$$g(p_X) = \begin{pmatrix} \tilde{I}(p_X, \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_0)) \\ \tilde{I}(p_X, \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_2)) \\ \vdots \\ \tilde{I}(p_X, \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_{|\mathcal{S}|-1})) \end{pmatrix}.$$

By Propositions B.1 and B.2, $\tilde{I}\left(p_X, \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s})\right)$ is uniformly continuous on $\mathcal{P}_X$ and is bounded by $\log(\min\{|\mathcal{X}|, |\mathcal{Y}|\})$. Then for $p_{\hat{S}}, p'_{\hat{S}} \in \mathcal{P}_{\hat{S}}$ and $p_X, p'_X \in \mathcal{P}_X$, we have

$$\begin{aligned}\left|p_{\hat{S}} g(p_X) - p'_{\hat{S}} g(p'_X)\right| &= \left|p_{\hat{S}} g(p_X) - p'_{\hat{S}} g(p_X) + p'_{\hat{S}} g(p_X) - p'_{\hat{S}} g(p'_X)\right| \\ &\leq \left|p_{\hat{S}} g(p_X) - p'_{\hat{S}} g(p_X)\right| + \left|p'_{\hat{S}} g(p_X) - p'_{\hat{S}} g(p'_X)\right| \\ &\leq \left|p_{\hat{S}} - p'_{\hat{S}}\right| g(p_X) + p'_{\hat{S}} \left|g(p_X) - g(p'_X)\right| \\ &\leq \log(\min\{|\mathcal{X}|, |\mathcal{Y}|\}) \, ||p_{\hat{S}} - p'_{\hat{S}}||_1 + ||g(p_X) - g(p'_X)||_1\end{aligned}$$

which implies that J_{D} is uniformly continuous. The remaining part is straightforward (Cover and Thomas, 2006, Th. 2.7.4). □

Proposition 3.2 *For $f \in \{\mathrm{E}, \mathrm{D}, \mathrm{ED}\}$, $C_f(p_{\hat{S}})$ is uniformly continuous and convex (and indeed linear for $f = \mathrm{ED}$).*

Proof: Use Proposition 3.1 and Proposition B.6 for $f = \mathrm{E}$ or D. The case of $f = \mathrm{ED}$ is trivial because $\mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}})$ is a linear function of $p_{\hat{S}}$. □

3.3 The Simplest Case

In order to gain some insights into the intrinsic capacity, we first consider the simplest case: a binary channel $p_{Y|X}$ with the stochastic matrix

$$\begin{pmatrix} 1-\epsilon_1 & \epsilon_1 \\ \epsilon_2 & 1-\epsilon_2 \end{pmatrix},$$

where $0 \leq \epsilon_1, \epsilon_2 \leq 1$, and we assume $\epsilon_1 + \epsilon_2 \leq 1$ without loss of generality.

Note that there are only four types of binary deterministic channels, and let us make them correspond to the intrinsic state as follows.

$$\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_0) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_1) = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$
$$\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}_3) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It can be verified that $\mathrm{dec}(p_{Y|X})$ is a convex set, more specifically, a line segment with endpoints $p_{\hat{S}}$ and $p'_{\hat{S}}$ as

$$p_{\hat{S}} = (\epsilon_2, \epsilon_1, 1-\epsilon_1-\epsilon_2, 0),$$
$$p'_{\hat{S}} = \begin{cases} (\epsilon_2-\epsilon_1, 0, 1-\epsilon_2, \epsilon_1), & \text{for } \epsilon_1 \le \epsilon_2, \\ (0, \epsilon_1-\epsilon_2, 1-\epsilon_1, \epsilon_2), & \text{for } \epsilon_1 > \epsilon_2. \end{cases}$$

It is easy to see that

$$\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}}) = 1-\epsilon_1-\epsilon_2,$$
$$\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}}) = 1-|\epsilon_1-\epsilon_2|.$$

Since the input is binary, the binary uniform distribution is always capacity-achieving for every deterministic channel, either rank 1 or rank 2, so that $\mathrm{IC}_{\mathrm{D}}(p_{\hat{S}}) = \mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}})$ for every $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$, and therefore $\underline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}}) = \underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$ and $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}}) = \overline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$. The case of $f = \mathrm{E}$ is slightly complicated. We have the following result.

Proposition 3.3

$$\{\mathrm{IC_E}(p_{\hat{S}})\} = \left\{C\left(p_{Y|X} + t\boldsymbol{M}\right) : 0 \leq t \leq \min\{\epsilon_1, \epsilon_2\}\right\}$$

where $\boldsymbol{M} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. *Then,*

$$\begin{aligned} \underline{\mathrm{IC}}_{\mathrm{E}}(p_{\hat{S}}) &= C(p_{Y|X}), \\ \overline{\mathrm{IC}}_{\mathrm{E}}(p_{\hat{S}}) &= C\left(p_{Y|X} + \min\{\epsilon_1, \epsilon_2\}\boldsymbol{M}\right). \end{aligned}$$

Proof: By the observation above, since $p_{\hat{S}} = (\epsilon_2, \epsilon_1, 1-\epsilon_1-\epsilon_2, 0)$, we have

$$\mathrm{dec}(p_{Y|X}) = \{p_{\hat{S}} + t(-1,-1,1,1) \colon 0 \leq t \leq \min\{\epsilon_1, \epsilon_2\}\}.$$

Since the size of output alphabet is two, we only need to choose two maps from all the 16 maps of $\hat{\mathcal{P}}_{Y|X}$ into $\mathcal{X} = \{0,1\}$ for constructing the capacity-achieving distributions. We denote these two maps by $\psi(u_0, \hat{S})$ and $\psi(u_1, \hat{S})$, and then the optimal strategy for choosing $\psi(u_0, \cdot)$ and $\psi(u_1, \cdot)$ is to maximize $p_{Y|U}(y_0|u_0) - p_{Y|U}(y_0|u_1)$ (Proposition B.7), where $p_{Y|U}(y|u) = \sum_{\hat{s}\in\hat{S}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(y|\psi(u,\hat{s}),\hat{s})$. One of such pairs is

$$\psi(u_0, \hat{S}) = \begin{cases} 0, & \hat{S} = \hat{s}_0, \\ 0, & \hat{S} = \hat{s}_1, \\ 0, & \hat{S} = \hat{s}_2, \\ 1, & \hat{S} = \hat{s}_3, \end{cases} \quad \text{and} \quad \psi(u_1, \hat{S}) = \begin{cases} 0, & \hat{S} = \hat{s}_0, \\ 0, & \hat{S} = \hat{s}_1, \\ 1, & \hat{S} = \hat{s}_2, \\ 0, & \hat{S} = \hat{s}_3. \end{cases}$$

Then we have $p_{Y|U}$ as

$$p_{Y|U} = \begin{pmatrix} 1-\epsilon_1+t & \epsilon_1-t \\ \epsilon_2-t & 1-\epsilon_2+t \end{pmatrix}$$

The remaining part of the proof is then straightforward. □

In summary, from this simplest case, we obtain some interesting results that may be extended to general cases, as follows:

- The set $\mathrm{dec}(p_{Y|X})$ is closed and convex, so that $\{\mathrm{IC}_f(p_{\hat{S}})\}$ is compact and connected.
- $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}}) = |1-\epsilon_1-\epsilon_2|$ and $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}}) = 1-|\epsilon_1-\epsilon_2|$.
- $\underline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}}) = \underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$ and $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}}) = \overline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$.
- $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{\hat{S}}) = C(p_{Y|X})$.

3.4 The General Case

Let $p_{Y|X}$ be an arbitrary channel with $|\mathcal{X}| \geq 2$ and $|\mathcal{Y}| \geq 2$. In this section we will study its lower and upper intrinsic capacities.

3.4.1 $\mathrm{dec}(p_{Y|X})$

First, we have the following fundamental result.

Theorem 3.1 *The set* $\mathrm{dec}(p_{Y|X})$ *is a bounded, closed convex polytope. For each* $f \in \{\mathrm{E}, \mathrm{D}, \mathrm{ED}\}$*,* $\{\mathrm{IC}_f(p_{\hat{S}})\}$ *is a closed interval and* $\overline{\mathrm{IC}}_f(p_{\hat{S}})$ *can be attained at some vertex of* $\mathrm{dec}(p_{Y|X})$*. Furthermore,* $\underline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}})$ *and* $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$ *can also be attained at some vertices of* $\mathrm{dec}(p_{Y|X})$*.*

Proof: By definition, it is clear that $\mathrm{dec}(p_{Y|X})$ is a bounded, closed convex polytope, so that $\{\mathrm{IC}_f(p_{\hat{S}})\}$ is a closed interval (Proposition 3.2).

By Proposition 3.2 and [Bertsekas *et al.* 2003, Prop. 3.4.1], it is easy to see that $\mathrm{IC}_f(p_{\hat{S}})$ attains its maximum $\overline{\mathrm{IC}}_f(p_{\hat{S}})$ at some vertex of $\mathrm{dec}(p_{Y|X})$ and that $\mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}})$ attains its minimum $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{\hat{S}})$ at some vertex of $\mathrm{dec}(p_{Y|X})$.

For $f = \mathrm{D}$, it follows from the minimax theorem (Sion, 1958, Th. 3.4) and Proposition 3.1 that

$$\begin{aligned}\underline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}}) &= \min_{p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})} \max_{p_X \in \mathcal{P}_X} J_{\mathrm{D}}(p_X, p_{\hat{S}}) \\ &= \max_{p_X \in \mathcal{P}_X} \min_{p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})} J_{\mathrm{D}}(p_X, p_{\hat{S}}) \\ &= \max_{p_X \in \mathcal{P}_X} g(p_X),\end{aligned}$$

where, for every fixed p_X, the value of $g(p_X)$ is always attained at some vertex of $\mathrm{dec}(p_{Y|X})$ (Proposition 3.1 and (Bertsekas *et al.*, 2003, Prop. 3.4.1)). Therefore, $\underline{\mathrm{IC}}_{\mathrm{D}}(p_{\hat{S}})$ is attained at some vertex of $\mathrm{dec}(p_{Y|X})$. □

In light of Theorem 3.1, we proceed to study the structure of $\mathrm{dec}(p_{Y|X})$, namely, its vertices. Our approach is analogous to [Jurkat and Ryser 1968].

Proposition 3.4 *Let*

$$\mathfrak{S} = \left\{ \operatorname{supp}(\alpha_{\hat{S}})\colon \alpha_{\hat{S}} \in \mathbb{R}^{|\hat{\mathcal{S}}|}, \sum_{\hat{s}\in\hat{\mathcal{S}}} \alpha_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) = 0 \right\}.$$

A probability distribution $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$ is a vertex iff for $T \in \mathfrak{S}$, $T \subseteq \operatorname{supp}(p_{\hat{S}})$ implies $T = \emptyset$.

Proof: (Sufficiency) Given $p_{\hat{S}}$, if it can be expressed as a linear combination of some $p'_{\hat{S}}, p''_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$ as follows,

$$p_{\hat{S}} = tp'_{\hat{S}} + (1-t)p''_{\hat{S}}, \quad \text{for some } 0 < t < 1, \tag{3.1}$$

then $p'_{\hat{S}} - p''_{\hat{S}} = (p_{\hat{S}} - p''_{\hat{S}})/t$. Since

$$\begin{aligned} \sum_{\hat{s}\in\hat{\mathcal{S}}} (p'_{\hat{S}} - p''_{\hat{S}})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) &= \sum_{\hat{s}\in\hat{\mathcal{S}}} p'_{\hat{S}}\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) - \sum_{\hat{s}\in\hat{\mathcal{S}}} p''_{\hat{S}}\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) \\ &= p_{Y|X} - p_{Y|X} \\ &= 0 \end{aligned}$$

we have $\operatorname{supp}(p'_{\hat{S}} - p''_{\hat{S}}) \in \mathfrak{S}$. Note that $\operatorname{supp}(p'_{\hat{S}} - p''_{\hat{S}}) \subseteq \operatorname{supp}(p_{\hat{S}})$ from (3.1). Hence, if for $T \in \mathfrak{S}$, $T \subseteq \operatorname{supp}(p_{\hat{S}})$ implies $T = \emptyset$, we have

$$p'_{\hat{S}} - p''_{\hat{S}} = 0, \quad \text{for any } p'_{\hat{S}}, p''_{\hat{S}} \in \operatorname{dec}(p_{Y|X}) \text{ and satisfy (3.1)},$$

in other words, $p_{\hat{S}} = p'_{\hat{S}} = p''_{\hat{S}}$ is a vertex.

(Necessity) For every nonempty $T \in \mathfrak{S}$, there is a vector $\alpha_{\hat{S}} \in \mathbb{R}^{|\hat{\mathcal{S}}|}$ such that $\operatorname{supp}(\alpha_{\hat{S}}) = T$ and

$$\sum_{\hat{s}\in\hat{\mathcal{S}}} \alpha_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) = 0.$$

Let $p'_{\hat{S}} = p_{\hat{S}} + t\alpha_{\hat{S}}$ and $p''_{\hat{S}} = p_{\hat{S}} - t\alpha_{\hat{S}}$ with $t \neq 0$, so that $p_{\hat{S}} = (p'_{\hat{S}} + p''_{\hat{S}})/2$ with $p'_{\hat{S}} \neq p''_{\hat{S}}$. Since $p_{\hat{S}}$ is a vertex, $p'_{\hat{S}}$ and $p''_{\hat{S}}$ must not be elements of $\operatorname{dec}(p_{Y|X})$ for all $t \neq 0$. Note that

$$\begin{aligned}\sum_{\hat{s}\in\hat{\mathcal{S}}} p'_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) &= \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) + \sum_{\hat{s}\in\hat{\mathcal{S}}} \alpha_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) \\ &= p_{Y|X},\end{aligned}$$

and $\sum_{\hat{s}\in\hat{\mathcal{S}}} p''_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) = p_{Y|X}$ similarly. Therefore, $p'_{\hat{S}}$ and $p''_{\hat{S}}$ must not be valid distributions in the same time, in other words, $T \not\subseteq \operatorname{supp}(p_{\hat{S}})$. □

Below are several easy consequences of Proposition 3.4.

Proposition 3.5 *A probability distribution $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$ is a vertex iff for every $p'_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$, $\operatorname{supp}(p'_{\hat{S}}) \subseteq \operatorname{supp}(p_{\hat{S}})$ implies $p'_{\hat{S}} = p_{\hat{S}}$.*

Proposition 3.6 *If $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$ is a vertex, then*

$$|\operatorname{supp}(p_{\hat{S}})| \leq \left|\operatorname{supp}(p_{Y|X})\right| - |\mathcal{X}| + 1.$$

Proof: For $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$, we have

$$p_{Y|X} = \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}). \tag{3.2}$$

Because of $\sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}} = 1$, the equations (3.2) have at most $|\mathcal{X}|(|\mathcal{Y}|-1)+1$ linearly independent equations. This number can be further reduced by utilizing the information of $p_{Y|X}$. Note that if for some $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$p_{Y|X}(y|x) = \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}) = 0,$$

$p_{\hat{S}}(\hat{s})$ must be zero for all $\hat{s}$ with $\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}) = 1$. Therefore, the number of linearly independent equations of (3.2) is no more than $\left|\operatorname{supp}(p_{Y|X})\right| - |\mathcal{X}| + 1$. In other words, if $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$ is a vertex, $|\operatorname{supp}(p_{\hat{S}})| \le \left|\operatorname{supp}(p_{Y|X})\right| - |\mathcal{X}| + 1$. □

Proposition 3.6 provides an upper bound for the support size of a vertex in $\operatorname{dec}(p_{Y|X})$. On the other hand, the following result provides a lower bound for the support size of points in $\operatorname{dec}(p_{Y|X})$, including all the vertices of $\operatorname{dec}(p_{Y|X})$.

Proposition 3.7 *For any $p_{\hat{S}} \in \operatorname{dec}(p_{Y|X})$,*

$$|\operatorname{supp}(p_{\hat{S}})| \ge \max\left\{ \left\lceil \log_2 \left|\operatorname{supp}(p_{Y|X})\right| \right\rceil, \max_{x\in\mathcal{X}} \left|\operatorname{supp}\left(p_{Y|X}(\cdot|x)\right)\right| \right\}.$$

Proof: By conditioning, we have

$$p_{Y|X}(y|x) = \sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}).$$

Since $\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s})$ is either 0 or 1, the right-hand side can yield at most $2^{\left|\mathrm{supp}(p_{\hat{S}})\right|}$ different values, so that

$$2^{\left|\mathrm{supp}(p_{\hat{S}})\right|} \geq \left|\mathrm{supp}(p_{Y|X})\right|,$$

or $|\mathrm{supp}(p_{\hat{S}})| \geq \left\lceil \log_2 \left|\mathrm{supp}(p_{Y|X})\right| \right\rceil$.

On the other hand, given $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, every equation

$$\sum_{\hat{s}\in\hat{\mathcal{S}}} p_{\hat{S}}(\hat{s})\hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}) = p_{Y|X}(y|x) > 0.$$

must have at least one positive $p_{\hat{S}}$ for some

$$\hat{s} \in \left\{\hat{s} \in \hat{\mathcal{S}} \colon \hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}) = 1\right\}.$$

Since for every x, the sets $\left\{\hat{s} \in \hat{\mathcal{S}} \colon \hat{p}_{Y|X,\hat{S}}(y|x,\hat{s}) = 1\right\}_{y\in\mathcal{Y}}$ are mutually disjoint, we conclude that $|\mathrm{supp}(p_{\hat{S}})| \geq \left|\mathrm{supp}\left(p_{Y|X}(\cdot|x)\right)\right|$ for each $x \in \mathcal{X}$. Therefore, the proof is complete. □

3.4.2 $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathbf{ED}}(p_{Y|X})$

If we can enumerate all the vertices of $\mathrm{dec}(p_{Y|X})$, then by Theorem 3.1, we can certainly obtain the exact values of $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$. However, because of the complex structure of $\mathrm{dec}(p_{Y|X})$, we turn to estimating $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ by other approaches. The next result is a generalization of the Birkhoff-von Neumann Theorem, which will be very useful for our purpose. Our approach is an extension of the ideas in [Jurkat and Ryser 1968; Caron *et al.* 1996].

Theorem 3.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two* $1 \times |\mathcal{Y}|$ *integer-valued row vectors such that* $\mathbf{a} \leq \mathbf{b}$*, namely,* $a_y \leq b_y$ *for* $y \in \mathcal{Y}$*. Define*

$$\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}] \triangleq \left\{ p_{Y|X} \in \mathcal{P}_{Y|X} \colon \mathbf{a} \leq \mathbf{1} p_{Y|X} \leq \mathbf{b} \right\}$$
$$\hat{\mathcal{P}}_{Y|X}[\mathbf{a}, \mathbf{b}] \triangleq \mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}] \cap \hat{\mathcal{P}}_{Y|X}$$

where $\mathbf{1}$ *denotes the* $1 \times |\mathcal{X}|$ *all-one row vector. If* $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$ *is not empty, then* $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$ *is convex and the vertices of* $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$ *are exactly the matrices in* $\hat{\mathcal{P}}_{Y|X}[\mathbf{a}, \mathbf{b}]$.

Proof: It is clear that $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$, if nonempty, is a convex set. We will show that any matrix $p_{Y|X} \in \mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$ having non-integer entries cannot be a vertex of $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{b}]$. There are two cases:

Case (a) There is a non-integer entry in a off-boundary column.

Case (b) All non-integer entries are in the on-boundary columns.

Here, let us say a column is on boundary if its sum is either a_j or b_j, where j is the index of the column, and such a column is called a on-boundary column. The column whose sum is neither a_j nor b_j is called an off-boundary column.

In whichever the case, we can pick a non-integer entry, say the (i_0, j_0)-th entry, which must be in a non-boundary column for Case (a). By the following searching process, we will find a chain or loop of non-integer entries of the matrix, which will be used to prove that the matrix is not extremal.

1. Pick a non-integer entry as the start entry, and record the index of this entry,

say (i_0, j_0). Note that this entry must be in a non-boundary column for Case (a).

2. Assuming that the last recorded index is (i_k, j_k) with $k = 0, 1, 2, \cdots$, pick a non-integer entry (i_k, j_k) in the same row excluding (i_0, j_0), say the (i_k, j_{k+1})-th entry. Determine if any of the following conditions is true, and execute the corresponding operations.

   - if the j_{k+1}-th column is not on boundary, record the index (i_k, j_{k+1}). Note that all the recorded indices form a chain. Return the index chain as follows.

$$(i_0, j_0), (i_0, j_1), (i_1, j_1), \cdots, (i_k, j_k), (i_k, j_{k+1});$$

   - if j_{k+1}-th column has already been visited, i.e. $j_{k+1} = j_l$ for some $0 \leq l \leq k-1$, but the (i_k, j_{k+1})-th entry hasn't been picked, record the index (i_k, j_{k+1}). Note that there is a loop formed with some of the recorded indices. Return the loop indices as follows;

$$(i_l, j_l), (i_l, j_{l+1}), (i_{l+1}, j_{l+1}), \cdots, (i_k, j_k), (i_k, j_{k+1});$$

   - if the (i_k, j_{k+1})-th entry has been picked, record the index (i_k, j_{k+1}). Note that a loop is formed with the most recently picked four entries. Return the loop indices as follows;

$$(i_{k-1}, j_{k-1}), (i_{k-1}, j_k), (i_k, j_k), (i_k, j_{k+1});$$

If any of the conditions above is true, the searching process is finished. Otherwise, record the index of the picked entry as (i_k, j_{k+1}) and move on to the next step.

Remark: Because the (i_k, j_k)-th entry is not an integer, there exists at least another entry in the same row that is also not an integer.

3. Assuming that the last recorded index is (i_k, j_{k+1}), pick a non-integer entry in the same column and record its index as (i_{k+1}, j_{k+1}). Jump back to Step 2.

   Remark: If the j_{k+1}-th column is on boundary, then there exists at least another non-integer entry in the same column.

Given the returned indices, either a chain or a loop, we can construct a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix $\mathbf{M}$ by setting the entry of $\mathbf{M}$ corresponding to the first returned index as 1, the entry of $\mathbf{M}$ corresponding to the second returned index as -1, the entry of $\mathbf{M}$ corresponding to the third returned index as 1, the entry of $\mathbf{M}$ corresponding to the forth returned index as -1, and so on until all the entries corresponding to the returned indices have been assigned values. And all other entries are set to be zero. It is clear that

$$\mathbf{1M} = \boldsymbol{e}_{j_0} - \boldsymbol{e}_{j_{k+1}}, \quad \mathbf{M1}^{\mathsf{T}} = 0$$

given the chain indices and

$$\mathbf{1M} = 0, \quad \mathbf{M1}^{\mathsf{T}} = 0$$

given the loop indices , where $\boldsymbol{e}_j$ is a $1 \times |\mathcal{Y}|$ row vector with the j-th entry as 1 and all the other entries as 0.

Let $p'_{Y|X} = p_{Y|X} + \epsilon\mathbf{M}$ and $p''_{Y|X} = p_{Y|X} - \epsilon\mathbf{M}$. It is clear that $p'_{Y|X}, p''_{Y|X} \in \mathcal{P}_{Y|X}[a, b]$ for sufficiently small $\epsilon > 0$. It is also clear that $p_{Y|X} = \frac{1}{2}p'_{Y|X} + \frac{1}{2}p''_{Y|X}$ and

$p'_{Y|X} \neq p''_{Y|X}$, that is, $p_{Y|X}$ is not a vertex of $\mathcal{P}_{Y|X}[a,b]$.

Therefore, if denoting the set of all vertices of $\mathcal{P}_{Y|X}[a,b]$ by $\mathcal{P}^{(v)}_{Y|X}[a,b]$, we have $\mathcal{P}^{(v)}_{Y|X}[a,b] \subseteq \hat{\mathcal{P}}_{Y|X}[a,b]$. It remains to show that $\hat{\mathcal{P}}_{Y|X}[a,b] \subseteq \mathcal{P}^{(v)}_{Y|X}[a,b]$. For any $p_{Y|X} \in \hat{\mathcal{P}}_{Y|X}[a,b]$, if $p_{Y|X} = \alpha p'_{Y|X} + (1-\alpha)p''_{Y|X}$ with $p'_{Y|X}, p''_{Y|X} \in \mathcal{P}_{Y|X}[a,b]$ and $\alpha \in (0,1)$, then for every $1 \leq j \leq |\mathcal{X}|$,

$$\boldsymbol{e}_i p_{Y|X} = \alpha \boldsymbol{e}_i p'_{Y|X} + (1-\alpha)\boldsymbol{e}_i p''_{Y|X},$$

which however implies that $\boldsymbol{e}_i p'_{Y|X} = \boldsymbol{e}_i p''_{Y|X}$ for every $1 \leq j \leq |\mathcal{X}|$, or $p'_{Y|X} = p''_{Y|X}$. Therefore, it has been shown that each $p_{Y|X} \in \hat{\mathcal{P}}_{Y|X}[a,b]$ is a vertex of $\mathcal{P}^{(v)}_{Y|X}[a,b]$ which completes our proof. □

Now we proceed to estimate the lower and the upper intrinsic capacities. Since $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$ is a probability mass function (or equivalently, a probability measure) over $\hat{\mathcal{P}}_{Y|X}$ and $\mathrm{rank}(\hat{p}_{Y|X,\hat{S}}) \leq \min\{|\mathcal{X}|,|\mathcal{Y}|\}$, we define a probability mass function $\Gamma(r, p_{\hat{S}})$ which is called rank probability as follows.

$$\Gamma(r, p_{\hat{S}}) = \sum_{\hat{s}:\mathrm{rank}(\hat{p}_{Y|X,\hat{S}})=r} p_{\hat{S}}(\hat{s}).$$

where $r \in \{1, 2, \cdots, \min\{|\mathcal{X}|,|\mathcal{Y}|\}\}$. Further more, given a channel $p_{Y|X}$, the lower and the upper rank probabilities are defined by

$$\underline{\Gamma}(r, p_{Y|X}) = \min_{p_{\hat{S}\in \mathrm{dec}(p_{Y|X})}} \Gamma(r, p_{\hat{S}}),$$

$$\overline{\Gamma}(r, p_{Y|X}) = \max_{p_{\hat{S}\in \mathrm{dec}(p_{Y|X})}} \Gamma(r, p_{\hat{S}}),$$

respectively.

Proposition 3.8

$$\underline{\Gamma}(1, p_{Y|X}) = \left(g(p_{Y|X}) - |\mathcal{X}| + 1\right)^+,$$
$$\overline{\Gamma}(1, p_{Y|X}) = \sum_{y\in\mathcal{Y}} \alpha(y).$$

where

$$g(p_{Y|X}) = \max_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{Y|X}(y|x), \tag{3.3}$$

$$\alpha(y) = \min_{x\in\mathcal{X}} p_{Y|X}(y|x). \tag{3.4}$$

Proof: Note that

$$g(\hat{p}_{Y|X,\hat{S}}) \leq |\mathcal{X}| - 1, \quad \text{for every } \hat{p}_{Y|X,\hat{S}} \in \left\{\hat{p}_{Y|X,\hat{S}} : \mathrm{rank}(\hat{p}_{Y|X,\hat{S}}) \geq 2\right\}.$$

Then, by Theorem 3.2, $p_{Y|X}$ can be expressed as a convex combination of deterministic channels of rank ≥ 2 if $g(p_{Y|X}) \leq |\mathcal{X}| - 1$, in which case, $\underline{\Gamma}(1, p_{Y|X}) = 0$. Otherwise, let the y_l-column be the one such that $g(p_{Y|X}) \geq |\mathcal{X}| - 1$. Consider the convex combination

$$p_{Y|X} = t\mathbf{U}_{y_l} + (1-t)p'_{Y|X},$$

where $\mathbf{U}_{y_l}$ is a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix in which the y_l-column is all one and all the other entries are zeros. It is clear that $p'_{Y|X}$ cannot be a convex combination of deterministic channels of rank ≥ 2 unless the sum of its l-th column is $\leq |\mathcal{X}| - 1$. To this end, we set $t = g(p_{Y|X}) - |\mathcal{X}| + 1$, which is the minimum value required such that $p'_{Y|X}$ can be a convex combination of deterministic channels of rank ≥ 2, and we have

$$\sum_{x\in\mathcal{X}} p'_{Y|X}(y_l|x) = \frac{\sum_{x\in\mathcal{X}} p_{Y|X}(y_l|x) - t|\mathcal{X}|}{1-t} = |\mathcal{X}| - 1$$

and

$$\sum_{x\in\mathcal{X}} p'_{Y|X}(y_j|x) = \frac{\sum_{x\in\mathcal{X}} p_{Y|X}(y_j|x)}{1-t} \leq 1, \quad \text{for } y_j \neq y_l,$$

so that $\underline{\Gamma}(1, p_{Y|X}) = g(p_{Y|X}) - |\mathcal{X}| + 1$.

Note that $p_{Y|X}$ has the following convex decomposition

$$p_{Y|X} = \left(1 - \sum_{y\in\mathcal{Y}} \alpha'(y)\right) p'_{Y|X} + \sum_{y\in\mathcal{Y}} \alpha'(y)\mathbf{U}_y,$$

in which $p'_{Y|X}$ is a valid stochastic matrix iff $\alpha'(y) \leq \alpha(y)$ for all $y \in \mathcal{Y}$. Therefore, $\overline{\Gamma}(1, p_{Y|X}) = \sum_{y\in\mathcal{Y}} \alpha(y)$. □

Proposition 3.9 *If $|\mathcal{X}| \leq |\mathcal{Y}|$, then*

$$\overline{\Gamma}(|\mathcal{X}|, p_{Y|X}) \leq 1 - \beta,$$

where

$$\beta = \max_{y\in\mathcal{Y}} \left(\frac{\left(\sum_{x\in\mathcal{X}} p_{Y|X}(y|x)\right) - 1}{\left|\text{supp}(p_{Y|X}(y|\cdot))\right| - 1} \right)^{+}. \tag{3.5}$$

Furthermore, if $\beta = 0$, then $\overline{\Gamma}(|\mathcal{X}|, p_{Y|X}) = 1$.

If $|\mathcal{X}| \geq |\mathcal{Y}|$*, then*

$$\overline{\Gamma}(|\mathcal{Y}|, p_{Y|X}) \leq h,$$

where

$$h = \min\left\{1, \min_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{Y|X}(y|x)\right\}. \tag{3.6}$$

If $h = 1$*, then* $\overline{\Gamma}(|\mathcal{Y}|, p_{Y|X}) = 1$.

Proof: If $|\mathcal{X}| \leq |\mathcal{Y}|$, then the sum of every column of a deterministic channel of rank $|\mathcal{X}|$ is at most 1, and for every $y \in \mathcal{Y}$, $p_{Y|X}$ allows a convex combination of deterministic channels with the y-column sum at most $\left|\text{supp}(p_{Y|X}(y|\cdot))\right|$. Thus for every $p_{\hat{S}} \in \text{dec}(p_{Y|X})$ and every $y \in \mathcal{Y}$, we have

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} p_{Y|X}(y|x) \\
&\leq \min\left\{1, \left|\text{supp}(p_{Y|X}(y|\cdot))\right|\right\} \Gamma(|\mathcal{X}|, p_{\hat{S}}) + \left|\text{supp}(p_{Y|X}(y|\cdot))\right| (1 - \Gamma(|\mathcal{X}|, p_{\hat{S}})) \\
&= \left|\text{supp}(p_{Y|X}(y|\cdot))\right| - \left(\left|\text{supp}(p_{Y|X}(y|\cdot))\right| - 1\right)^+ \Gamma(|\mathcal{X}|, p_{\hat{S}})
\end{aligned}$$

so that

$$\Gamma(|\mathcal{X}|, p_{\hat{S}}) \leq 1 - \frac{\left(\sum_{x\in\mathcal{X}} p_{Y|X}(y|x)\right) - 1}{\left|\text{supp}(p_{Y|X}(y|\cdot))\right| - 1}$$

for $\left|\text{supp}(p_{Y|X}(y|\cdot))\right| > 1$ and hence $\overline{\Gamma}(|\mathcal{X}|, p_{Y|X}) \leq 1 - \beta$. If $\beta = 0$, which implies that $\sum_{x\in\mathcal{X}} p_{Y|X}(y|x) \leq 1$ for all $y \in \mathcal{Y}$, then $\overline{\Gamma}(|\mathcal{X}|, p_{Y|X}) = 1$ (Theorem 3.2).

If $|\mathcal{X}| \geq |\mathcal{Y}|$, then the sum of every column of a deterministic channel of rank $|\mathcal{Y}|$ is at least 1, so that, for every $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$ and every $y \in \mathcal{Y}$, we have

$$\sum_{x \in \mathcal{X}} p_{Y|X}(y|x) \geq \Gamma(|\mathcal{Y}|, p_{\hat{S}}),$$

and hence $\overline{\Gamma}(|\mathcal{Y}|, p_{Y|X}) \leq h$. If $h = 1$, which implies $\sum_{x\in\mathcal{X}} p_{Y|X}(y|x) \geq 1$ for all $y \in \mathcal{Y}$, then $\overline{\Gamma}(|\mathcal{Y}|, p_{Y|X}) = 1$ (Theorem 3.2). □

Proposition 3.10

$$\begin{aligned}&\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})\\ \leq &\begin{cases} (1-\overline{\Gamma}(1,p_{Y|X}))\log\left(\min\left\{\left(|\mathcal{X}| + |\operatorname{supp}(\mathbf{a})| - \mathbf{a}\mathbf{1}^{\mathsf{T}}\right), |\mathcal{Y}|\right\}\right), & \overline{\Gamma}(1,p_{Y|X}) < 1,\\ 0, & \overline{\Gamma}(1,p_{Y|X}) = 1,\end{cases}\end{aligned}$$

$$\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) \geq 1 - \overline{\Gamma}(1, p_{Y|X}),$$

where

$$\mathbf{a} = \left\lfloor \mathbf{1} p'_{Y|X} \right\rfloor,$$
$$p'_{Y|X} = \frac{p_{Y|X} - \sum\limits_{y\in\mathcal{Y}} \alpha(y)\mathbf{U}_y}{1 - \overline{\Gamma}(1, p_{Y|X})},$$

and $\alpha(y)$ is defined by (3.4).

If $|\mathcal{X}| = 2$ or $|\mathcal{Y}| = 2$, then $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = 1 - \overline{\Gamma}(1, p_{Y|X})$.

Proof: One way to find an upper bound of the lower intrinsic capacity is to find a convex combination of $p_{Y|X}$ as bad as possible, while $\overline{\Gamma}(1, p_{Y|X})$ is achieved. To this end, we can first extract from $p_{Y|X}$ a collection of useless deterministic channels with the total probability $\overline{\Gamma}(1, p_{Y|X})$ according to Proposition 3.8, that is,

$$p_{Y|X} = \sum_{y \in \mathcal{Y}} \alpha(y) \mathbf{U}_y + (1 - \overline{\Gamma}(1, p_{Y|X})) p'_{Y|X}.$$

If $\overline{\Gamma}(1, p_{Y|X}) = 1$, then $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = 0$; otherwise,

$$\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = \left(1 - \overline{\Gamma}(1, p_{Y|X})\right) \underline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X}).$$

It is clear that $p'_{Y|X} \in \mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{x}]$, where $\mathbf{x}$ denotes the row vector with every entry equal $|\mathcal{X}|$. The best deterministic channels in $\mathcal{P}_{Y|X}[\mathbf{a}, \mathbf{x}]$ are those with the number of nonzero columns maximized. The rank of those matrices is

$$\min \left\{ \left(|\mathcal{X}| + |\operatorname{supp}(\mathbf{a})| - \mathbf{a}\mathbf{1}^{\mathsf{T}} \right), |\mathcal{Y}| \right\},$$

so $\underline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X}) \leq \log \left(\min \left\{ \left(|\mathcal{X}| + |\operatorname{supp}(\mathbf{a})| - \mathbf{a}\mathbf{1}^{\mathsf{T}} \right), |\mathcal{Y}| \right\} \right)$ (Theorem 3.2). Thus,

$$\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) \leq \left(1 - \overline{\Gamma}(1, p_{Y|X})\right) \log \left(\min \left\{ \left(|\mathcal{X}| + |\operatorname{supp}(\mathbf{a})| - \mathbf{a}\mathbf{1}^{\mathsf{T}} \right), |\mathcal{Y}| \right\} \right).$$

Let $p_{\hat{S}}$ be the vertex of $\operatorname{dec}(p_{Y|X})$ that attains $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$. Then

$$\begin{aligned} \underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = & \sum_{\hat{s} \in \hat{\mathcal{S}}} p_{\hat{S}}(\hat{s}) \log \operatorname{rank} \left(\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot, \hat{s}) \right) \\ \geq & 1 - \Gamma(1, p_{\hat{S}}) \end{aligned}$$

$$\geq 1 - \overline{\Gamma}(1, p_{Y|X}).$$

Finally, the special case of $|\mathcal{X}| = 2$ or $|\mathcal{Y}| = 2$ can be easily verified. □

Proposition 3.11

$$\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) \geq (1 - \underline{\Gamma}(1, p_{Y|X})) \log \left(|\operatorname{supp}(\mathbf{a})| + \left(|\mathcal{X}| - \sum_{y \in \operatorname{supp}(\mathbf{a})} b_y \right)^+ \right), \tag{3.7}$$

$$\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) \leq (1 - \underline{\Gamma}(1, p_{Y|X})) \log(m-1) + \overline{\Gamma}(m, p_{Y|X}) \log \frac{m}{m-1}, \tag{3.8}$$

where

$$\mathbf{a} = \left\lfloor \mathbf{1} p'_{Y|X} \right\rfloor,$$
$$\mathbf{b} = \left\lceil \mathbf{1} p'_{Y|X} \right\rceil,$$
$$p'_{Y|X} = \frac{p_{Y|X} - \underline{\Gamma}(1, p_{Y|X}) \mathbf{U}_{y_l}}{1 - \underline{\Gamma}(1, p_{Y|X})},$$
$$m = \min \left\{ |\mathcal{X}|, |\mathcal{Y}| \right\},$$

and y_l corresponds to the column of $p_{Y|X}$ such that $\mathbf{1} p_{Y|X}(y_l|\cdot) = g(p_{Y|X})$ with $g(p_{Y|X})$ defined by (3.3)*.*

If $|\mathcal{X}| = 2$ or $|\mathcal{Y}| = 2$, then $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = 1 - \overline{\Gamma}(1, p_{Y|X})$.

If $|\mathcal{X}| \leq |\mathcal{Y}|$ and $\beta = 0$ (see (3.5)*), then $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = \log |\mathcal{X}|$.*

If $|\mathcal{X}| \geq |\mathcal{Y}|$ and $h = 1$ (see (3.6)*), then $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = \log |\mathcal{Y}|$.*

Proof: In order to estimate a lower bound of $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$, we first extract from $p_{Y|X}$ the minimum required useless channels (Proposition 3.8), that is,

$$p_{Y|X} = \underline{\Gamma}(1, p_{Y|X})\mathbf{U}_{y_l} + \left(1 - \underline{\Gamma}(1, p_{Y|X})\right) p'_{Y|X},$$

so that $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) \geq (1 - \underline{\Gamma}(1, p_{Y|X}))\, \overline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X})$.

To find a lower bound of $\overline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X})$, we need to find a convex combination of $p'_{Y|X}$ as good as possible. It is clear that $p'_{Y|X} \in \mathcal{P}[\mathbf{a}, \mathbf{b}]$, so $\overline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X})$ is bounded below by the capacity of the worst deterministic channel in $\mathcal{P}[\mathbf{a}, \mathbf{b}]$ (Theorem 3.2), which are obviously those with the number of nonzero columns minimized. The capacity of such a channel is

$$\log\left(|\operatorname{supp}(\mathbf{a})| + \left(|\mathcal{X}| - \sum_{y \in \operatorname{supp}(\mathbf{a})} b_y\right)^+\right).$$

Since $\overline{\mathrm{IC}}_{\mathrm{ED}}(p'_{Y|X}) \geq \log\left(|\operatorname{supp}(\mathbf{a})| + \left(|\mathcal{X}| - \sum_{y \in \operatorname{supp}(\mathbf{a})} b_y\right)^+\right)$, (3.7) has been proved.

Let $p_{\hat{S}}$ be the vertex of $\operatorname{dec}(p_{Y|X})$ that attains $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$. Then

$$\begin{aligned}
\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) =& \sum_{\hat{s} \in \hat{\mathcal{S}}} p_{\hat{S}}(\hat{s}) \log \operatorname{rank}\left(\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot, \hat{s})\right) \\
\leq& (1 - \Gamma(1, p_{\hat{S}}) - \Gamma(m, p_{\hat{S}})) \log(m-1) + \Gamma(m, p_{\hat{S}}) \log(m) \\
=& (1 - \Gamma(1, p_{\hat{S}})) \log(m-1) + \Gamma(m, p_{\hat{S}}) \log \frac{m}{m-1} \\
\leq& (1 - \underline{\Gamma}(1, p_{Y|X})) \log(m-1) + \overline{\Gamma}(m, p_{Y|X}) \log \frac{m}{m-1}
\end{aligned}$$

where $m = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$. The remaining part of the proof is straightforward. □

3.4.3 $\underline{\mathrm{IC}}_f(p_{Y|X})$ and $\overline{\mathrm{IC}}_f(p_{Y|X})$ for $f = \mathrm{E}, \mathrm{D}$

Although it is more difficult to compute $\underline{\mathrm{IC}}_f(p_{Y|X})$ and $\overline{\mathrm{IC}}_f(p_{Y|X})$ for $f = \mathrm{E}, \mathrm{D}$ in general cases, we can still obtain some useful results for some special cases.

The first case to be considered is a channel with binary output, namely, $p_{Y|X}$ with $|\mathcal{Y}| = 2$.

Proposition 3.12 *If $|\mathcal{Y}| = 2$, then*

$$\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C(p_{Y|X}),$$
$$\overline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C(p^*_{Y|U}),$$

where $p^*_{Y|U} = \begin{pmatrix} 1 & 0 \\ \underline{\Gamma}(1, p_{Y|X}) & 1 - \underline{\Gamma}(1, p_{Y|X}) \end{pmatrix}$ *with $\underline{\Gamma}(1, p_{Y|X})$ given by Proposition 3.8.*

Proof: Since $|\mathcal{Y}| = 2$, we only need to choose two maps from all the $2^{|\mathcal{S}|} = 2^{2^{|\mathcal{X}|}}$ maps of $\hat{\mathcal{P}}_{Y|X}$ into $\mathcal{X}$ for constructing the capacity-achieving distributions. We denote these two maps by $\psi(u_0, \hat{S})$ and $\psi(u_1, \hat{S})$. Similar to the proof of Proposition 3.3, our strategy for choosing $\psi(u_0, \cdot)$ and $\psi(u_1, \cdot)$ is to maximize $p_{Y|U}(y_0|u_0) - p_{Y|U}(y_0|u_1)$, where $p_{Y|U}(y|u) = \sum_{\hat{s} \in \hat{S}} p_{\hat{S}}(\hat{s}) \hat{p}_{Y|X,\hat{S}}(y|\psi(u, \hat{s}), \hat{s})$.

There are only two classes of deterministic channels in $\hat{\mathcal{P}}_{Y|X}$, rank 1 and rank 2. For $\hat{s} \in \left\{\hat{s} \in \hat{S} : \mathrm{rank}\left(\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot, \hat{s})\right) = 1\right\}$, it does not matter to choose the values of $\psi(u_0, \hat{s})$ and $\psi(u_1, \hat{s})$. For $\hat{s} \in \left\{\hat{s} \in \hat{S} : \mathrm{rank}\left(\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot, \hat{s})\right) = 2\right\}$, however, let us choose $\psi(u_0, \hat{s})$ such that $\hat{p}_{Y|X,\hat{S}}(y_0|\psi(u_0, \hat{s}), \hat{s}) = 1$, and choose $\psi(u_1, \hat{s})$ such that

$\hat{p}_{Y|X,\hat{S}}(y_0|\psi(u_1,\hat{s}),\hat{s}) = 0$. Then we have $p_{Y|U}$ as

$$p_{Y|U} = \begin{pmatrix} 1-\lambda_{y_1} & \lambda_{y_1} \\ \lambda_{y_0} & 1-\lambda_{y_0} \end{pmatrix}$$

where

$$\begin{aligned} \lambda_{y_0} &= p_{\hat{S}}\left(\hat{s} : \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) = \mathbf{U}_{y_0}\right), \\ \lambda_{y_1} &= p_{\hat{S}}\left(\hat{s} : \hat{p}_{Y|X,\hat{S}}(\cdot|\cdot,\hat{s}) = \mathbf{U}_{y_1}\right). \end{aligned}$$

By Proposition 3.8, the maximum of $\Gamma(1, p_{\hat{S}}) = \lambda_{y_0} + \lambda_{y_1}$ is $\alpha(y_0) + \alpha(y_1)$ with each $\alpha(y)$ being the maximum of feasible values of λ_y, so that

$$\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C\left(\begin{pmatrix} 1-\alpha(y_1) & \alpha(y_1) \\ \alpha(y_0) & 1-\alpha(y_0) \end{pmatrix}\right).$$

Observing that these two rows are exactly two rows of $p_{Y|X}$, we further have $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C(p_{Y|X})$.

Again by Proposition 3.8, $\underline{\Gamma}(1, p_{Y|X}) = \left(g(p_{Y|X}) - |\mathcal{X}| + 1\right)^+$. Without loss of generality, we suppose $\sum_{x\in\mathcal{X}} p_{Y|X}(y_0|x) = g(p_{Y|X})$. Then the minima of feasible values of λ_{y_0} and λ_{y_1} are $\underline{\Gamma}(1, p_{Y|X})$ and 0, respectively. Therefore,

$$\overline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C\left(\begin{pmatrix} 1 & 0 \\ \underline{\Gamma}(1,p_{Y|X}) & 1-\underline{\Gamma}(1,p_{Y|X}) \end{pmatrix}\right).$$

□

The phenomenon $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = (p_{Y|X})$ implies that in some cases, the capacity of channel $p_{Y|X}$ cannot be increased, even if its exact mechanism is known at the encoder. The following result shows that it is not a special case and that a class of general channels with causal state information also has such a property.

Theorem 3.3 *Let $p_{Y|X} = p_{B|X}p_{Y|B}$, where $p_{B|X}$ is a channel with binary output and $p_{Y|B}$ is a channel with binary input in which $|\mathcal{B}| = 2$. Suppose*

$$p_{B|X} = \sum_{s\in\mathcal{S}} p_S(s)p_{B|X,S}(\cdot|\cdot, s)$$

where s denotes the state of channel. The capacity of $p_{Y|X}$ cannot be increased by the causal state information s available at the encoder iff all $p_{B|X,S}(\cdot|\cdot, s)$ are (x_i, x_j)-ended for some fixed $x_i, x_j \in \mathcal{X}$ for all $s \in \mathcal{S}$, where a binary output channel $p_{B|X,S}(\cdot|\cdot, s)$ is said to be (x_i, x_j)-ended if

$$\begin{aligned} & p_{B|X,S}(b_0|x_i, s) = \min_{x\in\mathcal{X}} p_{B|X,S}(b_0|x, s) \\ \textit{and} \qquad & p_{B|X,S}(b_0|x_j, s) = \max_{x\in\mathcal{X}} p_{B|X,S}(b_0|x, s). \end{aligned}$$

In other words, all row vectors of $p_{B|X,S}(\cdot|\cdot, s)$ are contained in the line segment between endpoints $p_{B|X,S}(\cdot|x_i, s)$ and $p_{B|X,S}(\cdot|x_j, s)$.

Proof: (Sufficiency) By [Gamal and Kim 2011, Th. 7.2], we consider the channel

$$p_{Y|U} = p_{B|U}p_{Y|B}$$

where the input alphabet is $\mathcal{U}$ with $|\mathcal{U}| = |\mathcal{X}|^{|\mathcal{S}|}$ and

$$p_{B|U} = \sum_{s\in\mathcal{S}} p_S(s)p_{B|X,S}(\cdot|\psi(u,s),s).$$

in which $\psi(u,s)$ maps from $\mathcal{S}$ to $\mathcal{X}$.

Because every channel $p_{B|X,S}(b_0|x,s)$ is (x_i,x_j)-ended, it is easy to show that $p_{B|U}$ is also (u_i,u_j)-ended, where u_i,u_j are regarded as two constant maps such that $\psi(u_i,s) = x_i$ and $\psi(u_i,s) = x_j$ for all $s \in \mathcal{S}$. Then every row vector of $p_{Y|U}$ is contained in the line segment between $p_{B|U}(\cdot|u_i)p_{Y|B}$ and $p_{B|U}(\cdot|u_j)p_{Y|B}$, hence $p_{Y|U}$ has a capacity-achieving input probability distribution supported on $\{u_i,u_j\}$ (Proposition B.7), and therefore the capacity of $p_{Y|X}$ cannot be increased by the causal state information at the encoder.

(Necessity) If the capacity of $p_{Y|X}$ cannot be increased by its causal state information at the encoder, then a capacity-achieving input probability distribution of $p_{Y|U}$ must have a support, say $\{u_i,u_j\}$, so that for every map $\psi(u,s)$, the vector

$$\begin{aligned} p_{Y|U}(\cdot|u) =& p_{B|U}(\cdot|u)p_{Y|B} \\ =& \left(\sum_{s\in\mathcal{S}} p_S(s)p_{B|X,S}(\cdot|\psi(u,s),s)\right)p_{Y|B} \end{aligned}$$

is contained in the line segment between $p_{Y|U}(\cdot|u_i)$ and $p_{Y|U}(\cdot|u_j)$ (Proposition B.8), where u_i and u_j are understood as two constant maps such that $\psi(u_i,s) = x_i$ and $\psi(u_i,s) = x_j$ for all $s \in \mathcal{S}$.

Without loss of generality, we assume $p_{B|U}(b_0|u_i) \le p_{B|U}(b_0|u_j)$. Then, for any $s' \in \mathcal{S}$ and any $x' \in \mathcal{X}$, we can take $\psi(u',s') = x'$ and $\psi(u',s) = x_i$ for any $s \ne s'$.

Since

$$\begin{aligned}0 &\geq p_{B|U}(b_0|u') - p_{B|U}(b_0|u_i)\\ &= \sum_{s\in\mathcal{S}} p_S(s)p_{B|X,S}(b_0|\psi(u',s),s) - \sum_{s\in\mathcal{S}} p_S(s)p_{B|X,S}(b_0|\psi(u,s),s)\\ &= p_S(s')p_{B|X,S}(b_0|\psi(u',s'),s') - p_S(s')p_{B|X,S}(b_0|\psi(u,s'),s')\\ &= p_S(s')\left(p_{B|X,S}(b_0|x',s') - p_{B|X,S}(b_0|x_i,s')\right)\end{aligned}$$

we have $p_{B|X,S}(b_0|x',s') \geq p_{B|X,S}(b_0|x_i,s')$ for any $s' \in \mathcal{S}$ and any $x' \in \mathcal{X}$. Similarly, we have $p_{B|X,S}(b_0|x',s') \leq p_{B|X,S}(b_0|x_j,s')$. Therefore, every $p_{B|X,S}(\cdot|\cdot,s)$ is (x_i,x_j)-ended. □

The second case to be considered is a channel with binary input, namely, $p_{Y|X}$ with $|\mathcal{X}| = 2$.

Proposition 3.13 *If $|\mathcal{X}| = 2$, then for every $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$, $\mathrm{IC}_{\mathrm{D}}(p_{\hat{S}}) = \mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}})$, so that $\underline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) = 1 - \overline{\Gamma}(1, p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) = 1 - \underline{\Gamma}(1, p_{Y|X})$.*

Proof: Because $|\mathcal{X}| = 2$, the binary uniform distribution is capacity-achieving for every deterministic channel, rank 1 or rank 2. Thus we have $\mathrm{IC}_{\mathrm{D}}(p_{\hat{S}}) = \mathrm{IC}_{\mathrm{ED}}(p_{\hat{S}})$ for every $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$. The remaining part is an easy consequence of Propositions 3.10 and 3.11. □

In the above two special cases, we notice that $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C(p_{Y|X})$ for $|\mathcal{Y}| = 2$ and $\overline{\mathrm{IC}}_{\mathrm{D}}(W) = \overline{\mathrm{IC}}(W)$ for $\mathcal{X} = 2$. However they are not true in general.

Example 3.1 *For*

$$p_{Y|X} = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.6 & 0.35 & 0.05 \end{pmatrix},$$

$\underline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) > C(p_{Y|X})$.

Proposition 3.14 *Let $p_{Y|X}$ be a ternary-input-binary-output channel. If all probabilities $p_{Y|X}(y|x)$ are distinct and the sum of each column of $p_{Y|X}$ is greater than or equal to* 1*, then $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) < \overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$.*

The proof of Proposition 3.14 is given in Appendix B.3.

3.5 Conclusion

We have shown that the intrinsic capacity of a channel can be any value between the lower and the upper intrinsic capacities. So, to some extent, the lower and the upper intrinsic capacities are important properties of a channel, reflecting the freedom of the underlying structure of the channel from an information-theoretic perspective.

There are three options for the availability of the causal state information, so to each option there corresponds a pair of lower and upper intrinsic capacities, which we denote by $\underline{\mathrm{IC}}_f(p_{Y|X})$ and $\overline{\mathrm{IC}}_f(p_{Y|X})$ for $f = \mathrm{E}, \mathrm{D}, \mathrm{ED}$, respectively. We determined their values in almost all cases when the input or the output are binary. Two exceptions are the binary-input nonbinary-output channels for $f = \mathrm{E}$ and the nonbinary-input-binary-output channels for $f = \mathrm{D}$. Example 3.1 and Proposition 3.14 show that these two cases are not as simple as other binary cases.

Our main approach involves determining the lower and the upper rank probabilities (Propositions 3.8 and 3.9). However, it is still unknown that, in general, the greedy strategy to put probability mass as much as possible on channels with priorities from the lowest rank to the highest or from the highest to the lowest will necessarily lead us to the lower or the upper intrinsic capacities. Anyway, this approach is useful for estimating the lower and the upper intrinsic capacities (Propositions 3.10 and 3.11).

This work may not be very useful for the real-world communications because it is usually difficult to get the full knowledge of a channel such that it degenerates to a deterministic channel. Storage applications may be one of the exceptions.

On the contrary, for information theory, this work is very important. In a coding system, any dependence can be modeled by a random map or a channel, and if using the terms in game theory, they are called a mixed strategy and a behavioral strategy, respectively. The relation between a random map and a channel is many-to-one, and note that a random map is nothing but a convex combination of deterministic channels. The problem studied in this paper is a bridge connecting these two objects. The convex-nature results of this topic, combined with traditional tools of information theory (e.g., the log-sum inequality), can provide powerful approaches to many information-theoretic problems.

Chapter 4

Conclusions and Future Work

This thesis focused on the capacity analysis for channels with state. Particularly, based on the availability of the state information at the encoder and/or decoder, the capacity-achieving coding schemes are different.

In Chapter 2, we studied the channel model when the decoder-side state information is always perfect and only a noisy version of state information is available at the encoder side. We have shown that when a fairly small amount of noise is added upon the encoder-side state information, the capacity of binary-input channels is as low as if the noisy state information is totally useless. On the contrary, the generalized probing capacity is as high as if the encoder has the knowledge of the perfect state information, even when a fairly large amount of noise is added to the encoder-side state information. Two explicit thresholds about the quality of the encoder-side state information have been derived in terms of $I(S;\tilde{S})$ and $H(S|\tilde{S})$. Extensions have also been made to non-binary-input cases and to the model when the state information is not available at the decoder. The necessary and sufficient conditions have been given on when such thresholds exist for the extensions. Considering that binary signalling is widely used, especially in wideband communications, our work might have some practical relevance. However, our results rely on the assumption that the channel has finite state which is not always true in reality; moreover, the freedom of power control in real communication systems is considered in our research. Thus, it will be interesting to take into account of either an infinite size of channel state or the freedom of power control. The potential counterpart result in source coding is also worthy to be investigated.

In Chapter 3, we introduced the idea of intrinsic capacity which can be seen as the potential of increasing the channel capacity by utilizing the channel state information. In pursuing a better understanding of intrinsic capacity, we first studied the structure of the convex polytope $\mathrm{dec}(p_{Y|X})$ consisting of all convex combinations of deterministic channels for channel $p_{Y|X}$. We proved that, except for $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X})$, all the other $\underline{\mathrm{IC}}_f(p_{Y|X})$ and $\overline{\mathrm{IC}}_f(p_{Y|X})$ are attained at some vertex of $\mathrm{dec}(p_{Y|X})$. Necessary and sufficient conditions for a vertex of $\mathrm{dec}(p_{Y|X})$, as well as a series of consequences, are also provided. Then, we proved a generalization of the Birkhoff-von Neumann Theorem for a family $\mathcal{P}_{Y|X}[\mathbf{a},\mathbf{b}]$ of channel matrices with integer-valued column-sum vector constraints $\mathbf{a}$ and $\mathbf{b}$ from below and above, respectively. It has been shown that $\mathcal{P}_{Y|X}[\mathbf{a},\mathbf{b}]$ is convex and its vertices are exactly all deterministic channels in $\mathcal{P}_{Y|X}[\mathbf{a},\mathbf{b}]$. Using this fundamental result, we have determined the exact values of $\underline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ when the input or the output is binary. General lower and upper bounds are further provided for the nonbinary cases and in some cases, the exact value of $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$ has also been determined. For a binary-output channel $p_{Y|X}$, we have derived the exact values of $\underline{\mathrm{IC}}_{\mathrm{E}}((p_{Y|X})$ and $\overline{\mathrm{IC}}_{\mathrm{E}}((p_{Y|X})$, and for a binary-input channel $p_{Y|X}$, the exact values of $\underline{\mathrm{IC}}_{01}(W)$ and $\overline{\mathrm{IC}}_{01}(W)$ have also been obtained.

Finally, an interesting phenomenon observed is that $\underline{\mathrm{IC}}_{\mathrm{E}}(p_{Y|X}) = C(p_{Y|X})$ for binary-output channels. In other words, the causal state information at the encoder is useless. We further proved that a class of general channels with causal state information available at the encoder also has such a property. In the future, it will be worthy to investigate if the greedy strategy to put probability mass as much as possible on channels with priorities from the lowest rank to the highest or from the highest to the lowest will necessarily lead us to the lower or the upper intrinsic capacities. Besides that, more research is required to derive the exact values of lower and upper intrinsic capacities for non-binary channels.

Appendix A

Proofs for Chapter 2

A.1 An Alternative Proof of Theorem 2.2

We shall show that, for any binary-input channel $p_{Y|X,S}$, state distribution p_S, and side channel $p_{\tilde{S}|S}$,

$$C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \overline{C}(p_{Y|X,S}, p_S)$$

if

$$H(S|\tilde{S}) \leq \frac{4\rho \log 2}{3 + 2(e-1)\sqrt{2|\mathcal{S}|}}. \tag{A.1}$$

Lemma A.1 *$p_{\hat{X}|S}$ is a stochastically degraded version of $p_{\tilde{S}|S}$ if*

$$H(S|\tilde{S}) \leq \frac{4\tau\rho \log 2}{3\tau + 2\sqrt{2|\mathcal{S}|}}, \tag{A.2}$$

where

$$\tau = \min_{x \in \mathcal{X}_+} \frac{\min_{s \in \mathcal{S}} p_{\hat{X}|S}(x|s)}{\max_{s \in \mathcal{S}} p_{\hat{X}|S}(x|s)}.$$

Proof: Let $\hat{S}$ denote the maximum likelihood estimate of S based on $\tilde{S}$. It suffices to show that $p_{\hat{S}|S}$ is invertible and $p_{\hat{S}|S}^{-1} p_{\hat{X}|S}$ is a valid probability transition matrix if (A.2) is satisfied.

Table A.1: Specification of $\psi(\cdot,\cdot)$ for $\mathcal{U} = \{0, 1, \cdots, 7\}$ and $\tilde{\mathcal{S}} = \{0, 1, *\}$

$\psi(u, \tilde{s})$	$\tilde{s} = 0$	$\tilde{s} = 1$	$\tilde{s} = *$
$u = 0$	0	0	0
$u = 1$	1	1	1
$u = 2$	1	1	0
$u = 3$	0	0	1
$u = 4$	0	1	0
$u = 5$	0	1	1
$u = 6$	1	0	0
$u = 7$	1	0	1

Let $\sigma_{\min}(p_{\hat{S}|S})$ denote the smallest singular value of $p_{\hat{S}|S}$. It follows from [Johnson 1989, Th. 3] that

$$\sigma_{\min}(p_{\hat{S}|S}) \geq \min_{s \in \mathcal{S}} \frac{1}{2}\left(2p_{\hat{S}|S}(s|s) - \sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(\hat{s}|s) - \sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(s|\hat{s})\right). \tag{A.3}$$

Clearly,

$$\begin{aligned} &\min_{s \in \mathcal{S}} \frac{1}{2}\left(2p_{\hat{S}|S}(s|s) - \sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(\hat{s}|s) - \sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(s|\hat{s})\right) \\ &= \min_{s \in \mathcal{S}} \frac{1}{2}\left(2 - 3\sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(\hat{s}|s) - \sum_{\hat{s} \in \mathcal{S}: \hat{s} \neq s} p_{\hat{S}|S}(s|\hat{s})\right) \end{aligned}$$

$$\geq 1 - \frac{3}{2} \sum_{s, \hat{s} \in \mathcal{S} : s \neq \hat{s}} p_{\hat{S}|S}(\hat{s}|s). \tag{A.4}$$

Substituting (A.4) into (A.3) and invoking (2.45) gives

$$\sigma_{\min}(p_{\hat{S}|S}) \geq 1 - \frac{3H(S|\tilde{S})}{4\rho \log 2}. \tag{A.5}$$

Therefore, $p_{\hat{S}|S}$ is invertible if $H(S|\tilde{S}) < \frac{4\rho \log 2}{3}$. Let $\|\cdot\|_\infty$, $\|\cdot\|_2$, and $\|\cdot\|_F$ denote the maximum row sum matrix norm, the spectral norm, and the Frobenius norm, respectively Horn and Johnson (1985). Note that

$$\begin{aligned}
&\|p^{-1}_{\hat{S}|S} - \mathrm{diag}(1, \cdots, 1)\|_\infty \\
&\leq \sqrt{|\mathcal{S}|}\|p^{-1}_{\hat{S}|S} - \mathrm{diag}(1, \cdots, 1)\|_2 \\
&\leq \sqrt{|\mathcal{S}|}\|p^{-1}_{\hat{S}|S}\|_2 \|p_{\hat{S}|S} - \mathrm{diag}(1, \cdots, 1)\|_2 \qquad \text{(A.6)} \\
&\leq \sqrt{|\mathcal{S}|}\|p^{-1}_{\hat{S}|S}\|_2 \|p_{\hat{S}|S} - \mathrm{diag}(1, \cdots, 1)\|_F, \qquad \text{(A.7)}
\end{aligned}$$

where (A.6) follows by the sub-multiplicative property of the spectral norm. We have

$$\begin{aligned}
\|p^{-1}_{\hat{S}|S}\|_2 &= \frac{1}{\sigma_{\min}(p_{\hat{S}|S})} \\
&\leq \left(1 - \frac{3H(S|\tilde{S})}{4\rho \log 2}\right)^{-1}, \qquad \text{(A.8)}
\end{aligned}$$

where (A.8) is due to (A.5). For $p_{\hat{S}|S} - \mathrm{diag}(1, \cdots, 1)$, it is clear that the diagonal entries are non-positive, the off-diagonal entries are non-negative, and the sum of all entries is equal to 0; moreover, the sum of its off-diagonal entries is bounded above by $\frac{H(S|\tilde{S})}{2\rho\log 2}$ (see (2.45)). Therefore,

$$\begin{aligned}
&\|p_{\hat{S}|S} - \mathrm{diag}(1,\cdots,1)\|_F \\
&= \sqrt{\sum_{s\in\mathcal{S}}(p_{\hat{S}|S}(s|s)-1)^2 + \sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}}(p_{\hat{S}|S}(\hat{s}|s))^2} \\
&\leq \sqrt{\bigg(\sum_{s\in\mathcal{S}}(p_{\hat{S}|S}(s|s)-1)\bigg)^2 + \bigg(\sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_{\hat{S}|S}(\hat{s}|s)\bigg)^2} \\
&= \sqrt{2\bigg(\sum_{s,\hat{s}\in\mathcal{S}:s\neq\hat{s}} p_{\hat{S}|S}(\hat{s}|s)\bigg)^2} \\
&\leq \frac{H(S|\tilde{S})}{\sqrt{2}\rho\log 2}. \qquad \text{(A.9)}
\end{aligned}$$

Substituting (A.8) and (A.9) into (A.7) yields

$$\|p^{-1}_{\hat{S}|S} - \mathrm{diag}(1,\cdots,1)\|_\infty \leq \frac{\sqrt{|\mathcal{S}|}H(S|\tilde{S})}{\sqrt{2}\rho\log 2}\bigg(1-\frac{3H(S|\tilde{S})}{4\rho\log 2}\bigg)^{-1}. \qquad \text{(A.10)}$$

To ensure that all entries of $p^{-1}_{\hat{S}|S}p_{\hat{X}|S}$ are non-negative (or equivalently $(\mathrm{diag}(1,\cdots,1)-p^{-1}_{\hat{S}|S})p_{\hat{X}|S}$ is component-wise dominated by $p_{\hat{X}|S}$), it suffices to have

$$\|p^{-1}_{\hat{S}|S} - \mathrm{diag}(1,\cdots,1)\|_\infty \leq \tau. \qquad \text{(A.11)}$$

Combining (A.10) and (A.11) shows that $p^{-1}_{\hat{S}|S}p_{\hat{X}|S}$ is a valid probability transition matrix[1] if (A.2) is satisfied[2]. □

[1] The requirement that the entries in each row of $p^{-1}_{\hat{S}|S}p_{\hat{X}|S}$ add up to 1 is automatically satisfied.

[2] Note that (A.2) implies $H(S|\tilde{S}) < \frac{4\rho\log 2}{3}$, which further implies the existence of $p^{-1}_{\hat{S}|S}$.

Since $|\mathcal{X}| = 2$, it follows from (Shulman and Feder, 2004, Th. 2) that there exists $p_{\hat{X}|S} \in \mathcal{P}$ satisfying

$$p_{\hat{X}|S}(x|s) > e^{-1}, \quad x \in \mathcal{X}, s \in \mathcal{S}.$$

For such $p_{\hat{X}|S}$, we have

$$\tau \geq \frac{1}{e-1}.$$

Invoking Lemma A.1 shows that $p_{\hat{X}|S}$ is a stochastically degraded version of $p_{\tilde{S}|S}$ (and consequently $C'(p_{Y|X,S}, p_S, p_{\tilde{S}|S}) = \overline{C}(p_{Y|X,S}, p_S)$) if (A.1) is satisfied.

A.2 Proof of (2.66) and (2.67)

Table A.2: Specification of $\psi(\cdot,\cdot)$ for $\mathcal{U} = \{0, 1, \cdots, 3\}$ and $\tilde{\mathcal{S}} = \{0, 1, *\}$

$\psi(u, \tilde{s})$	$\tilde{s} = 0$	$\tilde{s} = 1$
$u = 0$	0	0
$u = 1$	1	1
$u = 2$	0	1
$u = 3$	1	0

Lemma A.2 *For* $\theta \in (0, 1)$,

$$\eta(\theta) \triangleq (1-\theta)\log(1+\theta) + \theta \log \theta < 0.$$

Proof: We have

$$\begin{aligned}\frac{\mathrm{d}^2\eta(\theta)}{\mathrm{d}\theta^2} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(-\log(1+\theta)+\frac{1-\theta}{1+\theta}+\log\theta+1\right)\\ &= -\frac{1}{1+\theta}-\frac{2}{(1+\theta)^2}+\frac{1}{\theta}\\ &= \frac{1-\theta}{\theta(1+\theta)^2}\\ &> 0, \quad \theta\in(0,1),\end{aligned}$$

which, together with the fact $\eta(0)=\eta(1)=0$, implies the desired result. □

When $\theta = 0$ or $\theta = 1$, we have $\underline{C}(p_{Y|X,S}, p_S) = \overline{C}(p_{Y|X,S}, p_S)$, which implies $\underline{\epsilon}(p_{Y|X,S}, p_S) = \underline{q}(p_{Y|X,S}, p_S) = 0$. When $\theta \in (0,1)$, the maximizer of the optimization problem in (2.4), denoted by $p_{\hat{X}}$, is unique and is given by

$$p_{\hat{X}}(0) = p_{\hat{X}}(1) = \frac{1}{2}.$$

Now consider $\psi(\cdot,\cdot)$ specified by Table A.1. It can be verified that

$$\begin{aligned}D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) &= \frac{1}{2}\left((1-\theta)\log 2+\log\frac{2}{1+\theta}+\theta\log\frac{2\theta}{1+\theta}\right), \quad u=0,1,\\ D_{\mathrm{GE}}(p_{\hat{U}},\epsilon,u) &= \frac{1}{2}\left(\epsilon(1-\theta)\log 2\epsilon+(\theta+\epsilon(1-\theta))\log\frac{2(\theta+\epsilon(1-\theta))}{1+\theta}\right.\\ &\quad+(1-\epsilon(1-\theta))\log\frac{2(1-\epsilon(1-\theta))}{1+\theta}\\ &\quad\left.+(1-\epsilon)(1-\theta)\log 2(1-\epsilon)\right), \quad u=2,3,\end{aligned}$$

$$\begin{aligned}
D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) &= \frac{1}{2}\Bigg((1-\theta)\log 2 + (\theta + \epsilon(1-\theta))\log\frac{2(\theta+\epsilon(1-\theta))}{1+\theta} \\
&\quad + \theta \log \frac{2\theta}{1+\theta} + (1-\epsilon)(1-\theta)\log 2(1-\epsilon)\Bigg), \quad u = 4, 5, \\
D_{\mathrm{GE}}(p_{\hat{U}}, \epsilon, u) &= \frac{1}{2}\Bigg(\epsilon(1-\theta)\log 2\epsilon + \log\frac{2}{1+\theta} \\
&\quad + (1-\epsilon(1-\theta))\log\frac{2(1-\epsilon(1-\theta))}{1+\theta}\Bigg), \quad u = 6, 7.
\end{aligned}$$

Moreover,

$$\begin{aligned}
D_{\mathrm{GE}}(p_{\hat{U}}, 0, u) &= \frac{1}{2}\Bigg((1-\theta)\log 2 + \log\frac{2}{1+\theta} + \theta\log\frac{2\theta}{1+\theta}\Bigg) \\
&= \underline{C}(p_{Y|X,S}, p_S), \quad u = 0, 1, 2, 3, \\
D_{\mathrm{GE}}(p_{\hat{U}}, 0, u) &= (1-\theta)\log 2 + \theta\log\frac{2\theta}{1+\theta} \\
&< \underline{C}(p_{Y|X,S}, p_S), \quad u = 4, 5, \qquad \text{(A.12)} \\
D_{\mathrm{GE}}(p_{\hat{U}}, 0, u) &= \log\frac{2}{1+\theta} \\
&> \underline{C}(p_{Y|X,S}, p_S), \quad u = 6, 7, \qquad \text{(A.13)}
\end{aligned}$$

where (A.12) and (A.13) follow from Lemma A.2. Therefore, we have

$$\begin{aligned}
\epsilon(u) &= 0, \quad u = 0, 1, 2, 3, 4, 5, \\
\epsilon(u) &= \hat{\epsilon}(\theta), \quad u = 6, 7,
\end{aligned}$$

which, together with (2.64), proves (2.66) for $\theta \in (0,1)$. Next consider $\psi(\cdot,\cdot)$ specified by Table A.2. It can be verified that

$$\begin{aligned}
D_{\mathrm{GS}}(p_{\hat{U}}, q, u) &= \frac{1}{2}\left((1-\theta)\log 2 + \log\frac{2}{1+\theta} + \theta\log\frac{2\theta}{1+\theta}\right), \quad u = 0,1,\\
D_{\mathrm{GS}}(p_{\hat{U}}, q, 2) &= (1-q)(1-\theta)\log 2(1-q) + (\theta + q(1-\theta))\log\frac{2(\theta+q(1-\theta))}{1+\theta},\\
D_{\mathrm{GS}}(p_{\hat{U}}, q, 3) &= q(1-\theta)\log 2q + (1-q(1-\theta))\log\frac{2(1-q(1-\theta))}{1+\theta}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
D_{\mathrm{GS}}(p_{\hat{U}}, 0, u) &= \frac{1}{2}\left((1-\theta)\log 2 + \log\frac{2}{1+\theta} + \theta\log\frac{2\theta}{1+\theta}\right)\\
&= \underline{C}(p_{Y|X,S}, p_S), \quad u = 0,1,\\
D_{\mathrm{GS}}(p_{\hat{U}}, 0, 2) &= (1-\theta)\log 2 + \theta\log\frac{2\theta}{1+\theta}\\
&< \underline{C}(p_{Y|X,S}, p_S), \qquad \text{(A.14)}\\
D_{\mathrm{GS}}(p_{\hat{U}}, 0, 3) &= \log\frac{2}{1+\theta}\\
&> \underline{C}(p_{Y|X,S}, p_S), \qquad \text{(A.15)}
\end{aligned}$$

where (A.14) and (A.15) follow from Lemma A.2. Therefore, we have

$$\begin{aligned}
q(u) &= 0, \quad u = 0,1,2,\\
q(3) &= \hat{q}(\theta),
\end{aligned}$$

which, together with (2.65), proves (2.67) for $\theta \in (0,1)$.

A.3 Proof of (2.73) and (2.74)

When $\theta = 0$ or $\theta = 1$, we have $\underline{C}(p_{Y|X,S}, p_S) = \overline{C}(p_{Y|X,S}, p_S)$, which implies $\overline{\epsilon}(p_{Y|X,S}, p_S) = 1$ and $\overline{q}(p_{Y|X,S}, p_S) = \frac{1}{2}$. When $\theta \in (0,1)$, the maximizer of the optimization problem in (2.5), denoted by $p_{\hat{X}|S}$, is unique and is given by

$$\begin{aligned} &p_{\hat{X}|S}(x|s) \\ &= \begin{cases} \left(1 + (1-\theta)\theta^{\frac{\theta}{1-\theta}}\right)^{-1} \theta^{\frac{\theta}{1-\theta}}, & x = s, \\ \left(1 + (1-\theta)\theta^{\frac{\theta}{1-\theta}}\right)^{-1} \left(1 - \theta^{\frac{1}{1-\theta}}\right), & \text{otherwise.} \end{cases} \end{aligned}$$

In view of (2.71) and (2.72), it suffices to show that

$$\theta^{\frac{\theta}{1-\theta}} < 1 - \theta^{\frac{1}{1-\theta}}, \quad \theta \in (0,1).$$

Indeed, for $\theta \in (0,1)$,

$$\begin{aligned} &\theta^{\frac{\theta}{1-\theta}} < 1 - \theta^{\frac{1}{1-\theta}} \\ &\Leftrightarrow 1 < \theta^{-\frac{\theta}{1-\theta}} - \theta \\ &\Leftrightarrow (1-\theta)\log(1+\theta) + \theta \log \theta < 0, \end{aligned}$$

and the last inequality is true according to Lemma A.2.

Appendix B

Proofs for Chapter 3

B.1 Continuity of Mutual Information

Proposition B.1 *(Zhang, 2007, Th. 2) For $p_X, p'_X \in \mathcal{P}_X$ and $p_{Y|X}, p'_{Y|X} \in \mathcal{P}_{Y|X}$,*

$$\left|\tilde{I}(p_X, p_{Y|X}) - \tilde{I}(p'_X, p'_{Y|X})\right| \leq 3\delta \log(|\mathcal{X}||\mathcal{Y}| - 1) + 3H_B(\delta).$$

where $\delta = \mathrm{d}(\mathrm{diag}(p_X)p_{Y|X}, \mathrm{diag}(p'_X)p'_{Y|X})$ and $H_B(\cdot)$ is the binary entropy.

Proposition B.2 *(Yassaee* et al.*, 2014, cf. Lemma 3) For $p_X, p'_X \in \mathcal{P}_X$ and $p_{Y|X} \in \mathcal{P}_{Y|X}$,*

$$\mathrm{d}(\mathrm{diag}(p_X)p_{Y|X}, \mathrm{diag}(p'_X)p_{Y|X}) = \mathrm{d}(p_X, p'_X)$$

and

$$\mathrm{d}(p_X p_{Y|X}, p'_X p_{Y|X}) \leq \mathrm{d}(p_X, p'_X).$$

Proposition B.3 *For $p_X \in \mathcal{P}_X$ and $p_{Y|X}, p'_{Y|X} \in \mathcal{P}_{Y|X}$,*

$$\mathrm{d}(\mathrm{diag}(p_X)p_{Y|X}, \mathrm{diag}(p_X)p'_{Y|X}) \leq \mathrm{d}(p_{Y|X}, p'_{Y|X}).$$

Proof:

$$\begin{aligned}\mathrm{d}(\mathrm{diag}(p_X)p_{Y|X}, \mathrm{diag}(p_X)p'_{Y|X}) &= \frac{1}{2}\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} |p_X(x)p_{Y|X}(y|x) - p_X(x)p'_{Y|X}(y|x)| \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{Y}} |p_{Y|X}(y|x) - p'_{Y|X}(y|x)| \\ &= \sum_{x\in\mathcal{X}} p_X(x)\, \mathrm{d}(p_{Y|X}(\cdot|x), p'_{Y|X}(\cdot|x)) \\ &\leq \mathrm{d}(p_{Y|X}, p'_{Y|X})\end{aligned}$$

□

Proposition B.4 *(Yassaee* et al.*, 2014, cf. Lemma 3) For $p_X, p'_X \in \mathcal{P}_X$ and $p_{Y|X}, p'_{Y|X} \in \mathcal{P}_{Y|X}$,*

$$\begin{aligned}\mathrm{d}(\mathrm{diag}(p_X)p_{Y|X}, \mathrm{diag}(p'_X)p'_{Y|X}) &\leq \mathrm{d}(p_X, p'_X) + \mathrm{d}(p_{Y|X}, p'_{Y|X}) \\ &\leq 2\,\mathrm{d}\left((p_X, p_{Y|X}), (p'_X, p'_{Y|X})\right),\end{aligned}$$

so that $\tilde{I}(p_X, p_{Y|X})$ is uniformly continuous on $(\mathcal{P}_X \times \mathcal{P}_{Y|X})$.

Proof: This proposition is a direct result from the triangle inequality and Propositions B.1–B.3. □

Proposition B.5 *Let $g(\cdot)$ be a map from $\mathcal{P}_{\hat{S}}$ to $\mathcal{P}_{Y|X}$. If $g(\cdot)$ is uniformly continuous, then $\tilde{I}(p_X, g(\mathcal{P}_{\hat{S}}))$ is uniformly continuous on $(\mathcal{P}_X \times \mathcal{P}_{\hat{S}})$, where $p_X \in \mathcal{P}_X$ and $p_{\hat{S}} \in \mathcal{P}_{\hat{S}}$.*

Proof: This proposition is a direct result from B.4. □

Proposition B.6 *If $\tilde{I}(p_X, p_{Y|X})$ is uniformly continuous on $(\mathcal{P}_X \times \mathcal{P}_{Y|X})$, then $C(p_{Y|X}) = \max_{p_X} \tilde{I}(p_X, p_{Y|X})$ is uniformly continuous.*

Proof: Since $\tilde{I}(p_X, p_{Y|X})$ is uniformly continuous, for any $\epsilon > 0$, there is a $\delta > 0$ such that for any $p_{Y|X}, p'_{Y|X} \in \mathcal{P}_{Y|X}$ and any $p_X \in \mathcal{P}_X$,

$$\mathrm{d}\left((p_X, p_{Y|X}), (p_X, p'_{Y|X})\right) < \delta \quad \Rightarrow \quad \left|\tilde{I}(p_X, p_{Y|X}) - \tilde{I}(p_X, p'_{Y|X})\right| < \epsilon$$

In other words, for any $p_X \in \mathcal{P}_X$,

$$\mathrm{d}\left(p_{Y|X}, p'_{Y|X})\right) < \delta \quad \Rightarrow \quad \left|\tilde{I}(p_X, p_{Y|X}) - \tilde{I}(p_X, p'_{Y|X})\right| < \epsilon$$

Then

$$\left|\max_{p_X} \tilde{I}(p_X, p_{Y|X}) - \max_{p_X} \tilde{I}(p_X, p'_{Y|X})\right| \leq \max_{p_X} \left|\tilde{I}(p_X, p_{Y|X}) - \tilde{I}(p_X, p'_{Y|X})\right| < \epsilon,$$

so that $C(p_{Y|X})$ is uniformly continuous. □

B.2 Capacity-Achieving Input Probability Distributions

For a channel $p_{Y|X}$, according to (Gallager, 1968, Th. 4.5.1), an input probability distribution p_X^* maximizes the mutual information $I(X;Y)$ iff

$$D\left(p_{Y|X}(\cdot|x)||p_Y\right)\begin{cases} = C & \text{for } x \in \operatorname{supp}(p_X^*) \\ \leq C & \text{for } x \notin \operatorname{supp}(p_X^*) \end{cases}$$

where $p_Y = p_X^* p_{Y|X}$. Based on this sufficient and necessary condition, we have the following results concerning the support of capacity-achieving input probability distributions. In the sequel, we denote by $\operatorname{conv}(p_{Y|X})$ the convex hull of all row vectors in $p_{Y|X}$.

Proposition B.7 *Let $\mathcal{A} \subseteq \mathcal{X}$. If all row vectors of $p_{Y|X}$ are contained in the convex hull, $\operatorname{conv}(\{p_{Y|X}(\cdot|x)\}_{x\in\mathcal{A}})$, then there exists a capacity-achieving probability distribution p_X^* such that $\operatorname{supp}(p_X^*) \subseteq \mathcal{A}$.*

Proof: Let p_A be a capacity-achieving probability distribution of the sub matrix $p_{Y|A}$. Padding p_A with zeros for $x \in \mathcal{X} \setminus \mathcal{A}$, we obtain a probability distribution p_X^* over $\mathcal{X}$. It is clear that

$$D\left(p_{Y|X}(\cdot|x)||p_Y^*\right)\begin{cases} = C & \text{for } x \in \operatorname{supp}(p_X^*) \\ \leq C & \text{for } x \notin \mathcal{A} \setminus \operatorname{supp}(p_X^*) \end{cases}$$

where $p_Y^* = p_X^* p_{Y|X}$. It remains to show that

$$D\left(p_{Y|X}(\cdot|x)||p_Y\right) \leq C \quad \text{for } x \notin \mathcal{A}.$$

Note that all row vectors of $p_{Y|X}$ are contained in the convex hull, $\mathrm{conv}(\{p_{Y|X}(\cdot|x)\}_{x\in\mathcal{A}})$. Thus, there exists some distribution p_A over $\mathcal{A}$, such that

$$p_{Y|X}(\cdot|x) = \sum_{a\in\mathcal{A}} q_A(a) p_{Y|A}(\cdot|a) \quad \text{for each } x \in \mathcal{A}.$$

Since divergence is a convex function on the domain of probability distribution, we have

$$\begin{aligned} D\left(p_{Y|X}(\cdot|x)||p_Y\right) &= D\left(\sum_{a\in\mathcal{A}} q_A(a) p_{Y|A}(\cdot|a)\middle\| p_Y\right) \\ &\le \sum_{a\in\mathcal{A}} q_A(a) D\left(p_{Y|A}(\cdot|a)\|p_Y\right) \\ &\le C \end{aligned}$$

□

Proposition B.8 *Let p_X^* be a capacity-achieving probability distribution of $p_{Y|X}$ and let $\mathcal{A} = \mathrm{supp}(p_X^*)$. For any $x_a \in \mathcal{A}$, $p_{Y|X}(\cdot|x_a) \notin \mathrm{conv}\left(\{p_{Y|X}(\cdot|x)\}_{x\in\mathcal{X}} \setminus \{p_{Y|X}(\cdot|x_a)\}\right)$.*

Proof: It is clear that $D\left(p_{Y|X}(\cdot|x)\|p_Y^*\right) = C$ for all $x \in \mathcal{A}$, where $p_Y^* = p_X^* p_{Y|X}$. We first show that $p_{Y|X}(\cdot|x_a) \notin \mathrm{conv}\left(\{p_{Y|X}(\cdot|x)\}_{x\in\mathcal{A}} \setminus \{p_{Y|X}(\cdot|x_a)\}\right)$. If it is false, then

$$p_{Y|X}(\cdot|x_a) = \sum_{x\in\mathcal{A}'} p'_X(x) p_{Y|X}(\cdot|x)$$

where $\mathcal{A}' = \left\{x \in \mathcal{A} \colon p_{Y|X}(\cdot|x) \neq p_{Y|X}(\cdot|x_a)\right\}$ and p'_X is some strictly postive distribution over $\mathcal{A}'$. It is clear that $p'_X(x) < 1$ for all $x \in \mathcal{A}'$, so that

$$\begin{aligned} D\left(p_{Y|X}(\cdot|x_a)\|p^*_Y\right) =& D\left(\sum_{x\in\mathcal{A}'} p'_X(x)p_{Y|X}(\cdot|x)||p^*_Y\right) \\ <& \sum_{x\in\mathcal{A}'} p'_X(x)D\left(p_{Y|X}(\cdot|x)\|p^*_Y\right) \\ =& C, \end{aligned}$$

a contradiction.

Now suppose that

$$p_{Y|X}(\cdot|x_a) \notin \operatorname{conv}\left(\left\{p_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{A}} \setminus \left\{p_{Y|X}(\cdot|x_a)\right\}\right)$$

Then

$$\begin{aligned} p_{Y|X}(\cdot|x_a) &= \sum_{x\in\mathcal{A}''} p''_X(x)p_{Y|X}(\cdot|x) \\ &= \sum_{x\in\mathcal{A}'} p''_X(x)p_{Y|X}(\cdot|x) + \sum_{x\in\mathcal{A}''\setminus\mathcal{A}'} p''_X(x)p_{Y|X}(\cdot|x) \end{aligned}$$

where $\mathcal{A}'' = \left\{x \in \mathcal{X} \colon p_{Y|X}(\cdot|x) \neq p_{Y|X}(\cdot|x_a)\right\}$ and p''_X is some strictly postive distribution over $\mathcal{A}''$. It is clear that $0 < \sum_{x\in\mathcal{A}''\setminus\mathcal{A}'} p''_X(x) < 1$, and therefore

$$\begin{aligned} C =& D\left(p_{Y|X}(\cdot|x_a)\|p^*_Y\right) \\ =& D\left(\sum_{x\in\mathcal{A}'} p''_X(x)p_{Y|X}(\cdot|x) + \sum_{x\in\mathcal{A}''\setminus\mathcal{A}'} p''_X(x)p_{Y|X}(\cdot|x)\|p^*_Y\right) \end{aligned}$$

$$
\begin{aligned}
&< \left(1 - \sum_{x \in \mathcal{A}'' \setminus \mathcal{A}'} p''_X(x)\right) C + \sum_{x \in \mathcal{A}'' \setminus \mathcal{A}'} p''_X(x) D\left(p_{Y|X}(\cdot|x) \| p^*_Y\right) \\
&\leq \left(1 - \sum_{x \in \mathcal{A}'' \setminus \mathcal{A}'} p''_X(x)\right) C + \max_{x \in \mathcal{A}'' \setminus \mathcal{A}'} D\left(p_{Y|X}(\cdot|x) \| p^*_Y\right) \sum_{x \in \mathcal{A}'' \setminus \mathcal{A}'} p''_X(x),
\end{aligned}
$$

Thus, it leads to an absurd result that

$$
\max_{x \in \mathcal{A}'' \setminus \mathcal{A}'} D\left(p_{Y|X}(\cdot|x) \| p^*_Y\right) > C.
$$

It completes our proof. □

B.3 Proofs of Results in Proposition 3.14

By Proposition 3.9, $\overline{\Gamma}(2, p_{Y|X}) = 1$, so that $p_{Y|X}$ can be expressed as a convex combination of perfect channels and hence $\overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X}) = 1$.

Let

$$
\mathcal{P}'_{\hat{S}} = \left\{p_{\hat{S}} \in \mathrm{dec}(p_{Y|X}) \colon \Gamma(2, p_{\hat{S}}) = 1\right\}.
$$

If $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) = 1$, then there exists a $p_{\hat{S}} \in \mathcal{P}'_{\hat{S}}$ such that the capacity-achieving input distribution, denoted p^*_X, is capacity-achieving for every perfect channel in $\left\{\hat{p}_{Y|X,\hat{S}}(\cdot|\cdot, \hat{s}) \in \hat{\mathcal{P}} : \hat{s} \in \mathrm{supp}(p_{\hat{S}})\right\}$. Thus at least one entry of p^*_X must be $1/2$. Without loss of generality, we assume $p^*_X(x_0) = 1/2$.

If $p^*_X(x_1)$ and $p^*_X(x_2)$ are both positive, then p^*_X is capacity-achieving only for perfect channels

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

By Proposition 3.7, every $p_{\hat{S}} \in \mathrm{dec}(p_{Y|X})$ must satisfy that $|\operatorname{supp}(p_{\hat{S}})| \geq \lceil \log_2 6 \rceil = 3$, which implies that a positive distribution p_X^* is not capacity-achieving for $p_{\hat{S}} \in \mathcal{P}'_{\hat{S}}$.

If $p_X^*(x_1) = 0$, then p_X^* is capacity-achieving for perfect channels

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

However, any convex combination of these four matrices can only yield a channel matrix with at most four distinct probability values, and hence p_X^* is not capacity-achieving for $p_{\hat{S}} \in \mathcal{P}'_{\hat{S}}$.

In all cases, we have shown that p_X^* is not capacity-achieving, which contradicts to the assumption $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) = 1$. Therefore, we have $\overline{\mathrm{IC}}_{\mathrm{D}}(p_{Y|X}) < 1 = \overline{\mathrm{IC}}_{\mathrm{ED}}(p_{Y|X})$.

Bibliography

Asnani, H., Permuter, H., and Weissman, T. (2011). Probing capacity. *IEEE Trans. Inf. Theory*, **57**(11), 317–7332.

Bertsekas, D. P., Nedi, A., and Ozdaglar, A. E. (2003). *Convex Analysis and Optimization*. Athena Scientic.

Bhashyam, S., Sabharwal, A., and Aazhang, B. (2002). Feedback gain in multiple antenna systems. *IEEE Trans. Commun.*, **50**(5), 785 – 798.

Caire, G. and Shamai, S. (1999). On the capacity of some channels with channel state information. *IEEE Trans. Inf. Theory*, **45**(6), 2007–2019.

Caron, R. M., Li, X., Mikusiski, P., Sherwood, H., and Taylor, M. D. (1996). Nonsquare doubly stochastic matrices. *in Institute of Mathematical Statistics Lecture Notes - Monograph Series. Hayward, CA: Institute of Mathematical Statistics*, pages 65–75.

Cover, T. M. and Chiang, M. (2002). Duality between channel capacity and rate distortion with two-sided state infornmation. *IEEE Trans. Inf. Theory*, **48**(6), 1629–1638.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

Csiszár, I. and Körner, J. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press.

Gallager, R. (1968). *Information Theory and Reliable Communication*. Wiley.

Gamal, A. E. and Kim, Y.-H. (2011). *Network Information Theory*. Cambridge University Press.

Gel'fand, S. I. and Pinsker, M. S. (1980). Coding for channel with random parameters. *robl. Control Inf. Theory*, **9**(1), 19–31.

Heegard, C. and Gamal, A. E. (1983). On the capacity of computer memories with defects. *IEEE Trans. on Information Theory*, **29**(5), 731–739.

Ho, S.-W. and Verdú, S. (2010). On the interplay between conditional entropy and error probability. *IEEE Trans. Inf. Theory*, **56**(12), 5930–5942.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Johnson, C. R. (1989). A Gershgorin-type lower bound for the smallest singular value. *Linear Algebra Appl.*, **112**(1), 1–7.

Jöngren, C., Skoglund, M., and Ottersten, B. (2002). Combining beamforming and orthogonal space-time block coding. *IEEE Trans. Inf. Theory*, **48**(3), 611627.

Jurkat, W. and Ryser, H. (1968). Extremal configurations and decomposition theorems. *Problemy peredachi informatsii*, **8**(2), 194–222.

Kuznetsov, N. V. and Tsybakov, B. S. (1974). Coding in memories with defective cells. *Problemy peredachi informatsii*, **10**(2), 52–60.

Plemmons, R. J. (1977). M-matrices characterizations.I—non-signular M-matrices. *Linear Algebra Appl.*, **18**(2), 175–188.

Rosenzweig, A., Steinberg, Y., and S., S. (2005). On channels with partial channel state information at the transmitter. *IEEE Trans. Inf. Theory*, **51**(5), 1817–1830.

Sabharwal, A., Erkip, E., and Aazhang, B. (2000). On channel state information in multiple antenna block fading channels. *Proc. Int. Symp. Information Theory and Its Applications (ISITA), Honolulu, HI,*, page 116119.

Salehi, M. (1992). Capacity and coding for memories with real-time noisy defect information at encoder and decoder. *Proc. IEEE Pt. I*, **139**(2), 113–117.

Shannon, C. E. (1958). Channels with side information at the transmitter. *IBM J. Res. Devel.*, **2**(4), 289–293.

Shulman, N. and Feder, M. (2004). The uniform distribution as a universal prior. *IEEE Trans. Inf. Theory*, **50**(6), 1356–1362.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, **8**(1), 171–176.

Song, L. and Chen, J. (2011). On the capacity of finite-state channels with noisy state information at the encoder. *6th Intl. ICST Conf. Communications and Networking.*

Wang, J., Chen, J., Zhao, L., Cuff, P., and Permuter, H. (2011). On the role of the refinement layer in multiple description coding and scalable coding. *IEEE Trans. Inf. Theory*, **57**(3), 1443–1456.

Yassaee, M. H., Aref, M. R., and Gohari, A. (2014). Achievability proof via output statistics of random binning. *IEEE Trans. Inf. Theory*, **60**(11), 6760–6786.

Zhang, Z. (2007). Estimating mutual information via Kolmogorov distance. *IEEE Trans. Inf. Theory*, **53**(9), 3280–3282.