The University Of Sheffield.

On a Generalised Typicality and Its Applications in Information Theory

By:

Wuling Liu

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Engineering
Department of Electronic and Electrical Engineering

May 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Wuling Liu

Supervised by Dr. Xiaoli Chu

May 2019

Acknowledgements

I would like to express my most sincere gratitude to my supervisor Dr. Xiaoli Chu, for encouraging me to work on this information theoretical topic. Xiaoli not only provides beneficial suggestions on the research methodology and useful research skills, but also leads me to consider how my work might influence the information and communication society. Besides those, she always reminds me of the importance of planning and implementing, which will impact my career and life in the future.

I would like to express my gratitude to Prof. Jie Zhang and Dr. Wei Liu, for their kind help and suggestions on the choice of my topic. I am also very grateful to express my gratitude to Dr. Mikko Vehkapera, for his illuminating comments and advices during our discussion and collaboration.

I would like to thank all my colleagues from the Communications Research Group for the years we spent together in our research laboratory.

Abstract

Typicality lemmas have been successfully applied in many information theoretical problems. The conventional strong typicality is only defined for finite alphabets. Conditional typicality and Markov lemmas can be obtained for strong typicality. Weak typicality can be defined based on a measurable space without additional constraints, and can be easily defined based on a general stochastic process. However, to the best of our knowledge, no conditional typicality or strong Markov lemmas have been obtained for weak typicality in classic works. As a result, some important coding theorems can only be proved by strong typicality lemmas and using the discretisation-and-approximation-technique.

In order to solve the aforementioned problems, we will show that the conditional typicality lemma can be obtained for a generic typicality. We will then define a multivariate typicality for general alphabets and general probability measures on product spaces, based on the relative entropy, which can be a measure of the relevance between multiple sources. We will provide a series of multivariate typicality lemmas, including conditional and joint typicality lemmas, packing and covering lemmas, as well as the strong Markov lemma for our proposed generalised typicality. These typicality lemmas can be used to solve source and channel coding problems in a unified way for finite, continuous, or more general alphabets. We will present some coding theorems with general settings using the generalised multivariate typicality lemmas without using the discretisation-and-approximation technique. Generally, the proofs of the coding theorems in general settings are simpler by using the generalised typicality, than using strong typicality with the discretisation-and-approximation technique.

Table of contents

Chapter 1

Introduction

1.1 Motivations

Typicality lemmas have been successfully applied in many information theoretical problems. Meanwhile, there still exist some limitations, which we will discuss in this section, due to the conventional definitions of typicality. This inspires us to introduce a generalised definition of typicality.

The conventional strong typicality is only defined for finite alphabets. Conditional typicality and Markov lemmas can be obtained for strong typicality, due to the property of empirical distribution in its definition. However, it is difficult to generalise this to continuous or more general alphabets. In [1, Chap. 3], continuous coding theorems were proved by strong typicality lemma and a discretisation-and-approximation technique, which originated in the problem of the semi-continuous channel [2, Chap. 5] ([cf. [3, Chap. 6]).

Mathematically, the applicability of the discretisation-and-approximation technique needs to be verified for every continuous coding theorem. The verification should not be ignored for rigorousness, if strong typicality lemmas are used. Hence, if we can find a new definition of typicality based on general alphabets with all necessary typicality lemmas, these procedures will be circumvented or simplified.

Weak typicality [4] can be defined based on a measurable space without additional constraints, and can be easily defined based on a general stochastic process. However, to the

best of our knowledge, no conditional typicality or strong Markov lemmas have been obtained for weak typicality in classic works of probability theory or information theory. As a result, some important coding theorems can only be proved by strong typicality lemmas and using the discretisation-and-approximation technique.

In order to solve the aforementioned problems, we intend to define a typicality for general alphabets and probability measures, such that the discretisation and approximation can be circumvented in proofs of general coding theorems. Besides this, the proposed typicality will be defined based on the multivariate measure. Conventionally, the weak typicality is defined by the difference between the empirical entropy and the entropy of a given probability distribution. Similarly, the joint weak typicality can be defined by the difference between the information density and the mutual information determined by a given joint probability measure (see [4, Sec. 3.1]). The entropy is a measure of the uncertainty of a single source, and mutual information is a measure of the relevance of two sources. Hence, we are inspired to propose a multivariate typicality definition based on the relative entropy, which can be a measure of the relevance between multiple sources (cf. [5]).

1.2 Contributions

In this thesis, we will generalise the notion of weak typicality. This new definition of typicality is for any general alphabet and any general measure on a product space. This will be based on a study of a generic typicality. We will provide a series of multivariate typicality lemmas, including conditional and joint typicality lemmas, packing and covering lemmas, as well as the strong Markov lemma for our proposed generalised multivariate typicality. This will change the current status that some basic lemmas including the conditional typicality lemma and the strong Markov lemma had only been obtained for strong typicality and thus was restricted in applications with general alphabets and probability measures. These typicality lemmas can be used to solve source and channel coding problems in a unified way for finite, continuous or more general alphabets. We will present some coding theorems with general settings using the generalised multivariate typicality lemmas without

using the discretisation-and-approximation-technique. Generally, the proofs of the coding theorems in general settings are simpler by using the generalised typicality, than using strong typicality with the discretisation-and-approximation technique.

1.3 Thesis Outline

The first two articles in my publication list will contribute most parts of this thesis. The rest which are co-authored will be irrelevant.

In Chap. 2, we will introduce basic mathematical notations and information theoretical definitions for this thesis. We will also review previous works on typicality.

In Chap. 3, we will revisit some approaches to the achievability proof of the channel coding theorem. We will specify the relevance of different approaches and then provide a unified method of the achievability proof, where the typicality plays a key role in this unified method, hence it is sufficient to study the typicality for various probability measures, in order to solve channel coding problems with various general settings.

In Chap. 4, we will focus on the typicality according to the conclusion in Chap. 3. We will first study a generic typicality and specify that a property similar to the renowned conditional typicality lemma exists for the generic typicality. We will then propose a generalised definition of weak typicality for general multivariate alphabets and general measures on product spaces. Based on this new definition, we will derive several typicality lemmas, including conditional and joint typicality lemmas, packing and covering lemmas, as well as the strong Markov lemma. Most of the lemmas have not been defined for the conventional typicality. My publication 1 will contribute to this part. For the convenience of application, we will also provide bivariate versions of the typicality lemmas, some of which have been presented in my publication 2.

In Chap. 5, we will follow the typicality lemmas in Chap. 4 and study several source or channel coding problems. Achievablily proofs of these problems are provided based on generalised typicality lemmas. Our generalised conditional typicality lemma will be applicable for Cover's strong rate-distortion theorem, channel with input constraint, Gelfand

-Pinsker coding for the channel with state, multi-user broadcast channel (BC) coding, all with general settings. Our generalised strong Markov lemma will be applicable for the general Berger-Tung problem. We specify that in most cases it will be simpler to prove the coding theorems by using the generalised typicality, than using strong typicality with the discretizsation-and-approximation technique. We also provide a second-order analysis on multiple access channel (MAC), and specify that a second-order weak typicality is also useful. Most examples have been presented in my publications 1 and 2.

In Chap. 6, we will conclude this thesis and provide some possible topics for future researches.

1.4 List of Publications

As a First Author

1. **W. Liu**, X. Chu and M. Vehkapera, "On generalised Multivariate Typicality Lemmas," under correction and to be resubmitted, 2018.

2. **W. Liu**, X. Chu and J. Zhang, "On a generalised typicality with respect to general probability distributions," 14th Canadian Workshop Inf. Theory, St. John's, NL, Canada, 6-9 July 2015.

As a Co-Author

1. Y. Wu, S. Wang, **W. Liu**, W. Guo and X. Chu, "Iunius: A cross layer peer-to-peer system with device-to-device communications," IEEE Trans. Wireless Commun., vol. 15, no. 10, pp. 7005-7017, Oct. 2016.

2. J. Weng, **W. Liu**, H. Hu, X. Chu and J. Zhang, "A mathematical formulation for channel map and its application in MIMO systems," Loughborough Antennas Propag. Conf., Loughborough, UK, 2-3 Nov. 2015.

3. H. Shao, D. Wu, Y. Li, **W. Liu** and X. Chu, "An improved SNR estimation algorithm for asymmetric pulse-shaped signals," IET Commun., vol.9, no. 14, pp. 1788-1792, Sept. 2015.

4. Y. Wu, **W. Liu**, S. Wang, W. Guo and X. Chu, "Network coding in device-to-device (D2D) communications underlaying cellular networks," IEEE ICC'15, London, UK, 8-12 June 2015.

Chapter 2

Preliminary

2.1 Mathematical Notations

In this thesis, we consider general alphabets and probability measures represented by a probability space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_X)$. $\mathcal{X}$ can be any abstract alphabet and $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_X)$ can be expressed by a product space in this thesis. $\mathbf{P}\{\cdot\}$ denotes the probability of an event, $\mathbf{E}(\cdot)$ denotes the mathematical expectation and $\mathbf{D}(\cdot)$ denotes the variance. We also borrow '$\times$' and '$\prod$' notations from [6] to denote the product probability and the conditional product probability. We follow the convention that letters in upper and lower cases stand for random variables and corresponding realizations, respectively. In order to introduce and study the multivariate typical sequence, we denote by $X^n = (X_1, X_2, \cdots, X_n)$ an n-length sequence, and by $\mathbf{X} = (X_1, X_2, \cdots)$ a semi-infinite sequence, where $\mathbf{X}$ and X^n are $\mathcal{X}$-valued stochastic processes. $P_{\mathbf{X}}$ denotes the infinite sequence $\{P_{X^n}\}_{n=1}^{\infty}$ of probability measures, where X^n is a subsequence of $\mathbf{X}$ for each n. We denote a joint probability measure as P_{XY} and $P_{\mathbf{XY}} = \{P_{X^nY^n}\}_{n=1}^{\infty}$ and transition probability measure as $P_{Y|X}$ and $P_{\mathbf{Y}|\mathbf{X}} = \{P_{Y^n|X^n}\}_{n=1}^{\infty}$. All logarithms are taken to the base e. $P_{X^nY^n}$ is not necessarily a product probability measure denoted as $P_{X^n} \times P_{Y^n}$. Sometimes we use a subscript to distinguish different probability spaces and sequences, for example, an $\mathcal{X}_k$-valued sequence is denoted as $X_k^n = (X_{k,1}, X_{k,2}, \cdots, X_{k,n})$. For brevity, we denote a tuplet of sequences $\{X_k^n\}_{k\in\mathcal{K}}$ by $X_{\mathcal{K}}^n$, and $\{\mathbf{X}_k\}_{k\in\mathcal{K}}$ by $\mathbf{X}_{\mathcal{K}}$. $\mathcal{K}$ always stands for a finite index set throughout this thesis. Besides the sequences of random

variables, we will also introduce sequences of sets, which is denoted as $\mathcal{A}^n(X)$ or $\mathcal{A}^{(n)}$, where we use brackets in the superscript when it might be confused with product spaces.

As in [7], we only use n once in a single notation, for example,

$$P_{XY}^n(B) \doteq P_{X^nY^n}(B^{(n)}),$$
$$d_{P_X||Q_X}^n(x) \doteq d_{P_{X^n}||Q_{X^n}}(x^n),$$
$$\mathcal{A}^n(X,Y) \doteq \mathcal{A}^n(X^n,Y^n).$$

2.2 Probability and Information Theory

A communication system is basically pictured by models including information sources, information channels, encoders and decoders. Messages are generated at information sources and next encoded into transmittable signals for information channels, the transmitted signals will then be received and decoded by receivers. Since statistical communication theory and information theory were established in 1940s, probability notions have been introduced to rigorously describe information and communication models. Then it is possible to analyse coding problems with mathematical and probabilistic tools. In this section, we will provide or reformulate a group of probabilistic definitions from [6], and those definitions will be fundamental components of information models which will be used in this thesis.

Definition 1 (Measurable Space) *Given a set $\mathcal{X}$, a non-empty collection $\mathcal{B}$ of subsets of $\mathcal{X}$ is called a* σ-algebra *on $\mathcal{X}$ if*

1. $\mathcal{A}^{\complement} \in \mathcal{B}$ *for every* $\mathcal{A} \in \mathcal{B}$,
2. $\bigcup_{n=1}^{\infty} \mathcal{A}_n \in \mathcal{B}$ *for every* $\{\mathcal{A}_n \in \mathcal{B}\}_{n=1}^{\infty}$.

The pair $(\mathcal{X},\mathcal{B})$ is called a measurable space.

Definition 2 (Signed Measure) *Given a measurable space , then a* signed measure *on $(\mathcal{X},\mathcal{B})$ is a set function $\mu : \mathcal{B} \mapsto \mathbb{R}$ such that*

1. $\mu(\emptyset) = 0$,

2. $\mu\left(\bigcup_{n=1}^{\infty} A^{(n)}\right) = \sum_{n=1}^{\infty} \mu\left(A^{(n)}\right)$ *for every disjoint sequence* $\{A^{(n)}\}_{n=1}^{\infty}$ *in* $\mathcal{B}$.

If $\mu(A) \geq 0$ *for every* $A \in \mathcal{B}$*, then* μ *is called a* positive measure.

If μ *is positive and* $\mu(X) = 1$*, then* μ *is called a* probability measure.

Definition 3 (Indicator Function) *Given a set* $\mathcal{X}$ *and let* $\mathcal{A} \subset \mathcal{X}$*, then an* indicator function *on* $\mathcal{X}$ *w.r.t.* $\mathcal{A}$ *is defined by*

$$\chi_{\mathcal{A}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A}, \\ 0 & \text{if } x \notin \mathcal{A}. \end{cases}$$

Definition 4 (Measurable Function) *Given two measurable spaces* $(\mathcal{X}, \mathcal{B})$ *and* $(\mathcal{Y}, \mathcal{F})$*, the a mapping from* $\mathcal{X}$ *into* $\mathcal{Y}$ *is called a* measurable function *relative to* $\mathcal{B}$ *and* $\mathcal{F}$ *if the inverse image* $f^{-1}(F) \in \mathcal{B}$ *for every* $F \in \mathcal{F}$.

Definition 5 (Probability Space) *Given a measurable space* $(\Omega, \mathcal{H})$ *and a probability measure* $\mathbf{P}$ *on* $(\mathcal{X}, \mathcal{H})$*, then the triplet* $(\Omega, \mathcal{H}, \mathbf{P})$ *is called a* probability space. *The set* Ω *is called the* sample space *and elements of the* σ*-algebra* $\mathcal{H}$ *are called* events.

Definition 6 (Random Variable) *Given a probability space* $(\Omega, \mathcal{H}, \mathbf{P})$ *and a measurable space* $(\mathcal{X}, \mathcal{B})$*, then a function X from* Ω *into* $\mathcal{X}$ *and relative to* $\mathcal{H}$ *and* $\mathcal{B}$*) is called a* random variable *taking values in* $(\mathcal{X}, \mathcal{B})$.

Definition 7 (Stochastic Process) *Given a measurable space* $(\mathcal{X}, \mathcal{B})$ *and an arbitrary set* T*, and for each* $t \in T$ *let* X_t *be a random variable taking values in* $(\mathcal{X}, \mathcal{B})$*, then the collection* $\{X_t | t \in T\}$ *is called a* stochastic process *with* state space $(\mathcal{X}, \mathcal{B})$ *and* parameter set T.

Equivalently, a stochastic process $\{X_t | t \in T\}$ *can be defined as a random variable* $\mathbf{X}$ *taking values in the product space* $(\mathcal{X}^T, \mathcal{B}^T)$.

Particularly, if $T = \mathbf{N}^+$*, we denote the stochastic process by* $\mathbf{X} = (X_n)_{n=1}^{\infty}$.

Definition 8 (Probability Law) *Given a random variable defined on a probability space* $(\Omega, \mathcal{H}, \mathbf{P})$ *and taking values in* $(\mathcal{X}, \mathcal{B})$*, and let* P_X *be an image measure of* $\mathbf{P}$ *under X, namely,*

$$P_X(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}\{X \in A\},$$

for every $A \in \mathcal{B}$, then P_X is a probability measure on $(\mathcal{X}, \mathcal{B})$ and is called distribution.

Given a stochastic process $\{X_t | t \in T\}$, or equivalently, $\mathbf{X}$*, the* probability law $\mathbf{X}$ *is defined as the distribution $P_{\mathbf{X}}$ of* $\mathbf{X}$*. $P_{\mathbf{X}}$ determines every* joint distribution *defined by*

$$P_{X_{t_1} \cdots X_{t_n}}(A_{t_1} \times \cdots \times A_{t_n}) = \mathbf{P}\{X_{t_1} \in A_{t_1}, \cdots, X_{t_n} \in A_{t_n}\}$$

with n over $\mathbf{N}^+$, and $t_1, \cdots, t_n$ over T, and $A_{t_1}, \cdots, A_{t_n}$ over $\mathcal{B}$.

Throughout this thesis, we will denote random variables by $X, Y, \cdots$, and the corresponding measurable spaces as $(\mathcal{X}, \mathcal{B}_X), (\mathcal{Y}, \mathcal{B}_Y), \cdots$.

We will then employ the above probabilistic notion to define the basic information theoretical notions needed in this thesis.

In information theory, information source is described as a source generating random signals over from a collection of symbols called source alphabet. It can be formally defined as follows.

Definition 9 (Information Source) *Given a source alphabet $\mathcal{X}$, an information source is a stochastic process $\mathbf{X} = (X_n)_{n=1}^{\infty}$ with state space $(\mathcal{X}, \mathcal{B}_X)$ and probability law $P_{\mathbf{X}}$.*

Specifically, if $P_{\mathbf{X}}$ is i.i.d., namely, for all n, $P_X^n = \prod_{k=1}^n P_{X_k}$, where $P_{X_k} = P_X$, then this source is stationary memoryless with respect to P_X.

Definition 10 (Information Channel) *Given a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{XY})$. An information channel is a set of conditional probability measures*

$$\{P_{Y|X}(\cdot|x) | x \in \mathcal{X}\},$$

where $P_{Y|X}(\cdot|x)$ is a probability measure defined on $(\mathcal{Y}, \mathcal{B}_Y)$.

Specifically, a channel with respect to the code-length n is defined by

$$\{P_{Y|X}^n(\cdot|x) | x^n \in \mathcal{X}^n\},$$

and by

$$\{P_{\mathbf{Y}|\mathbf{X}}(\cdot|\mathbf{x})|\mathbf{x}\in\mathcal{X}^{\mathbb{N}^+}\}$$

when the code-length goes to infinity. Throughout this thesis, we briefly use $P^n_{Y|X}$ or $P_{\mathbf{Y}|\mathbf{X}}$ to specify a channel when it is not confusing.

Remark 1 *In some books, the information channel is defined in terms of regular conditional probability. This is stricter than Defn. 10, which is not necessary a regular conditional probability on $(\mathcal{X}\times\mathcal{Y},\mathcal{B}_{XY})$.*

We here present a general point-to-point channel coding problem in terms of the decoding error probability and channel capacity.

Definition 11 (Channel Coding) *Given a channel $P_{\mathbf{Y}|\mathbf{X}}$, and for every fixed code-length n, given a message index set $\mathcal{M}^{(n)}=\{1,2,\cdots,M^{(n)}\}$, we set an encoding function $\varphi^{(n)}:\mathcal{X}^n\to\mathcal{M}^{(n)}$ and a decoding function $\psi^{(n)}:\mathcal{M}^{(n)}\to\mathcal{Y}^n$, then the average decoding error probability $\varepsilon^{(n)}$ is calculated by*

$$\varepsilon^{(n)}=\frac{1}{M^{(n)}}\sum_{k=1}^{M^{(n)}}\mathbf{P}\{\psi^{(n)}(\varphi^{(n)}(k))\neq k\}.$$

We say that $\varphi^{(n)}$ and $\psi^{(n)}$ specifies an $(n,N^{(n)},\varepsilon^{(n)})$-code. If there exists a sequence of $(n,e^{nR},\varepsilon^{(n)})$-code in code-length n, then R is called an achievable rate. The channel capacity C is defined as the supremum of all achievable rates.

2.3 A Review on the Method of Typicality

The notion of typicality, which is more precisely described by various definitions of typical sets and typical sequences, plays a essential and profound role in the probabilistic information theory. All these definitions are based on some certain divergence between the empirical measure of a sequence and a given probability measure.

Strong typicality, which originated from Wolfowitz's work [8] and was well studied in Csiszár and Körner's book [9], is defined by the *uniform distance*[1] (see [10], [11, Sec. 10.6]

[1]The uniform distance is the maximum $|P(x)-Q(x)|$ over all $x\in\mathcal{X}$ given probability measures P and Q on the measurable space $(\mathcal{X},\mathcal{B})$.

and [1, Chap. 2]) or the *total variation distance* (see [12]) between the *empirical measure* and a given discrete probability measure directly. We will first introduce the relevant notions.

The empirical measure is defined as follows.

Definition 12 (Empirical Measure) *Given mutually independent random variables*

$$X_1, X_2, \cdots, X_n$$

taking values in the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ *and having the identical distribution* P_X*, then the* empirical measure *or the* empirical distribution *determined by the realisation* $x_k = X_k(\omega), k = 1, 2, \cdots, n$ *is defined as*

$$\bar{P}_X(A) = \frac{1}{n}\sum_{k=1}^{n} \chi_A(X_k(\omega)) = \frac{1}{n}\sum_{k=1}^{n} \chi_A(x_k), \qquad n \in \mathbf{N}^+, \omega \in \Omega, A \in \mathcal{B}_\mathcal{X}.$$

Before defining the total variation distance, we need to define the partition of a set.

Definition 13 (Partition) *Given a set* $\mathcal{X}$*, then a countable collections of subsets* $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ *of* $\mathcal{X}$ *is called a partition of* $\mathcal{X}$ *if* $\mathcal{A}^{(i)} \cap \mathcal{A}^{(j)} = \emptyset$ *whenever* $i \neq j$*, and if* $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{A}^{(n)}$*. Given a measurable space* $(\mathcal{X}, \mathcal{B})$*, if* $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ *is a partition of* $\mathcal{X}$ *and* $\mathcal{A}^{(n)} \in \mathcal{B}$ *for every n, then* $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ *is called a* $\mathcal{B}$*-measurable partition of* $\mathcal{X}$*.*

In general, the total variation distance is defined as follows (see [13, p. 7]).

Definition 14 (Total variation distance) *Given two probability measures P and Q on the measurable space* $(\mathcal{X}, \mathcal{B}_\mathcal{X})$*, the total variation distance between P and Q is*

$$\|P - Q\| = \sup \sum_k |P(A^{(n)}) - Q(A^{(n)})|,$$

where the supremum is taken over all partitions $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ *of* $\mathcal{X}$*.*

For the total variation distance, more details will be discussed in Sec. 2.4.

Conventionally, the strong typicality can only be defined for discrete probability measures [14]. Recently, Mitran has provided a generalised definition of the empirical measure and then proposed a weak* typicality for general measures [15]. Though this typicality includes a "weak*" in its name, it is actually a generalisation of strong typicality, because it eventually considers a metric of distance between the empirical measure and a given measure based on the Polish space.

Distinct from strong typicality, *weak typicality* is defined by the difference between the empirical entropy and the entropy of a given probability measure (cf. [11, Section 3.1]). As the entropy is a functional of the measure, the empirical entropy is a functional of the empirical measure. Fundamental properties of both weak and strong typicality are rooted in the weak law of large numbers (LLN).

Instead of the entropic functional, more general functionals can also be introduced in the definition of typicality. In [16], Raginsky proposed a generalised strong typicality based on a class of measurable functions of the standard Borel space. In [17], Jeon defined a generalised typicality on the measurable space by introducing an abstract typicality criteria determined by an integrable function.

From the definition of strong typicality and weak typicality, and using the typical average lemma and Pinsker's inequality, it can be deduced that every strongly typical set is a subset of some weakly typical set. This was discussed in [1, Sec. 2.4 and Bibliographic Notes].

The *conditional typicality lemma* is one of the commonly used lemmas in the method of typical sequences. It was formally proposed in [1, Sec. 2.5], but had actually been implied in El Gamal and van der Meulen's alternative proof [18] of Marton's inner bound on the capacity region of the broadcast channel (BC) [19]. In addition to Marton's inner bound, the conditional typicality lemma can be applied to derive, e.g., the Gelfand-Pinsker theorem [20]. Historically, Wolfowitz had considered the binary chanel coding problem, and proposed a binary and second-order version of the conditional strong typicality lemma [8, Lem. 2.1.2]. This has been pointed out and restated by Ahlswede in [3, 21].

However, the conventional conditional typicality lemma is based on strong typicality, which is only defined for discrete probability measures. This is because in the proof of the

conditional strong typicality lemma in [1, Appd. 2A] strictly relies on the strong typicality definition. We can not simply replace the probability measure and the empirical measure with the entropy and the empirical entropy to obtain a proof in the same form for the weak typicality.

Cover proposed a trivariate version of conditional typicality lemma, which he called "Markov lemma" [11, Lem. 15.8.1]. Although Mitran has generalised this Lemma in [15], his result is restricted to Polish alphabets, because the weak* typicality is based on the weak* convergence property on a Polish space. On the other hand, though weak typicality can be defined for general code alphabets and general probability measures, no conditional typicality lemma based on weak typicality has been proposed.

Similar to the conditional typicality lemma, the mutual covering lemma was implied in [18], and then summarized in [1, Lem. 8.1]. A multivariate version of the mutual covering lemma was also obtained in [1, Lem. 8.2]. In [22], a multivariate covering lemma for the discrete memoryless case was proposed. Previously in [23] a stronger version which is called subset typicality lemma by the authors was considered.

The strong Markov lemma [1, Lem 12.1], which is considered as a strictly stronger version of the most commonly used conditional typicality lemma and Cover's Markov lemma, plays a unique role in some scenarios stricter than those where the conditional typicality lemma is sufficient. Recently, some researchers have extended the Markov lemma for countable alphabets [24, 25]. However, there is still a lack of the Markov lemma based on a definition of typicality with respect to a general joint measure on a product space.

Besides aforementioned asymptotic typicality lemmas, there are recently some works on non-asymptotic ones [26, 27]. Essentially, Feinstein-type lemmas are expressed in a non-asymptotic fashion, and it is possible to obtain non-asymptotic typicality lemmas from Feinstein-type lemmas.

2.4 Equivalent Definitions of Total Variation Distance

As mentioned in Sec. 2.3, total variation distance is used by Cuff in his typicality definition. Total variation distance is generally used in information theoretical works. However, total variation distance is often defined by various authors in different forms, and the equivalence of those definitions were rarely discussed in literatures. Hence, in this chapter, we will introduce the equivalent definitions of total variation norm from mathematical literatures and explicitly prove the equivalence of three forms of total variation distance definitions.

We will reformulate and reorganise definitions and theorems from [6, 28, 29], including Defns. 15-20 and Thms. 1-7, as follows.

Definition 15 (Normed Linear Space) *A real (or complex) vector space $\mathcal{X}$ is called a* normed linear space *if to each $x \in \mathcal{X}$ there is associated a real number $\|x\|$, called the* norm *of x, such that*

1. $\|x\| \geq 0$ *for all* $x \in \mathcal{X}$ *and* $\|x\| = 0$ *if and only if* $x = 0$,
2. $\|x+y\| \leq \|x\| + \|y\|$ *for all* $x, y \in \mathcal{X}$,
3. $\|\alpha x\| = |\alpha| \, \|x\|$ *for all* $x \in \mathcal{X}$ *and all* $\alpha \in \mathbb{R}$ *(or* $\alpha \in \mathbb{C}$*).*

Definition 16 (Total Variation) *Given a measurable space $(\mathcal{X}, \mathcal{B})$ with a signed measure μ, for every $A \in \mathcal{B}$, we define a set function $|\mu|$ on $\mathcal{B}$ by*

$$|\mu|(A) = \sup \sum_{n=1}^{\infty} \left| \mu(A^{(n)}) \right|,$$

where the supremum is taken over all $\mathcal{B}$-measurable partitions $\{A^{(n)}\}_{n=1}^{\infty}$ of A, then $|\mu|$ is called the total variation *of μ. More specifically, we denote $\|\mu\| = |\mu|(\mathcal{X})$.*

Theorem 1 (Total Variation Measure) *Given a measurable space $(\mathcal{X}, \mathcal{B})$ with a total variation $|\mu|$, then $|\mu|$ is positive measure on $\mathcal{B}$ and is thus called a* total variation measure.

Theorem 2 (Total Variation Norm) *Given a measurable space $(\mathcal{X}, \mathcal{B})$, then the set of all signed measures on $\mathcal{B}$ is a vector space, $\|\mu\|$ is a norm on this vector space and is thus called a* total variation norm.

Definition 17 (Positive and Negative Variations) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a signed measure* μ *and the associated total variation measure* $|\mu|$*, and define*

$$\mu^+ = \frac{1}{2}(|\mu|+\mu),$$
$$\mu^- = \frac{1}{2}(|\mu|-\mu),$$

μ^+ *and* μ^- *are called* positive and negative variations *of* μ*, respectively. Thus*

$$\mu = \mu^+ - \mu^-,$$
$$|\mu| = \mu^+ + \mu^-,$$

From Thm. 2, μ^+ *and* μ^- *are both positive measures on* $\mathcal{B}$*.*

The following theorem provides an equivalent definition of the total variation norm.

Theorem 3 (Hahn-Jordan Decomposition) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a signed measure* μ*, then there exists a set* $B \in \mathcal{B}$ *with* $B^\complement = \mathcal{X} \setminus B$*, and such that the positive and negative variations* μ^+ *and* μ^- *of* μ *satisfy*

$$\mu^+(A) = \mu(A \cap B),$$
$$\mu^-(A) = \mu(A \cap B^\complement)$$

for every $A \in \mathcal{B}$*.*

$(B, B^\complement)$ *is called a* Hahn decomposition *of* $\mathcal{X}$ *induced by* μ*, and* (μ^+,μ^-) *is called the* Jordan decomposition *of* μ*.*

Corollary 1 (Decompositional Definition of Total Variation Norm) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a signed measure* μ*, the associated total variation norm* $\|\mu\|$ *and the associated positive and negative variations* μ^+ *and* μ^-*, then*

$$\|\mu\| = \mu^+(\mathcal{X}) - \mu^-(\mathcal{X}) = \sup_{B\in\mathcal{B}} \Big(\mu(B) - \mu(B^\complement)\Big)$$

Definition 18 (Absolute Continuity) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *and singed measures* μ *and* ν*, then* ν *is called* absolutely continuous *w.r.t.* μ *(denoted by* $\nu \ll \mu$*) if* $\mu(A)=0 \Rightarrow \nu(A)=0$ *for every* $A \in \mathcal{B}$*. Furthermore,* μ *and* ν *are called* equivalent *(denoted by* $\nu \approx \mu$*) if* $\nu \ll \mu$ *and* $\mu \ll \nu$*.*

Definition 19 (Concentration) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *and a singed measure* μ*, then* μ *is called* concentrated *on* $C \in \mathcal{B}$ *if* $\mu(A)=\mu(A\cap C)$ *for every* $A \in \mathcal{B}$*.*

Definition 20 (Mutual Singularity) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *and singed measures* μ *and* ν*, then* μ *and* ν *are called* mutually singular *(denoted by* $\mu \perp \nu$*) if there exists* $C \in \mathcal{B}$ *such that* $\mu(C)=\nu(C^{\complement})=0$*.*

Theorem 4 (Lebesgue Decomposition) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a positive* σ*-finite measure* μ *and a signed measure* ν*, there exits a unique pair of signed measures* ν_a *and* ν_s *on* $\mathcal{B}$ *such that* $\nu_a \ll \mu$*,* $\nu_s \perp \mu$ *and* $\nu=\nu_a+\nu_s$*. Thus there exists* $C \in \mathcal{B}$ *such that* $\mu(C)=\nu_a(C)=\nu_s(C^{\complement})=0$*.*$(\nu_a,\nu_s)$ *is called the* Lebesgue decomposition *of* ν *w.r.t.* μ*. Furthermore, if* ν *is positive, then* ν_a *and* ν_s *are also positive.*

Theorem 5 (Radon-Nikodym Derivative) *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a positive* σ*-finite measure* μ *and a signed measure* ν *such that* $\nu \ll \mu$*, then there exists a* $\mathcal{B}$*-measurable function* g *such that for every* $\mathcal{B}$*-measurable function* f*,*

$$\int f \,\mathrm{d}\nu = \int fg \,\mathrm{d}\mu.$$

Such g *is unique* μ*-a.e., and is called a* Radon-Nikodym derivative *denoted by* $g=\mathrm{d}\nu/\mathrm{d}\mu$*.*

Theorem 6 *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a signed measure* μ*, then* $\mu \ll |\mu|$ *and* $h=\mathrm{d}\mu/\mathrm{d}|\mu|$ *satisfies* $|h(x)|=1$ *for every* $x \in \mathcal{X}$*.*

Theorem 7 *Given a measurable space* $(\mathcal{X},\mathcal{B})$ *with a positive measure* μ *and a signed measure* ν *such that* $\nu \ll \mu$*, let* $g=\mathrm{d}\nu/\mathrm{d}\mu$*, then for every* $A \in \mathcal{B}$*,*

$$|\nu|(A)=\int_A |g| \,\mathrm{d}\mu.$$

Corollary 2 (Integral Definition of Total Variation Norm for Positive Measure) *Given a measurable space $(\mathcal{X},\mathcal{B})$ with a positive measure μ and a signed measure ν such that $\nu \ll \mu$, let $g = \mathrm{d}\nu/\mathrm{d}\mu$, then*

$$\|\nu\| = \int |g| \, \mathrm{d}\mu.$$

We will then prove the following theorems, which provides another equivalent definition of the total variation norm.

Theorem 8 *Given a measurable space $(\mathcal{X},\mathcal{B})$ with a positive measure μ and a signed measure ν, let (ν_a,ν_s) be the Lebesgue decomposition of ν w.r.t. μ, $g = \mathrm{d}\nu_a/\mathrm{d}\mu$, $h = \mathrm{d}\nu/\mathrm{d}|\nu|$, $\bar{h} = 1/h$, and C be any set in $\mathcal{B}$ such that $\mu(C) = \nu_a(C) = \nu_s(C^{\complement}) = 0$, then for every $A \in \mathcal{B}$,*

$$|\nu|(A) = \int_A |g| \, \mathrm{d}\mu + \int_A \bar{h} \, \mathrm{d}\nu_s. \tag{2.1}$$

Furthermore, if $\nu(B) \geq 0$ for every $B \in \mathcal{B}$ such that $B \subseteq C$, then

$$|\nu|(A) = \int_A |g| \, \mathrm{d}\mu + \nu(A \cap C). \tag{2.2}$$

Proof: For every $\mathcal{B}$-measurable function f,

$$\int f \, \mathrm{d}\nu = \int fh \, \mathrm{d}|\nu|, \tag{2.3}$$

and on the other hand,

$$\int f \, \mathrm{d}\nu = \int f \, \mathrm{d}\nu_a + \int f \, \mathrm{d}\nu_s = \int fg \, \mathrm{d}\mu + \int f \, \mathrm{d}\nu_s. \tag{2.4}$$

For every $A \in \mathcal{B}$, let $f = \bar{h}\chi_A$ in (2.3) and (2.4), then

$$\int fh \, \mathrm{d}|\nu| = |\nu|(A), \tag{2.5}$$

$$\int fg \, \mathrm{d}\mu = \int_A \bar{h}g \, \mathrm{d}\mu, \tag{2.6}$$

$$\int f \, \mathrm{d}\nu_s = \int_A \bar{h} \, \mathrm{d}\nu_s. \tag{2.7}$$

And for every $A \in \mathcal{B}$ such that $A \subseteq C^{\complement}$, from (2.3)-(2.7) and $\nu_s(C^{\complement}) = 0$, we have

$$|\nu|(A) = \int_A \bar{h} g \, \mathrm{d}\mu,$$

μ and $|\nu|$ are both positive, thus $\bar{h}g \geq 0$ μ-a.e. on $C^{\complement}$. Because $|h| = 1$, we have $\bar{h}g = |g|$ μ-a.e. on $C^{\complement}$. Hence, for every $A \in \mathcal{B}$,

$$\int_A \bar{h} g \, \mathrm{d}\mu \overset{(a)}{=} \int_{A \cap C^{\complement}} \bar{h} g \, \mathrm{d}\mu = \int_{A \cap C^{\complement}} |g| \, \mathrm{d}\mu \overset{(b)}{=} \int_A |g| \, \mathrm{d}\mu, \tag{2.8}$$

where (a) and (b) are both from $\mu(C) = 0$.

Finally, (2.1) is obtained from (2.3)-(2.7) and (2.8). And if $\nu(B) \geq 0$ for every $B \in \mathcal{B}$ such that $B \subseteq C$, then again from $|\nu|$ is positive, we have $h \geq 0$ $|\nu|$-a.e. on C. From $|h| = 1$ and $\nu_a(C) = 0$, we have $|\nu| = 0 \Rightarrow \nu = 0 \Rightarrow \nu_s = 0$ and thus $h = 1$ ν_s-a.e. on C. Hence, $\int_A \bar{h} \, \mathrm{d}\nu_s = \nu_s(A \cap C) = \nu(A \cap C)$, and (2.2) is obtained. ■

Corollary 3 (Integral Definition of Total Variation Norm for Signed Measure) *Given a measurable space $(\mathcal{X}, \mathcal{B})$ with a positive measure μ and a signed measure ν, let (ν_a, ν_s) be the Lebesgue decomposition of ν w.r.t. μ, $g = \mathrm{d}\nu_a/\mathrm{d}\mu$, $h = \mathrm{d}\nu/\mathrm{d}|\nu|$, $\bar{h} = 1/h$, and C be any set in $\mathcal{B}$ such that $\mu(C) = \nu_a(C) = \nu_s(C^{\complement}) = 0$, then*

$$\|\nu\| = \int |g| \, \mathrm{d}\mu + \int \bar{h} \, \mathrm{d}\nu_s.$$

Furthermore, if $\nu(B) \geq 0$ for every $B \in \mathcal{B}$ such that $B \subseteq C$, then

$$\|\nu\| = \int |g| \, \mathrm{d}\mu + \nu(C).$$

In general, for a vector space $\mathcal{X}$, the *distance* between two vectors $x, y \in \mathcal{X}$ induced by a norm $\|\cdot\|$ is defined as $\|x - y\|$. Hence we can define the total variation distance from the equivalent definitions of total variation norm in theorem 2 and corollaries 1 and 2.

Definition 21 (Total Variation Distance) *Given a measurable space $(\mathcal{X}, \mathcal{B})$ with a positive σ-finite measure μ and signed measures φ and λ, then the total variation distance $\|\varphi - \lambda\|$*

between φ and λ is defined in three equivalent forms as follows.

1. $$\|\varphi-\lambda\| = \sup \sum_{n=1}^{\infty} \left|\varphi(A^{(n)}) - \lambda(A^{(n)})\right|,$$

   where the supremum is taken over all $\mathcal{B}$-measurable partitions $\{A^{(n)}\}_{n=1}^{\infty}$ of $\mathcal{X}$.

2. $$\|\varphi-\lambda\| = \sup_{B\in\mathcal{B}} \left((\varphi(B)-\lambda(B)) - \left(\varphi(B^{\complement})-\lambda(B^{\complement})\right)\right) = 2\sup_{B\in\mathcal{B}} (\varphi(B)-\lambda(B)).$$

3. *Let $\nu = \varphi - \lambda$, (ν_a, ν_s) be the Lebesgue decomposition of ν w.r.t. μ, $g = \mathrm{d}\nu_a/\mathrm{d}\mu$, $h = \mathrm{d}\nu/\mathrm{d}|\nu|$, $\bar{h} = 1/h$, then define*

   $$\|\varphi-\lambda\| = \int |g|\,\mathrm{d}\mu + \int \bar{h}\,\mathrm{d}\nu_s.$$

   If $\varphi \ll \mu$ and $\lambda \ll \mu$, then $\nu_a = \varphi - \lambda$ and $\nu_s = 0$, let $g_1 = \mathrm{d}\varphi/\mathrm{d}\mu$ and $g_2 = \mathrm{d}\lambda/\mathrm{d}\mu$, then

   $$\|\varphi-\lambda\| = \int |g_1 - g_2|\,\mathrm{d}\mu.$$

   If φ is positive and λ is positive and σ-finite, and let $\mu = \lambda$, $f = \mathrm{d}\varphi_a/\mathrm{d}\mu$, (φ_a, φ_s) be the Lebesgue decomposition of φ w.r.t. μ, and C be any set in $\mathcal{B}$ such that $\mu(C) = \varphi_a(C) = \varphi_s(C^{\complement}) = 0$, then $\nu_a = \varphi_a - \lambda$, $\nu_s = \varphi_s$, $g = f - 1$ μ-a.e., $\nu_a(C) = \nu_s(C^{\complement}) = 0$, ν is positive on C and $\nu(C) = \varphi(C)$. Hence,

   $$\|\varphi-\lambda\| = \int |f-1|\,\mathrm{d}\lambda + \nu(C),$$

   And more specifically, if φ is positive,λ is positive and σ-finite, and $\varphi \ll \lambda$, let $f = \mathrm{d}\varphi/\mathrm{d}\lambda$, then

   $$\|\varphi-\lambda\| = \int |f-1|\,\mathrm{d}\lambda.$$

More specifically, under the conditions of Defn. 21, it can be verified that the total variation distance between two discrete probability measures φ and λ has the equivalent

definitions as follows.

1.
$$\|\varphi - \lambda\| = \sup \sum_{n=1}^{\infty} \left| \varphi(A^{(n)}) - \lambda(A^{(n)}) \right|,$$

where the supremum is taken over all $\mathcal{B}$-measurable partitions $\{A^{(n)}\}_{n=1}^{\infty}$ of $\mathcal{X}$.

2.
$$\|\varphi - \lambda\| = 2 \sup_{B \in \mathcal{B}} (\varphi(B) - \lambda(B)).$$

3.
$$\|\varphi - \lambda\| = \sum_{x \in \mathcal{X}} |\varphi(x) - \lambda(x)|.$$

The first definition is used in Pinsker's book [13], where Pinsker's inequality was originally proposed; the second definition is commonly used in later works; the third definition is used by Cuff for his typicality definition [12], where he used a normalised coefficient $\frac{1}{2}$.

Chapter 3

Typicality in Coding Problems

In [30], random coding, Feinstein-type lemmas, and typicality method were treated as different approaches to the achievability proof of the channel coding theorem. They have been considered as different methods providing different error probability bounds in later books or lecture notes (see [11, p. 240] and [31, Sec. 17.4]). In this chapter, we will revisit some of the achievability proofs and show the relevance of different approaches. We will also point out that typicality plays a fundamental role in these proofs.

3.1 Asymptotic Equipartition Property and Typicality

The asymptotic equipartition property (AEP) or entropy equipartition property expresses a property of an $\mathcal{X}$-valued random sequence X^n with probability distribution P_X^n that the logarithm of its probability $P^n(X)$ is close to $-H(X^n)$ defined with P_X^n, and the probability of the event that X^n satisfies AEP approaches 1 when n tends to infinity [32]. We illustrate the mathematical definition of AEP for the i.i.d. case as follows (cf. [11, 33–35]).

Theorem 9 (AEP) *If X^n is an i.i.d. sequence generated according to P_X, then*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|-\frac{1}{n}\log P_X^n(X) - H(X)\right| < \varepsilon\right\} = 1$$

for any $\varepsilon > 0$. *We denote*

$$\mathcal{A}_\varepsilon^n(X) = \left\{ x^n \mid \left| -\frac{1}{n}\log P_X^n(x) - H(X) \right| < \varepsilon \right\},$$

and we have $(1-\varepsilon)e^{n(H(X)-\varepsilon)} \leq |\mathcal{A}_\varepsilon^n(X)| \leq e^{n(H(X)+\varepsilon)}$ *for sufficiently large n.*

The theorem can be naturally derived from the weak law of large numbers (WLLN) or Chebychev's inequality. If we replace X with a pair (X,Y) in Thm. 9, we will obtain the joint AEP.

Typicality describes the property of a set with a P_X-probability close to 1 of sequences x^n's, and this set is naturally called a typical set [36]. Both notions of the AEP and the typicality originated in Shannon's celebrated work [37] (see [38]).We can see that AEP has already provided a possible approach to define the typicality. The AEP and typicality have been used in many pioneering works (ex. [39, 40]) and current textbooks (ex. [11, 14, 41–43]) as major tools for deriving coding theorems.

Theorem 10 *Given a stationary and memoryless channel with* $\mathcal{X}, \mathcal{Y}$ *and* $P_{Y|X}(\cdot|x)$*, then the channel capacity* $C = \sup_{P_X} I(X;Y)$

We first provide a proof based on the AEP in terms of entropy, when the channel alphabet is finite (cf. [35]).

Proof:

Random coding For fixed P_X and n, randomly and independently generate e^{nR} many $x^n(m)$'s according to the n-fold product of P_X, i.e. $x^n(m)$ is a realization of the i.i.d. $\mathcal{X}^n$-valued stochastic process $X^n(m)$. Send $x^n(m)$ when $M = m$.

Decoding Assume y^n is received. If $(x^n(\hat{m}), y^n) \in \mathcal{A}_\varepsilon^{(n)}$ where $\mathcal{A}_\varepsilon^{(n)} = \mathcal{A}_\varepsilon^{(n)}(X,Y) \cap (\mathcal{A}_\varepsilon^{(n)}(X) \times \mathcal{A}_\varepsilon^{(n)}(Y))$, then declare $\hat{m}$ is sent.

Error analysis $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where $\mathcal{E}_1 = \{(x^n(m), y^n) \notin \mathcal{A}_\varepsilon^{(n)}\}$ and $\mathcal{E}_2 = \{(x^n(\hat{m}), y^n) \in \mathcal{A}_\varepsilon^{(n)}\}$ for some $\hat{m} \neq m$. $\lim_{n\to\infty} \mathbf{P}(\mathcal{E}_1) = 0$ following the AEP and joint AEP. As for $\mathcal{E}_2$, we have

$$\mathbf{P}(\mathcal{E}_2) \leq \sum_{\hat{m}\neq m} \mathbf{P}\{(X^n(\hat{m}), Y^n) \in \mathcal{A}_\varepsilon^{(n)}\} \leq e^{nR}\mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\varepsilon^{(n)}\},$$

where $\tilde{X}^n \sim P_X^n$ and $\tilde{Y}^n \sim P_Y^n$ are independent, Furthermore,

$$\begin{aligned}\mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\varepsilon^{(n)}\} \leq & \sum_{\mathcal{A}_\varepsilon^{(n)}} P_X^n(x) P_Y^n(y) \\ \leq & |\mathcal{A}_\varepsilon^{(n)}(X,Y)| e^{-n(H(X)-\varepsilon)} e^{-n(H(Y)-\varepsilon)} \\ \leq & e^{n(H(X,Y)+\varepsilon)} e^{-n(H(X)-\varepsilon)} e^{-n(H(Y)-\varepsilon)}\end{aligned}$$

according to the AEP and joint AEP. Hence, $\lim_{n\to\infty} \mathbf{P}(\mathcal{E}_2) = 0$ when $R < I(X;Y) = H(X) + H(Y) - H(X,Y)$. ■

Joint typicality can be defined based on the joint AEP. In fact, $\mathcal{A}_\varepsilon^{(n)} = \mathcal{A}_\varepsilon^{(n)}(X,Y) \cap (\mathcal{A}_\varepsilon^{(n)}(X) \times \mathcal{A}_\varepsilon^{(n)}(Y))$ was exactly the definition of jointly typical set in [11]. In [4], Han defined the jointly typical set as

$$\mathcal{T}_\varepsilon^n(X,Y) = \left\{(x^n, y^n) \mid \left|\frac{1}{n} i^n(x;y) - I(X;Y)\right| < \varepsilon\right\},$$

where

$$i^n(x;y) = \log \frac{P_{XY}^n(x,y)}{P_X^n(x) P_X^n(y)}$$

is called information density[1]. It is evident that $\mathcal{A}_\varepsilon^{(n)} \subset \mathcal{T}_\delta^n(X,Y)$ where $\delta = 3\varepsilon$. Similar to the joint AEP,

$$\lim_{n\to\infty} \mathbf{P}\{(X^n, Y^n) \in \mathcal{T}_\varepsilon^n(X,Y)\} = 1,$$

where $(X^n, Y^n) \sim P_{XY}^n$ and P_{XY}^n is n-fold product of P_{XY}. We then provide an alternative proof based on Han's typicality definition.

Proof:

[1] We will formally introduce the notion of information density for general random variables. Here we only use the discrete version.

Random coding The same as the first proof.

Decoding Assume y^n is received. If $(x^n(\hat{m}), y^n) \in \mathcal{T}_\varepsilon^n(X,Y)$, then declare $\hat{m}$ is sent.

Error analysis $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where $\mathcal{E}_1 = \{(x^n(m), y^n) \notin \mathcal{T}_\varepsilon^n(X,Y)\}$ and $\mathcal{E}_2 = \{(x^n(\hat{m}), y^n) \in \mathcal{T}_\varepsilon^n(X,Y)\}$ for some $\hat{m} \neq m$. We have

$$\lim_{n\to\infty} \mathbf{P}(\mathcal{E}_1) = \lim_{n\to\infty} \mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \notin \mathcal{T}_\varepsilon^n(X,Y)\} = 0 \tag{3.1}$$

following the property of the jointly typical set. Similar to the first proof, we have

$$\mathbf{P}(\mathcal{E}_2) \leq e^{nR} \mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{T}_\varepsilon^n(X,Y)\},$$

and

$$\begin{aligned} \mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{T}_\varepsilon^n(X,Y)\} \leq & \sum_{\mathcal{T}_\varepsilon^n(X,Y)} P_X^n(x) P_Y^n(y) \\ = & \sum_{\mathcal{T}_\varepsilon^n(X,Y)} \frac{P_X^n(x) P_Y^n(y)}{P_{XY}^n(x,y)} P_{XY}^n(x,y) \\ \leq & e^{-n(I(X;Y)-\varepsilon)} \end{aligned} \tag{3.2}$$

according to the definition of jointly typical set. Hence, $\lim_{n\to\infty} \mathbf{P}(\mathcal{E}_2) = 0$ when $R < I(X;Y)$. ■

In the alternative proof, we do not use the entropy and get rid of bounding the cardinality of the typical set. As is generally known, information entropy is defined only for finite and countably infinite valued random variables, and the AEP is well defined only when the entropy with respect to the given probability distribution is finite. Although the AEP was generalised to the continuous alphabet by replacing the entropy and the cardinal of typical set with the differential entropy and volume of typical set [11], it is not plausible for the general alphabet. Besides, the differential entropy of a continuous probability is not a measure of uncertainty as exact as entropy, and was considered 'not the prettiest object in

information theory' [44]. On the contrary, $i^n(X;Y)$ in Han's joint typicality definition can be generalised to any abstract alphabet by replacing the fractional with the Radon-Nikodym derivative. Hence we intend to employ similar notions of typicality in this thesis.

3.2 Feinstein-Type Lemma and Typicality

In [45] Feinsten provided a rigorous proof of Shannon's channel coding theorem. The key steps had then been extracted and reformulated by different authors (including Feinsten himself) in [2, 46–48]. This is usually called Feinstein's fundamental lemma, and then was used in many books (ex. [49, 50]). Other forms of Feinstein-type lemmas were obtained in [4, 7, 8, 51, 52].

Basically, we can restate the Feinstein-type lemma as follows.

Lemma 1 *Given a stationary and memoryless channel with $\mathcal{X}, \mathcal{Y}$ and $P_{Y|X}$, and $\varepsilon > 0$, then for sufficiently large n, there exits a set $\{x_k^n\}_{k=1}^{M_n}$ and a measurable partition $\{\mathcal{D}_k^{(n)}\}_{k=1}^{M_n}$ of $\mathcal{Y}^n$ with $P^n(\mathcal{D}_k|x_k) > 1-\varepsilon$ for all $k = 1,2,\cdots,M_n$, and $P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) \geq \frac{\varepsilon}{2}$, where P_Y^n is a channel output distribution induced by any channel input distribution P_X^n.*

Proof: The details of the proof can be found in the aforementioned works. Here we state the major steps extracted from those proofs.

Fix P_X^n and denote $P_{XY}^n = P_X^n \times P_{Y|X}^n$. Let $\{\mathcal{A}^n\}_{n=1}^{\infty}$ be a set sequence satisfying $\mathcal{A}^n \in \mathcal{B}_{XY}^n$ and $P_{XY}^n(\mathcal{A}) > 1 - \frac{\varepsilon}{2}$ for sufficiently large n, and

$$\mathcal{A}^{(n)}(x) = \{y^n \mid (x^n, y^n) \in \mathcal{A}^{(n)}\}.$$

Then we choose $\{x_k^n\}_{k=1}^{M_n}$ from all $x^n \in \mathcal{X}^n$ and $\{\mathcal{D}_k^{(n)}\}_{k=1}^{M_n}$ as

$$\begin{aligned}
\mathcal{D}_1^{(n)} &= \mathcal{A}^{(n)}(x_1), \\
\mathcal{D}_k^{(n)} &= \mathcal{A}^{(n)}(x_k) \setminus \bigcup_{j=1}^{k-1} \mathcal{D}_j^{(n)}, \quad k = 2,\cdots,M_n,
\end{aligned}$$

subject to

$$P_{Y|X}^n(\mathcal{D}_k^{(n)}|x_k) > 1-\varepsilon, \quad k = 1, \cdots, M_n.$$

Due to the assumptions on $\{\mathcal{D}_k^{(n)}\}_{k=1}^{M_n}$, there always exists a maximum M_n.

Fix P_X^n, and let P_Y^n be the marginal distribution of $P_X^n \times P_{Y|X}^n$. When M_n is the largest as mentioned above, we have

$$P_{Y|X}^n(\mathcal{A}(x)|x) = P_{Y|X}^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) + P_{Y|X}^n(\mathcal{A}(x) \setminus \bigcup_{k=1}^{M_n} \mathcal{D}_k) \leq P_{Y|X}^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) + 1 - \varepsilon,$$

thus we have

$$P_{XY}^n(\mathcal{A}) \leq P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) + 1 - \varepsilon.$$

Then from the condition on which we set $\{\mathcal{A}^n\}_{n=1}^{\infty}$, we have for sufficiently large n

$$P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) \geq \frac{\varepsilon}{2}. \tag{3.3}$$

■

Furthermore, if we let $\{\mathcal{D}_k^{(n)}\}_{k=1}^{M_n}$ be taken in different forms, we can estimate M_n and ε in various ways.

Theorem 11 *In Lem. 1, M_n and ε_n can be estimated as $M_n > e^{n(C-\varepsilon)}$ for all $0 < \varepsilon \leq 1$, where $C = \sup_{P_X} I(X;Y)$.*

Proof: Following the proof of Lem. 1, let $\mathcal{A}^{(n)}(x)$ be taken as

$$\mathcal{A}^{(n)}(x) = \left\{ y^n \mid \left| -\frac{1}{n}\log P_Y^n(y) - H(Y) \right| < \delta_1 \wedge \left| -\frac{1}{n}\log P_{Y|X}^n(y|x) - H(Y|X) \right| < \delta_2 \right\},$$

for some $\delta_1, \delta_2 > 0$, or

$$\mathcal{A}^{(n)}(x) = \left\{ y^n \mid \left| \frac{1}{n} i^n(x;y) - I(X;Y) \right| < \frac{\varepsilon}{2} \right\},$$

for some $\varepsilon > 0$. For both cases, by the AEP or the property of typicality, we have

$$P_Y^n(\mathcal{A}(x)) = \sum_{\mathcal{A}^{(n)}(x)} P_Y^n(y^n) < e^{-n(H(Y)-H(Y|X)-\delta_1-\delta_2)}$$

or

$$P_Y^n(\mathcal{A}(x)) < e^{-n(I(X;Y)-\varepsilon)} \tag{3.4}$$

We set $\delta_1 + \delta_2 < \frac{\varepsilon}{2}$ and take the maximizing P_X^n, then we have for both cases that

$$P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) \leq P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{A}(x_k)) < M_n e^{-n(C-\frac{\varepsilon}{2})}. \tag{3.5}$$

Then following (3.3) and (3.5), we have

$$M_n > \frac{\varepsilon}{2} e^{-n(C-\frac{\varepsilon}{2})} > e^{-n(C-\varepsilon)}$$

for sufficiently large n and when $n > \frac{2}{\varepsilon}\log\frac{2}{\varepsilon}$.

More details of the proof can be found in [46, Chap. IV] and [2, Sec. 4.1], [47, Sec. 3] (see also [49, Thm. 7.7] and [53, Lem. 3.7.1]). ■

Theorem 12 *In Lem. 1, M_n and ε_n can be estimated as $\varepsilon_n \leq M_n e^{-a} + P_{XY}^n\{i \leq a\}$ for all $a > 0$, where $i = \frac{1}{n} i^n(x;y)$.*

Proof: Following the proof of Lem. 1, let $\mathcal{A}^{(n)}(x)$ be taken as

$$\mathcal{A}^{(n)}(x) = \left\{ y^n \mid \frac{1}{n} i^n(x;y) > a \right\},$$

for some $a > 0$. We only specify that equation (3.3) is equivalent to the lower bound inequalities in both [47] and [48]. Hence, Thm. 12 is essentially equivalent to Thm. 11. More details of the proof can be found in [48, Thm. 2] (see also [50, Lem. 14.1]) and [4, Lem. 3.4.1]. ■

Theorem 13 *In Lem. 1 M_n and ε_n can be estimated as $M_n > e^{(nC-c\sqrt{n})}$ for all $0 < \varepsilon_n \leq 1$ and some constant $c = c(\varepsilon_n)$.*

Proof: Following the proof of Lem. 1, let $\mathcal{A}^{(n)}(x)$ be taken as

$$\mathcal{A}^{(n)}(x) = \left\{ y^n \mid \|\pi(x,\cdot|x^n,y^n) - P_{XY}^n(x,\cdot)\| < \gamma\sqrt{n} \right\}, \tag{3.6}$$

thus following the lemmas in [21, Sec. 5.2.7], we have for sufficiently large n that

$$P_Y^n(\bigcup_{k=1}^{M_n} \mathcal{D}_k) < M_n e^{-(nC - c(\gamma)\sqrt{n})}. \tag{3.7}$$

It should be noted that we remedy the definition of $\mathcal{T}_{W,\delta}^n(x)$ in [21, Sec. 5.2.7] following a similar way of the strong typicality in [1], and obtain the $\mathcal{A}^{(n)}(x)$ in (3.6), without any changes in the conclusion.

Then similar to the proof of Thm. 11, Thm. 13 can be established following (3.3) and (3.7). More details of the proof can be found in [21, Thm. 29] (see also [8, Thms. 3.2.1 and 7.2.1])

■

The Feinstein-type lemma essentially provides a group of codewords and corresponding decision regions of a channel coding scheme with a bounded decoding error probability. In the above statement, ε is a maximal error probability, while in some works [4, 7], ε was an average error probability. Specifically, in [4, Lem. 3.4.1], Han took the decision regions basically according to the jointly typical set, hence established a direct link between the Feinstein-type lemma and the joint typicality, despite the minor difference in error probability. In [51], Shannon provided an average error probability by employing random coding and the maximum likelihood decoding.

Although the above derivations of error probability were through different approaches, we can establish a link between those results. Here we derive the Feinstein-type lemma in terms of a maximal error probability from the one in terms of an average error probability, without any difficulty.

Lemma 2 *Given a memoryless channel with $\mathcal{X}, \mathcal{Y}$ and $P_{Y|X}$, and a set $\{x_k^n\}_{k=1}^{M_n}$ and a partition $\{\mathcal{D}_k^{(n)}\}_{k=1}^{M_n}$ of $\mathcal{Y}^n$ with $\frac{1}{M_n}\sum_{k=1}^{M_n} P^n(\mathcal{D}_k|x_k) > 1-\varepsilon_n$ for any ε_n, where $M_n > e^{n(C-\varepsilon_n)}$. then*

we can obtain a set $\{\tilde{x}_k^n\}_{k=1}^{\tilde{M}_n}$ *and a partition* $\{\mathcal{D}_k^{(n)}\}_{k=1}^{\tilde{M}_n}$ *with* $P^n(\tilde{\mathcal{D}}_k|\tilde{x}_k) > 1-\gamma_n$ *for all* k *for any* γ_n *for sufficiently large* n*, where* $\tilde{M}_n > e^{n(C-\gamma_n)}$.

Proof: Because $\frac{1}{M_n}\sum_{k=1}^{M_n} P^n(\mathcal{D}_k|x_k) > 1-\varepsilon_n$, then we have at least $\frac{M_n}{2}$ many x_k^n's with $P^n(\mathcal{D}_k|x_k) > 1-\frac{\varepsilon_n}{2}$ in $\{x_k^n\}_{k=1}^{M_n}$, otherwise $\frac{1}{M_n}\sum_{k=1}^{M_n} P^n(\mathcal{D}_k|x_k)$ will be no larger than $1-\varepsilon_n$. We let $\tilde{M}_n = \frac{M_n}{2} > e^{n(C-\varepsilon_n)-\log 2}$ and $\gamma_n = \frac{\varepsilon_n}{2} + \frac{\log 2}{n}$, thus the lemma is proved. ■

3.3 Information Stability, Information-Spectrum and Typicality

We will study coding problems relevant to general stochastic processes rather than i.i.d. ones in this thesis. Hence, it is necessary to extend the AEP and the joint typicality, as well as the Feinstein-type lemma, to more general cases. Historically, the AEP for Markov, stationary and ergodic sources were studied in [33, 37, 54]. A description of the AEP for general stochastic processes can be found in [53, Sec. 2.5]. Feinstein-type lemmas were also studied for stationary processes [4, 52] or more general ones without stationarity or ergodicity [4, 52].

The notions of information stability [7, 13, 55] and information-spectrum [4, 56, 57] are relevant to this topic. We introduce them as follows.

Definition 22 (Information Stability) *Given stochastic processes* $\mathbf{X} = \{X_t|t\in T\}$ *and* $\mathbf{Y} = \{Y_t|t\in T\}$ *with state spaces* $(\mathcal{X},\mathcal{B}_X)$ *and* $(\mathcal{Y},\mathcal{B}_Y)$ *and parameter set* T*,* $T=\mathbf{R}$ *or* $T=\mathbf{N}^+$*, then* $(\mathbf{X},\mathbf{Y})$ *is called* information stable *if information density* $i(x^t;y^t)$ *exists and mutual information between satisfying* X^t *and* Y^t *satisfies* $0 < I(X^t;Y^t) < \infty$ *for sufficiently large* t*, and if* $\frac{i(X^t;Y^t)}{I(X^t;Y^t)}$ *converges to* 1 *in probability when* $t\to\infty$*, namely, if*

$$\lim_{t\to\infty} \mathbf{P}\left\{\left|\frac{i(X^t;Y^t)}{I(X^t;Y^t)} - 1\right| > \varepsilon\right\} = 0 \tag{3.8}$$

for every $\varepsilon > 0$.

Definition 23 (Information Spectrum) *Given stochastic processes* $\mathbf{X} = (X_n)_{n=1}^{\infty}$ *and* $\mathbf{Y} = (Y_n)_{n=1}^{\infty}$ *with state spaces* $(\mathcal{X}, \mathcal{B}_X)$ *and* $(\mathcal{Y}, \mathcal{B}_Y)$*, then the* spectral inf-mutual information rate *between* $\mathbf{X}$ *and* $\mathbf{Y}$ *is defined as* [2]

$$\underline{I}(\mathbf{X};\mathbf{Y}) = \mathrm{p}\text{-}\liminf_{n\to\infty} \frac{1}{n} i(X^n;Y^n).$$

According to Defn. 23, we have

$$\lim_{n\to\infty} \mathbf{P}\{i(X^n;Y^n) < \underline{I}(\mathbf{X};\mathbf{Y}) - \varepsilon\} = 0 \tag{3.9}$$

for every $\varepsilon > 0$.

We will show how to apply these notions to achievability proofs for more general cases. For the case of the stationary and memoryless channel, from equations (3.2) and (3.4), we can see that typicality plays essential roles in the two approaches to the achievability proof of the channel coding theorem. We still use $\mathcal{T}^n(X,Y)$ to denote the jointly typical set, and define

$$\mathcal{T}^n(Y|x) = \{y^n \mid (x^n, y^n) \in \mathcal{T}^n(X,Y)\}.$$

Then equation (3.1) can be reformulated as

$$\lim_{n\to\infty} P_{XY}^n(\mathcal{T}(X,Y)) = 1, \tag{3.10}$$

equation (3.2) can be reformulated as

$$(P_X^n \times P_Y^n)(\mathcal{T}(X,Y)) < e^{-n(I(X;Y)-\varepsilon)} \tag{3.11}$$

for sufficiently large n, and equation (3.4) can be reformulated as

$$P_Y^n(\mathcal{T}^n(Y|x)) < e^{-n(I(X;Y)-\varepsilon)} \tag{3.12}$$

[2] The limit inferior in probability of $\mathbf{X}$ is defined as $\mathrm{p}\text{-}\liminf_{n\to\infty} X_n = \sup\{\beta \mid \lim_{n\to\infty} \mathbf{P}\{X_n < \beta\} = 0\}$.

for sufficiently large n. It is evident that (3.12) implies (3.11).

As for more general channel models, if we keep the structure of the proof in the last two sections, replace $I(X;Y)$ with $\lim_{n\to\infty}\frac{1}{n}I(X^n;Y^n)$, and redefine the jointly typical set as

$$\mathcal{T}^n(X,Y)=\left\{(x^n,y^n)\mid\left|\frac{i(x^n;y^n)}{I(X^n;Y^n)}-1\right|\leq\varepsilon\right\},$$

where $(\mathbf{X},\mathbf{Y})$ is information stable, then (3.10) holds following (3.8) in the definition of information stability, and (3.11) holds following the new definition of joint typicality, by a derivation similar to that of (3.2). Therefore, the achievability proof of a more general channel coding theorem is obtained. Similarly, we can also replace $I(X;Y)$ with $\underline{I}(\mathbf{X};\mathbf{Y})$, and redefine the jointly typical set as

$$\mathcal{T}^n(X,Y)=\{(x^n,y^n)\mid i(X^n;Y^n)\geq\underline{I}(\mathbf{X};\mathbf{Y})-\varepsilon\},$$

then (3.10) holds following (3.9), and (3.11) holds following the new typicality definition. This shed a light to a unified method of achievability proofs of general coding problems.

3.4 Conclusion

In this chapter, we have specified the relevance of different approaches to the achievability proof of the channel coding theorem. We have then proposed a unified method of the achievability proof, where the typicality plays a key role in this unified method, hence it is sufficient to study the typicality for various probability distributions, in order to solve channel coding problems with various general settings.

Chapter 4

Generalised Typicality Lemmas

4.1 A Generic Typicality

The conditional strong-typicality lemma was formally proposed in [1]. The proof in [1, Appd. 2A] followed the authors' typicality definition and probabilistic bounding methods.

The conditional strong-typicality lemma states a fact in probability that for each set sequence $\{\mathcal{T}_\varepsilon^{(n)}(X,Y) \in \mathcal{X}^n \times \mathcal{Y}^n\}_{n=1}^\infty$ satisfying $\lim_{n\to\infty} P_{XY}^{(n)}(\mathcal{T}_\varepsilon(X,Y)) = 1$, there exists a set sequence $\{\mathcal{U}^{(n)} \in \mathcal{X}^n\}_{n=1}^\infty$ satisfying $\lim_{n\to\infty} P_X^{(n)}(\mathcal{U}) = 1$, and the marginal set sequence $\{\mathcal{V}_x^{(n)} \in \mathcal{Y}^n\}_{n=1}^\infty$ determined by an arbitrary sequence $x^n \in \mathcal{U}^{(n)}$ satisfies $\lim_{n\to\infty} P_{Y|X}^{(n)}(\mathcal{V}_x|x) = 1$.

Intuitively, there should exist a set sequence $\{\mathcal{U}^{(n)} \in \mathcal{X}^n\}_{n=1}^\infty$ satisfying the aforementioned properties for any $\{\mathcal{A}^{(n)} \in \mathcal{X}^n \times \mathcal{Y}^n\}_{n=1}^\infty$ satisfying $\lim_{n\to\infty} P_{XY}^{(n)}(\mathcal{U}) = 1$; otherwise $\lim_{n\to\infty} P_{XY}^{(n)}(\mathcal{A}) = 1$ would not be established. However, there is no direct proof for this conjecture. The proof in [1, Appd. 2A] is not applicable for a general case, because it strictly relies on the strong typicality definition.

Motivated by those, we wish to study whether the property showed in the conditional strong-typicality lemma can be generalised and is irrelevant to the definition of typicality. If it is the case, we can obtain similar conditional typicality lemmas for various typicality definitions, which will find applications in general coding problems. We will first introduce the notion of generic typical set sequence, which extracts parts of asymptotic properties of conventional definitions typical sets.

Definition 24 (Generic typical set sequence) *Given a sequence of probability spaces* $\{(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, P^{(n)})\}_{n=1}^{\infty}$, *then we call* $\{\mathcal{A}^{(n)} \in \mathcal{B}^{(n)}\}_{n=1}^{\infty}$ *a* generic typical set sequence *w.r.t.* $\{P^{(n)}\}_{n=1}^{\infty}$ *if it holds asymptotically almost surely* [1] *in terms of* $\mathbf{P}(\mathcal{A}^{(n)}) = P^{(n)}(\mathcal{A})$, *namely,*

$$\lim_{n\to\infty} P^{(n)}(\mathcal{A}^{(n)}) = 1. \tag{4.1}$$

Remark 2 *In Lem. 5, we will set a sequence of the proposed generalised typical sets and prove that this sequence satisfies condition* (4.1). *Besides, in [59], Somekh-Baruch proposed a sequence* $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ *satisfying condition* (4.1).

Next, we will extend Lem. 4.3.2 in [2] to the general case. The original version of this lemma played a key role in the proof of Feinstein's fundamental lemma. The original version is based on the discrete probability space, and it is non-trivial to generalise its proof if we intend to obtain a similar conclusion based on a general probability space. We will prove the generalised version by using a theorem related to the conditional expectation.

Lemma 3 (Feinstein) *Fix* P_U *and* $P_{V|U}$ *and let* $P_{UV} = P_U \times P_{V|U}$. *Let* $\mathcal{A} \in \mathcal{U} \times \mathcal{V}$ *be a set such that* $P_{UV}(\mathcal{A}) > 1-\varepsilon$, *and* $\mathcal{E} \in \mathcal{U}$ *be a set such that* $P_U(\mathcal{E}) > 1-\delta$. *For each* $u \in \mathcal{U}$, *let* $\mathcal{A}_u = \{v \in \mathcal{V} \mid (u,v) \in \mathcal{A}\}$. *Let* $\mathcal{F} = \{u \in \mathcal{U} \mid P_{V|U}(\mathcal{A}_u|u) \geq 1-\gamma\}$, *then* $P_U(\mathcal{E} \cap \mathcal{F}) > 1-\delta-\frac{\varepsilon}{\gamma}$.

Proof: We set $X = \mathbf{E}[\chi_{\mathcal{A}}(U,V)|U]$, thus $X \leq 1$ and $\mathbf{E}(X) = \mathbf{E}[\chi_{\mathcal{A}}(U,V)] = P_{UV}(\mathcal{A})$. Then we have

$$\mathbf{P}\{X \geq 1-\gamma\} \geq 1 - \mathbf{P}\{1-X \geq \gamma\} \geq 1 - \frac{\mathbf{E}(1-X)}{\gamma} > 1 - \frac{\varepsilon}{\gamma}, \tag{4.2}$$

where the second to the last inequality follows the Markov inequality for non-negative real random variables.

On the other hand, according to the relation between conditional distribution and condi-

[1] In probability theory, an event sequence $\{E^{(n)}\}$ holds asymptotically almost surely if and only if $\lim_{n\to\infty} \mathbf{P}(E^{(n)}) = 1$ (see [58, p. 6]).

tional expectation (see [6][Thm. 2.19] [2]), we have

$$\mathbf{E}[\chi_{\mathcal{A}}(U,V)|U] = \int \chi_{\mathcal{A}}(U,v)\, \mathrm{d}P_{V|U}(v|U) \stackrel{\Delta}{=} h(U), \tag{4.3}$$

where

$$h(u) = \int \chi_{\mathcal{A}}(u,v)\, \mathrm{d}P_{V|U}(v|u) = P_{V|U}(\mathcal{A}_u|u). \tag{4.4}$$

Thus

$$\mathbf{P}\{X \geq 1-\gamma\} = P_U(\{u \mid h(u) \geq 1-\gamma\}) = P_U(\mathcal{F}) \tag{4.5}$$

From (4.2) and (4.5), we obtain

$$P_U(\mathcal{F}) > 1 - \frac{\varepsilon}{\gamma}. \tag{4.6}$$

Finally, we bound the probability of $\mathcal{E} \cap \mathcal{F}$ by

$$P_U(\mathcal{E} \cap \mathcal{F}) = P_U(\mathcal{E}) - P_U(\mathcal{E} \setminus \mathcal{F}) \geq P_U(\mathcal{E}) - P_U(\mathcal{U} \setminus \mathcal{F}) > 1 - \delta - \frac{\varepsilon}{\gamma}, \tag{4.7}$$

which establishes this lemma. ■

Based on Lem. 3 study an asymptotic property of the generic typical set sequence. This will be a fundamental conclusion of our main results.

Lemma 4 (Conditional Generic Typicality) *Given a sequence of probabilities* $\{P_X^{(n)}\}_{n=1}^{\infty}$, *a sequence of transition probabilities* $\{P_{Y|X}^{(n)}\}_{n=1}^{\infty}$ *and a sequence of product probability spaces* $\{(\mathcal{X}^{(n)} \times \mathcal{Y}^{(n)}, \mathcal{B}_X^{(n)} \otimes \mathcal{B}_Y^{(n)}, P_{XY}^{(n)})\}_{n=1}^{\infty}$ *where* $P_{XY}^{(n)} = P_X^{(n)} \times P_{Y|X}^{(n)}$ *for each n, and let* $\{\mathcal{A}^{(n)}(X,Y) \in \mathcal{B}_X^{(n)} \otimes \mathcal{B}_Y^{(n)}\}_{n=1}^{\infty}$ *be a generic typical set sequence w.r.t.* $\{P_{XY}^{(n)}\}_{n=1}^{\infty}$, *namely,*

$$\lim_{n\to\infty} P_{XY}^{(n)}(\mathcal{A}(X,Y)) = 1, \tag{4.8}$$

let $\mathcal{A}^{(n)}(Y|x) = \{y^{(n)} \in \mathcal{Y}^{(n)} \mid (x^{(n)}, y^{(n)}) \in \mathcal{A}^{(n)}(X,Y)\}$ *for each* $x^{(n)} \in \mathcal{X}^{(n)}$ *and* $\mathcal{A}^{(n)}(X|Y) =$

[2] It implies that given random variables U and V with the conditional distribution $P_{V|U}$ and a measurable function $f(U,V)$, then

$$\mathbf{E}[f(U,V)|U] = \int f(U,v)\, \mathrm{d}P_{V|U}(v|U).$$

$\{x^{(n)} \in \mathcal{X}^{(n)} \mid P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x) > 0\}$, *then* $\{\mathcal{A}^{(n)}(X,Y)\}_{n=1}^{\infty}$ *has the following properties that*

$$\lim_{n\to\infty} P_X^{(n)}(\mathcal{A}(X|Y)) = 1, \tag{4.9}$$

and there exists a sequence of $\{\mathcal{F}^{(n)} \in \mathcal{B}_X^{(n)}\}_{n=1}^{\infty}$ *such that*

$$\lim_{n\to\infty} P_X^{(n)}(\mathcal{F}) = \lim_{n\to\infty} P_X^{(n)}(\mathcal{A}(X|Y) \cap \mathcal{F}) = 1, \tag{4.10}$$

and for every $\{x^{(n)} \in \mathcal{F}^{(n)}\}_{n=1}^{\infty}$,

$$\lim_{n\to\infty} P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x) = 1. \tag{4.11}$$

Proof: For every n, we have

$$\begin{aligned} P_{XY}^{(n)}(\mathcal{A}(X,Y)) &= \int_{\mathcal{X}^{(n)}} P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x)\, \mathrm{d}P_X^{(n)}(x) \\ &= \int_{\mathcal{A}^{(n)}(X|Y)} P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x)\, \mathrm{d}P_X^{(n)}(x) \qquad (4.12) \\ &\leq \int_{\mathcal{A}^{(n)}(X|Y)} \mathrm{d}P_X^{(n)}(x) = P_X^{(n)}(\mathcal{A}(X|Y)) \qquad (4.13) \end{aligned}$$

From (4.8) and (4.13), we obtain that $\lim_{n\to\infty} P_X^{(n)}(\mathcal{A}(X|Y)) = 1$.

Because $\lim_{n\to\infty} P_{XY}^{(n)}(\mathcal{A}(X,Y)) = \lim_{n\to\infty} P_X^{(n)}(\mathcal{A}(X|Y)) = 1$, there must exist real sequences $\{\varepsilon_n\}_{n=1}^{\infty}$ and $\{\delta_n\}_{n=1}^{\infty}$ with $0 < \varepsilon_n, \delta < 1$ for every n, such that $\lim_{n\to\infty} \varepsilon_n = \lim_{n\to\infty} \delta_n = 0$ and

$$P_{XY}^{(n)}(\mathcal{A}(X,Y)) > 1-\varepsilon_n, P_X^{(n)}(\mathcal{A}(X|Y)) > 1-\delta_n$$

for every n. Let $\gamma_n = \sqrt{\varepsilon}_n$ for every n, then

$$\lim_{n\to\infty} \gamma_n = \lim_{n\to\infty} \frac{\varepsilon_n}{\gamma_n} = 0.$$

Under the assumption of Lem. 4, we make substitutions in Lem. 3 as follows. For each n let $\mathcal{U} = \mathcal{X}^{(n)}, \mathcal{V} = \mathcal{Y}^{(n)}, \varepsilon = \varepsilon_n, \delta = \delta_n . \gamma = \gamma_n, \mathcal{A} = \mathcal{A}^{(n)}(X,Y), \mathcal{E} = \mathcal{A}^{(n)}(X|Y), \mathcal{A}_u =$

$\mathcal{A}^{(n)}(Y|x)$, and

$$\mathcal{F} = \mathcal{F}^{(n)} = \{x^{(n)} \in \mathcal{X}^{(n)} \mid P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x) \geq 1 - \gamma_n\}.$$

Because $P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x) \leq 1$ for every n, we obtain

$$\lim_{n\to\infty} P_{Y|X}^{(n)}(\mathcal{A}(Y|x)|x) = 1$$

for every $\{x^{(n)} \in \mathcal{F}^{(n)}\}_{n=1}^{\infty}$. From Lem. 4 we have

$$P_X^{(n)}(\mathcal{A}^{(n)}(X|Y) \cap \mathcal{F}) > 1 - \frac{\varepsilon_n}{\gamma_n}$$

for every n. Because

$$P_X^{(n)}(\mathcal{A}^{(n)}(X|Y) \cap \mathcal{F}) \leq P_X^{(n)}(\mathcal{F}) \leq 1$$

for every n, we obtain

$$\lim_{n\to\infty} P_X^{(n)}(\mathcal{F}) = \lim_{n\to\infty} P_X^{(n)}(\mathcal{A}(X|Y) \cap \mathcal{F}) = 1.$$

■

4.1.1 A Necessary Presumption of Lemma 3

It is evident that the sequence of robust typicality sets constructed in the conventional conditional typicality lemma [1, Section 2.5] satisfies condition (4.8) in Lem. 3. In [1, Problem 2.17], it is shown that the conditional typicality lemma is not necessarily established for a given $\{x^n\}_{n=1}^{\infty}$, where $x^n \in \mathcal{T}_\varepsilon^n(X)$ (for robust typicality) for all $n \in \mathbb{Z}^+$. We can prove that the given x^n actually falls out of $\mathcal{T}_\varepsilon^n(X|Y)$ asymptotically.

In the above problem, P_{XY} is given as the production probability measure of two Bernoulli measures $\mathrm{B}(1/2)$, and x^n is given as a binary sequence with k_n 1's followed by $(n-k_n)$ 0's, where $k_n = \lfloor (n/2)(1+\varepsilon) \rfloor$. According to the definition of robust jointly typical set, x^n will fall out of $\mathcal{T}_\varepsilon^n(X|Y)$ if $\lceil k_n/2 \rceil / n - 1/4 > \varepsilon/4$. Let $\{n_l'\}$ be a subsequence of all n's

satisfying $n = 4l_n + 1$. Assume that $k_n = (n/2)(1+\varepsilon) - \delta_n$ and $\lceil k_n/2 \rceil = k_n/2 + \gamma_n$. Because $k_n = 2l_n + 1/2 + (2l_n + 1/2)\varepsilon - \delta_n$, $(2l_n + 1/2)\varepsilon - \delta_n + 1/2$ is an integer denoted by N_{l_n}. Then because $\lceil k_n/2 \rceil = l_n + N_{l_n}/2 + \gamma_n$, $\gamma_n = 1/2$ if N_{l_n} is odd, thus $\lceil k_n/2 \rceil / n - 1/4 = \varepsilon/4 + (1-\delta_n)/(2n) > \varepsilon/4$. The range $0 < \varepsilon < 1$ implies that there exist infinitely many odd N_{l_n}'s in $\{N_{l_n}\}$. Hence, there exist infinitely many n's such that $x^n \notin \mathcal{T}_\varepsilon^n(X^n|Y^n)$.

4.2 A Generalised Multivariate Typicality

Although the generic typicality provides the conditional property, it is not related to any information quantities in its definition, thus we can not directly obtain relevant lemmas that is applicable in information theoretical problems. In this chapter, we will introduce a generalised weak typical set, which can be used to construct a generic typical set sequence. We will then provide typicality lemmas that can then be applied to coding problems.

Conventionally, the weakly typical set $\mathcal{A}_\varepsilon^n(X)$, also called the entropy-typical set, with respect to a probability distribution $P_X^n = \prod_{k=1}^n P_X$ of which the entropy is finite, is defined as the set of all $\mathcal{X}^n$-valued sequences such that

$$|\frac{1}{n}h(x^n) - H(X)| \leq \varepsilon, \tag{4.14}$$

where

$$h(x^n) = \sum_{j=1}^n \log \frac{1}{P_X(x_j)} \tag{4.15}$$

is the entropy function of x^n with respect to P_X^n, and $H(X)$ is the entropy with respect to P_X.

There also exists weak typicality defined for pairs of sequences. This is called joint typicality. In [4, Sec. 3.1], the mutual information density between two random variables, which is defined as

$$i(x;y) = \log \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X \times P_Y}(x,y), \tag{4.16}$$

plays a key role in the definition of joint weak typicality. The jointly weakly typical set $\mathcal{A}_\varepsilon^n(XY)$ with respect to $P_{XY}^n = \prod_k = 1^n P_{XY}$ is defined as the set of all sequence pairs (x^n, y^n)

such that

$$|\frac{1}{n}i(x^n,y^n)-I(X;Y)|\leq\varepsilon, \tag{4.17}$$

where $I(X;Y)$ is the mutual information between X and Y. The asymptotic property of $i(X^n;Y^n)$, or more specifically, the LLN is the fundamental of definitions of typicality and joint typicality. Inspired by [4, Sec. 3.1], we proposed a jointly weakly typicality with respect to the distribution of a general stochastic process, by replacing the mutual information rate with the spectral mutual information which is an asymptotic measure of the relevance of two stochastic processes [60].

As is generally understood, entropy is a measure of the uncertainty of a single source, and mutual information is a measure of the relevance of two sources. It is natural for us to obtain a multivariate typicality definition based on the relative entropy, which is a measure of multiple dependence and relevance of multiple sources (cf. [5]).

First we introduce the the notions of relative entropy and relative entropy density (see [54]).

Definition 25 (Relative Entropy Density) *Given P_X and Q_X satisfying $P_X \ll Q_X$, then the relative entropy density $d_{P_X||Q_X}(x)$ is defined as*

$$d_{P_X||Q_X}(x)=\log\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(x). \tag{4.18}$$

The relative entropy is the expectation of the relative entropy density, i.e.,

$$D(P_X||Q_X)=\int d_{P_X||Q_X}(x)\ \mathrm{d}P_X(x). \tag{4.19}$$

We also introduce the spectral relative entropy rates following the information spectrum fashion (see [4, Sec. 4.1]). This is an extension of relative entropy for general stochastic processes.

Definition 26 (Spectral Conditional Inf- and Sup-Relative Entropy Rate) *With given $P_{\mathbf{X}}$ and $Q_{\mathbf{X}}$ and let $\mathbf{X}\sim P_{\mathbf{X}}$, we define the spectral inf- (or sup-)relative entropy rate as limit*

inferior (or superior) in probability[3]*, i.e.,*

$$\underline{D}(P_{\mathbf{X}}||Q_{\mathbf{X}}) = \mathrm{p}-\liminf_{n\to\infty} \frac{1}{n} d^n_{P_X||Q_X}(X), \tag{4.20}$$

$$\overline{D}(P_{\mathbf{X}}||Q_{\mathbf{X}}) = \mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} d^n_{P_X||Q_X}(X). \tag{4.21}$$

Similar to joint weak typicality, the multivariate typicality is defined based on the distance between the relative entropy rate of a tuplet of sequences and the spectral relative entropy rate.

Definition 27 (Multivariate Inf- and Sup-Typicality) *Given a probability measure $P_{\mathbf{X}_\mathcal{K}}$ where $\mathcal{K}$ is an index set, let $\mathcal{S} \subsetneq \mathcal{K}$ and $\mathcal{S}^\complement = \mathcal{K} \setminus \mathcal{S}$, we define the multivariate ε-inf-typicality sequence w.r.t. $P_{\mathbf{X}_\mathcal{K}}$ as any $x^n_\mathcal{K}$ satisfying*

$$\left| \frac{1}{n} d^n_{P_{X_{\mathcal{S}^\complement}|X_\mathcal{S}}||\prod_{k\in\mathcal{S}^\complement} P_{X_k|X_\mathcal{S}}}(x_\mathcal{K}) - \underline{D}(P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}}||\prod_{k\in\mathcal{S}^\complement} P_{\mathbf{X}_k|\mathbf{X}_\mathcal{S}}|P_{\mathbf{X}_\mathcal{S}}) \right| \leq \varepsilon \tag{4.22}$$

for all $\mathcal{S} \subsetneq \mathcal{K}$. Similarly, we define the multivariate ε-sup-typicality sequence w.r.t. $P_{\mathbf{X}_\mathcal{K}}$ as any $x^n_\mathcal{K}$ satisfying

$$\left| \frac{1}{n} d^n_{P_{X_{\mathcal{S}^\complement}|X_\mathcal{S}}||\prod_{k\in\mathcal{S}^\complement} P_{X_k|X_\mathcal{S}}}(x_\mathcal{K}) - \overline{D}(P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}}||\prod_{k\in\mathcal{S}^\complement} P_{\mathbf{X}_k|\mathbf{X}_\mathcal{S}}|P_{\mathbf{X}_\mathcal{S}}) \right| \leq \varepsilon \tag{4.23}$$

for all $\mathcal{S} \subsetneq \mathcal{K}$. The multivariate inf- and sup-typicality set are denoted as $\underline{\mathcal{T}}^n_\varepsilon(X_\mathcal{K})$ and $\overline{\mathcal{T}}^n_\varepsilon(X_\mathcal{K})$, respectively.

The conditional typicality lemma plays significant roles in proofs of some coding problems. In order to generalise the lemma to multivariate cases, we first introduce the notion of conditional multivariate typicality.

Definition 28 (Conditional Multivariate Sup- and Inf-Typicality) *Under the assumptions of Defn. 27, for any $\mathcal{S} \subsetneq \mathcal{K}$, the conditionally sup- and inf-typical sets $\overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S})$ and*

[3]The limit inferior in probability of $\mathbf{X}$ is defined as $\mathrm{p}-\liminf_{n\to\infty} X_n = \sup\{\beta \mid \lim_{n\to\infty} \mathbf{P}\{X_n < \beta\} = 0\}$, and the limit superior in probability of $\mathbf{X}$ is defined as $\mathrm{p}-\limsup_{n\to\infty} X_n = \inf\{\alpha \mid \lim_{n\to\infty} \mathbf{P}\{X_n > \alpha\} = 0\}$.

$\underline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_{\mathcal{S}})$ with respect to a general joint probability distribution $P_{X^n_{\mathcal{K}}}$ and a given sequence $x^n_{\mathcal{S}}$ are defined as

$$\underline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_{\mathcal{S}}) = \{x^n_{\mathcal{S}^\complement} \in \mathcal{X}^n_{\mathcal{S}^\complement} | \Pi(x^n_{\mathcal{S}}, x^n_{\mathcal{S}^\complement}) \in \underline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{K}})\},$$
$$\overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_{\mathcal{S}}) = \{x^n_{\mathcal{S}^\complement} \in \mathcal{X}^n_{\mathcal{S}^\complement} | \Pi(x^n_{\mathcal{S}}, x^n_{\mathcal{S}^\complement}) \in \overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{K}})\},$$

where $\mathcal{S}^\complement = \mathcal{K} \setminus \mathcal{S}$ and $\Pi(\cdot,\cdot)$ is a permutation function that rearranges the elements of $(x^n_{\mathcal{S}}, x^n_{\mathcal{S}^\complement})$ according to $\mathcal{K}$.

4.3 Generalised Typicality Lemmas

Following Lem. 3, we will derive several elementary typicality lemmas, including the conditional typicality lemma and the strong Markov lemma, for the generalised multivariate typicality. Joint typicality and covering lemmas will be further results based on the conditional typicality lemma. We will collectively call our results as typicality lemmas. The generalised multivariate typicality lemmas can be applied to multivariate coding problems with general settings, whereas their counterparts in discrete or memoryless cases are applied to more restricted coding problems.

4.3.1 Generalised Conditional and Joint Typicality Lemmas

The following lemma is a direct corollary of Lem. 3 and the definition of the multivariate typicality.

Lemma 5 (Multivariate Conditional Typicality) *Given a probability measure $P_{\mathbf{X}_{\mathcal{K}}}$ where $\mathcal{K}$ is an index set, let $\mathcal{S} \subset \mathcal{K}$ and $\mathcal{S}^\complement = \mathcal{K} \setminus \mathcal{S}$, we set*

$$\underline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}}|X_{\mathcal{S}^\complement}) = \{x^n_{\mathcal{S}} \in \mathcal{X}^n_{\mathcal{S}} | P^n_{X_{\mathcal{S}^\complement}|X_{\mathcal{S}}}(\underline{\mathcal{T}}_\varepsilon(X_{\mathcal{S}^\complement}|x_{\mathcal{S}})|x_{\mathcal{S}}) > 0\}.$$

Then

$$\lim_{n\to\infty} P^n_{X_{\mathcal{S}}}(\underline{\mathcal{T}}_\varepsilon(X_{\mathcal{S}}|X_{\mathcal{S}^\complement})) = 1, \tag{4.24}$$

and there exists a sequence of $\{\mathcal{F}^{(n)} \in \mathcal{B}^{(n)}_{X_\mathcal{S}}\}_{n=1}^{\infty}$ *such that*

$$\lim_{n\to\infty} P^{(n)}_{X_\mathcal{S}}(\mathcal{F}) = 1, \tag{4.25}$$

and for every n and every $\{x^{(n)}_\mathcal{S} \in \mathcal{F}^{(n)}\}_{n=1}^{\infty}$,

$$\lim_{n\to\infty} P^n_{X_{\mathcal{S}^\complement}|X_\mathcal{S}}(\underline{\mathcal{T}}_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S})|x_\mathcal{S}) = 1. \tag{4.26}$$

The following multivariate joint typicality lemma is a counterpart of [1, Problem 2.16], which is a corollary of the discrete and i.i.d. version of conditional typicality lemma [1]. Similarly, the multivariate joint typicality lemma is derived from the multivariate conditional typicality lemma.

Lemma 6 (Multivariate Joint Sup-Typicality) *Given a probability measure* $P_{\mathbf{X}_\mathcal{K}}$ *where* $\mathcal{K}$ *is an index set, we set a sequence of sets* $\{\mathcal{H}_n \in \mathcal{X}^n_\mathcal{K}\}_{n=1}^{\infty}$ *satisfying* $\lim_{n\to\infty} P_{\mathbf{X}_\mathcal{K}}(\mathcal{H}_n) = 1$, *let* $\mathcal{S} \subset \mathcal{K}, \mathcal{S}^\complement = \mathcal{K} \setminus \mathcal{S}$ *and* $\mathcal{H}^n(x_{\mathcal{S}^\complement}) = \{x^n_\mathcal{S} | (x^n_\mathcal{S}, x^n_{\mathcal{S}^\complement}) \in \mathcal{H}_n\}$, *then for sufficiently large n and all* $x^n_\mathcal{S} \in \overline{\mathcal{T}}^n_\varepsilon(X_\mathcal{S}|X_{\mathcal{S}^\complement})$,

$$\begin{aligned}
&(1-\varepsilon)\exp(-n(\overline{D}(P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}} || \prod_{k\in\mathcal{S}^\complement} P_{\mathbf{X}_k|\mathbf{X}_\mathcal{S}}|P_{\mathbf{X}_\mathcal{S}}) + \varepsilon)) && (4.27)\\
\leq &(\prod_{k\in\mathcal{S}^\complement} P^n_{X_k|X_\mathcal{S}})(\overline{\mathcal{T}}_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S}) \cap \mathcal{H}(x_{\mathcal{S}^\complement}))\\
\leq &\exp(-n(\overline{D}(P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}} || \prod_{k\in\mathcal{S}^\complement} P_{\mathbf{X}_k|\mathbf{X}_\mathcal{S}}|P_{\mathbf{X}_\mathcal{S}}) - \varepsilon)). && (4.28)
\end{aligned}$$

Proof: Following Defn. 27 and multivariate conditional sup-typicality lemma, then we have

$$\begin{aligned}
(\prod_{k\in\mathcal{S}^\complement} P^n_{X_k|X_\mathcal{S}})(\overline{\mathcal{T}}_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S}) \cap \mathcal{H}(x_{\mathcal{S}^\complement})) &= \int_{\overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S}) \cap \mathcal{H}^n(x_{\mathcal{S}^\complement})} \mathrm{d} \prod_{k\in\mathcal{S}^\complement} P^n_{X_k|X_\mathcal{S}}\\
&= \int_{\overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_\mathcal{S}) \cap \mathcal{H}^n(x_{\mathcal{S}^\complement})} \frac{\mathrm{d}\prod_{k\in\mathcal{S}^\complement} P^n_{X_k|X_\mathcal{S}}}{\mathrm{d}P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}}} \mathrm{d}P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_\mathcal{S}},
\end{aligned}$$

which establishes the lemma. ■

4.3.2 A Generalised Multivariate Covering Lemma

In this section, we will introduce a multivariate version of the covering lemma based on the generalised multivariate typicality. This lemma will be applicable in multi-terminal coding problems.

Lemma 7 (Multivariate Covering) *Given a probability measure $P_{\mathbf{X}_\mathcal{K}}$ where $\mathcal{K}=\{0,1,\cdots,K\}$, let*

$$(X_0^n, X_1^n(m_1),\cdots,X_K^n(m_K)) \sim P_{X_0}^n \times \prod_{k=1}^{K} P_{X_k|X_0}^n,$$

where $m_k \in \mathcal{M}_{kn}$ with $|\mathcal{M}_{kn}| = e^{nR_k}$ for all $k \in \mathcal{K}\setminus\{0\}$ and for all n. If

$$\sum_{k\in\mathcal{S}} R_k > \overline{D}(P_{\mathbf{X}_\mathcal{S}|\mathbf{X}_0}||\prod_{k\in\mathcal{S}} P_{\mathbf{X}_k|\mathbf{X}_0}|P_{\mathbf{X}_0}), \tag{4.29}$$

for all $\mathcal{S}\subseteq\mathcal{K}\setminus\{0\}$ and $\mathcal{S}\neq\emptyset$, then there exists an $\varepsilon>0$ such that

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcap_{\prod_{k\in\mathcal{K}}\mathcal{M}_k}\bigcap_{\mathcal{L}\subseteq\mathcal{K}} \{(X_0^n, X_{l_1}^n(m_{l_1}),\cdots,X_{l_L}^n(m_{l_L})) \notin \overline{\mathcal{T}}_\varepsilon^n(X_\mathcal{L})\}\right) = 0,$$

where $\mathcal{L}=\{l_1,\cdots,l_L\}$.

Proof: We employ the combinatorial counting and bounding techniques from [1] and [22]. Similar to [1, Appendix 8A], we set

$$\mathcal{M}_n = \{m_\mathcal{K} \in \prod_{k\in\mathcal{K}}\mathcal{M}_k \mid \bigcap_{\mathcal{L}\subseteq\mathcal{K}} \{(X_0^n, X_{l_1}^n(m_{l_1}),\cdots,X_{l_L}^n(m_{l_L})) \in \overline{\mathcal{T}}_\varepsilon^n(X_\mathcal{K})\},$$

and obtain that

$$\mathbf{P}\{|\mathcal{M}_n|=0\} \le \frac{\mathbf{D}(|\mathcal{M}_n|)}{\mathbf{E}^2(|\mathcal{M}_n|)}.$$

We can express $|\mathcal{M}_n|$ as

$$|\mathcal{M}_n| = \sum_{m_\mathcal{K}\in\prod_{k\in\mathcal{K}}\mathcal{M}_k} \chi_\varepsilon(m_\mathcal{K}),$$

where the indicator function is $\chi_\varepsilon(m_\mathcal{K})$ is defined as $\chi_{\overline{\mathcal{T}}^n_\varepsilon(X_\mathcal{K})}(X_0^n, X_1^n(m_1), \cdots, X_K^n(m_K))$. Hence we obtain that

$$|\mathcal{M}_n|^2 = \sum_{m_\mathcal{K} \in \prod\limits_{k\in\mathcal{K}} \mathcal{M}_k} \chi_\varepsilon(m_\mathcal{K}) + \sum_{\substack{m'_\mathcal{K} \neq m_\mathcal{K} \\ m'_\mathcal{K} \in \prod\limits_{k\in\mathcal{K}} \mathcal{M}_k}} \sum_{m_\mathcal{K} \in \prod\limits_{k\in\mathcal{K}} \mathcal{M}_k} \chi_\varepsilon(m_\mathcal{K})\chi_\varepsilon(m'_\mathcal{K}).$$

Then we calculate for any $m'_\mathcal{K} \neq m_\mathcal{K}$ with $\mathcal{S} = \{k \in \mathcal{K} | m_k = m'_k\}$ and $\mathcal{S}^\complement = \mathcal{K} \setminus \{0\} \setminus \mathcal{S}$ that

$$\begin{aligned}
&\mathbf{E}[\chi_\varepsilon(m_\mathcal{K})\chi_\varepsilon(m'_\mathcal{K})] \\
=& \int_{\mathcal{T}^n_\varepsilon(X_0|X_{\mathcal{K}\setminus\{0\}})} \int_{\mathcal{T}^n_\varepsilon(X_\mathcal{S}|x_0 X_{\mathcal{S}^\complement})} \Big(\int_{\mathcal{T}^n_\varepsilon(X_{\mathcal{S}^\complement}|x_0 x_\mathcal{S})} \mathrm{d}P^n_{X_{\mathcal{S}^\complement}|X_0 X_\mathcal{S}} \Big)^2 \mathrm{d}P^n_{X_\mathcal{S}|X_0} \mathrm{d}P^n_{X_0} \\
\leq& \int_{\mathcal{T}^n_\varepsilon(X_0|X_{\mathcal{K}\setminus\{0\}})} \Big(\big(\prod_{k\in\mathcal{S}^\complement} P^n_{X_k|X_0} \big)\big(\overline{\mathcal{T}}^n_\varepsilon(X_{\mathcal{K}\setminus\{0\}}|x_0 x_\mathcal{S}) \big) \Big)^2 \Big(\prod_{k\in\mathcal{S}} P^n_{X_k|X_0} \Big)\big(\overline{\mathcal{T}}^n_\varepsilon(X_\mathcal{S}|x_0) \big) \, \mathrm{d}P^n_{X_0} \\
\leq& \exp\Big(-n\big(\overline{D}(P_{\mathbf{X}_\mathcal{S}|\mathbf{X}_0} || \prod_{k\in\mathcal{S}} P_{\mathbf{X}_k|\mathbf{X}_0} | P_{\mathbf{X}_0}) + 2\overline{D}(P_{\mathbf{X}_{\mathcal{S}^\complement}|\mathbf{X}_0\mathbf{X}_\mathcal{S}} || \prod_{k\in\mathcal{S}^\complement} P_{\mathbf{X}_k|\mathbf{X}_0} | P_{\mathbf{X}_0\mathbf{X}_\mathcal{S}}) - 3\varepsilon \big) \Big),
\end{aligned}$$

where the last step follows the multivariate joint sup-typicality lemma. We can then prove this lemma similarly to last steps in [1, Appendix 8A]. ■

In the discrete version of multivariate lemma [22], the constraint on $\sum\limits_{k\in\mathcal{S}} R_k$ is in terms of calculation of entropy terms, which are not well defined for general cases. From eqn. (4.29), we can see the entropy terms has been circumvented. This is because we introduce the relative entropy rate in our multivariate typicality definition. We can also expect that both the expression and the proof will be much more complicated if we use mutual information density instead in the typicality definition.

4.3.3 A Generalised Markov Lemma

In this section, we will generalise the strong version of Markov lemma based on our generalised multivariate typicality. Similar to its special case that is the conditional typicality lemma, the strong Markov lemma can also be derived from Lem. 3.

Lemma 8 (Markov) *Given a probability measure $P_{\mathbf{XYZ}}$ satisfying that $P^n_{XYZ} = P^n_{X|Y} \times P^n_Y \times$*

$P^n_{Z|Y}$ for all n, let $(x^n, y^n) \in \underline{\mathcal{T}}^n_\varepsilon(XY|Z)$[4] and $Q_{Z^n|Y^n}$ be a conditional probability satisfying that

$$\lim_{n\to\infty} Q^n_{Z|Y} = \lim_{n\to\infty} P^n_{Z|Y} \tag{4.30}$$

almost everywhere. Then

$$\lim_{n\to\infty} Q^n_{Z|Y}(\underline{\mathcal{T}}_\varepsilon(Z|xy)|y) = 1. \tag{4.31}$$

Proof: From Lem. 3 and the Markovity assumption on the joint probability, it is apparent that

$$\lim_{n\to\infty} P^n_{Z|Y}(\underline{\mathcal{T}}_\varepsilon(Z|xy)|y) = 1, \tag{4.32}$$

which establishes this lemma due to the convergence presumption on $Q^n_{Z|Y}$. ■

Remark 3 *If we set $Q^n_{Z|Y} = P^n_{Z|Y}$ in Lem. 8, we will obtain a weak version of Markov Lemma, which is an extension of the weak Markov lemma based on strong typicality and for the memoryless case [11, Lem. 15.8.1].*

Remark 4 *In the strong version of the Markov Lemma (see [1, Lem. 12.1]) based on the strong typicality and finite alphabet, the second condition implies the convergence presumption in Lem. 8.*

4.4 Bivariate Typicality Lemmas

For the convenience of applications in point-to-point coding problems, we will introduce a couple of bivariate typicality lemmas based on our generalized typicality definition. Most of the proofs are based on the proofs of the multivariate typicality lemmas in Sec. 4.3.

Definition 29 (Sup- and Inf-Typicality) *A sequence pair (x^n, y^n) is called* jointly ε-inf-typical *with respect to a general probability distribution P^n_{XY} if*

$$\left| \frac{1}{n} i^n(x,y) - \underline{I}(\mathbf{X};\mathbf{Y}) \right| \leq \varepsilon,$$

[4] In this notation of typical set, we omit the comma between X and Y. We might use similar notations later in this thesis when there would be no confusion.

when $P_{Y^n|X^n}(\cdot|x^n)$ is absolutely continuous with respect to P_Y^n, and the spectral inf-mutual information rate $\underline{I}(X^n;Y^n)$ *(see [4, Def. 3.2.1]) is defined as the limit inferior in probability of $\frac{1}{n}i^n(x,y)$.*

Similarly, a sequence pair (x^n,y^n) is called jointly ε-sup-typical *with respect to P_{XY}^n if*

$$\left|\frac{1}{n}i^n(x,y)-\bar{I}(\mathbf{X};\mathbf{Y})\right|\leq\varepsilon,$$

where the spectral sup-mutual information rate $\bar{I}(X^n;Y^n)$ *(see [4, Def. 3.5.2]) is defined as the limit superior in probability of $\frac{1}{n}i^n(x,y)$.*

Let $\underline{\mathcal{T}}_\varepsilon^n(XY)$ and $\overline{\mathcal{T}}_\varepsilon^n(XY)$ denote the set of all general jointly ε-inf-typical sequences and the set of all general jointly ε-sup-typical sequences, respectively.

Remark 5 *Different from [15] and [16], the proposed generalised typicality in Definition 29 is not based on the measure of a Polish or a Borel space, which introduces a metric.*

Definition 30 (Conditionally Sup- and Inf-Typicality) *The conditionally sup- and inf-typical sets $\overline{\mathcal{T}}_\varepsilon^n(Y|x)$ and $\underline{\mathcal{T}}_\varepsilon^n(Y|x)$ with respect to a general joint probability distribution P_{XY}^n and a given sequence x^n are defined as*

$$\overline{\mathcal{T}}_\varepsilon^n(Y|x)=\{y^n\in\mathcal{Y}^n|(x^n,y^n)\in\overline{\mathcal{T}}_\varepsilon^n(XY)\},$$
$$\underline{\mathcal{T}}_\varepsilon^n(Y|x)=\{y^n\in\mathcal{Y}^n|(x^n,y^n)\in\underline{\mathcal{T}}_\varepsilon^n(XY)\}.$$

Lemma 9 (Conditional Typicality) *Given $P_{\mathbf{XY}}$, we set*

$$\underline{\mathcal{T}}_\varepsilon^n(X|Y)=\{x^n\in\mathcal{X}^n|P_{Y|X}^n(\underline{\mathcal{T}}_\varepsilon(Y|x)|x)>0\},$$
$$\overline{\mathcal{T}}_\varepsilon^n(X|Y)=\{x^n\in\mathcal{X}^n|P_{Y|X}^n(\overline{\mathcal{T}}_\varepsilon(Y|x)|x)>0\}.$$

Then

$$\lim_{n\to\infty}P_X^n(\underline{\mathcal{T}}_\varepsilon^n(X|Y))=\lim_{n\to\infty}P_X^n(\overline{\mathcal{T}}_\varepsilon^n(X|Y))=1;\ \lim_{n\to\infty}P_{Y|X}^n(\underline{\mathcal{T}}_\varepsilon(Y|x)|x)=1,$$

for any $\{x^n\}_{n=1}^{\infty}$*, where* $x^n \in \underline{\mathcal{T}}_{\varepsilon}^n(X|Y)$ *for all n, and*

$$\lim_{n\to\infty} P_{Y|X}^n(\overline{\mathcal{T}}_{\varepsilon}(Y|x)|x) = 1,$$

for any $\{x^n\}_{n=1}^{\infty}$*, where* $x^n \in \overline{\mathcal{T}}_{\varepsilon}^n(X|Y)$ *for all n.*

Proof: Similar to Lem. 5, this lemma is a corollary of Lem. 3. ■

Remark 6 *Under the condition of Lemma 9, if given an* x^n *and let* $Y^n \sim P_{Y^n|X^n}(\cdot|x^n)$*, then we have*

$$P_{Y|X}^n(\underline{\mathcal{T}}_{\varepsilon}(Y|x)|x) = \mathbf{P}\{(x^n, Y^n) \in \underline{\mathcal{T}}_{\varepsilon}^n(XY)\},$$
$$P_{Y|X}^n(\overline{\mathcal{T}}_{\varepsilon}(Y|x)|x), = \mathbf{P}\{(x^n, Y^n) \in \overline{\mathcal{T}}_{\varepsilon}^n(XY)\}.$$

Lemma 10 (Joint Inf-Typicality) *Given* $P_{\mathbf{XY}}$*, for all n and for all* $x^n \in \underline{\mathcal{T}}_{\varepsilon}(X^n|Y^n)$*,*

$$e^{-n(\underline{I}(\mathbf{X};\mathbf{Y})+\varepsilon)} \le P_Y^n(\underline{\mathcal{T}}_{\varepsilon}(Y|x)) \le e^{-n(\underline{I}(\mathbf{X};\mathbf{Y})-\varepsilon)};$$
$$e^{-n(\underline{I}(\mathbf{X};\mathbf{Y})+\varepsilon)} \le (P_X^n \times P_Y^n)(\underline{\mathcal{T}}_{\varepsilon}(XY)) \le e^{-n(\underline{I}(\mathbf{X};\mathbf{Y})-\varepsilon)}.$$

Proof: Let $X_{\mathcal{S}} = X$ and $X_{\mathcal{SC}} = Y$ in Lem. 6, then this conclusion is obtained. ■

Remark 7 *Under the condition of Lemma 10, if given* $\bar{Y}^n \sim P_{\bar{Y}^n|X^n}(\cdot|x^n)$ *where* $P_{\bar{Y}^n|X^n}(\cdot|x^n) = P_Y^n(\cdot)$ *and* $(\tilde{X}^n, \tilde{Y}^n) \sim P_{\tilde{X}^n\tilde{Y}^n}$ *where* $P_{\tilde{X}^n\tilde{Y}^n} = P_X^n \times P_Y^n$*, we have*

$$P_Y^n(\underline{\mathcal{T}}_{\varepsilon}(Y|x)) = \mathbf{P}\{(x^n, \bar{Y}^n) \in \underline{\mathcal{T}}_{\varepsilon}^n(XY)\},$$
$$(P_X^n \times P_Y^n)(\underline{\mathcal{T}}_{\varepsilon}(XY)) = \mathbf{P}\{(\tilde{X}^n, \tilde{Y}^n) \in \underline{\mathcal{T}}_{\varepsilon}(X^nY^n)\}.$$

Similarly, we obtain the joint sup-typicality lemma.

Lemma 11 (Joint Sup-Typicality) *Given* $P_{\mathbf{XY}}$*, for all n and for all* $x^n \in \overline{\mathcal{T}}_{\varepsilon}(X^n|Y^n)$

$$e^{-n(\bar{I}(\mathbf{X};\mathbf{Y})+\varepsilon)} \le P_Y^n(\overline{\mathcal{T}}_{\varepsilon}^n(Y|x)) \le e^{-n(\bar{I}(\mathbf{X};\mathbf{Y})-\varepsilon)},$$
$$e^{-n(\bar{I}(\mathbf{X};\mathbf{Y})+\varepsilon)} \le (P_X^n \times P_Y^n)(\overline{\mathcal{T}}_{\varepsilon}(XY)) \le e^{-n(\bar{I}(\mathbf{X};\mathbf{Y})-\varepsilon)}.$$

As counterparts to the conventional method of typical sequences, we will first prove the following two lemmas which will be used in our achievability proof.

Lemma 12 (Packing) *Given $P_{\mathbf{XY}}$, for all $n \in \mathbb{Z}^+$, let $\tilde{Y}^n \sim P_{\tilde{Y}^n}$, which is not necessarily equal to P_Y^n, $X^n(m) \sim P_X^n, m \in \mathcal{M}_n$ with $|\mathcal{M}_n| = e^{nR}$, and $X^n(m)$ are independent of $\tilde{Y}^n$ for all $m \in \mathcal{M}_n$. If $R < \underline{I}(\mathbf{X};\mathbf{Y})$, then there exists an $\varepsilon > 0$ such that*

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcup_{m\in\mathcal{M}_n} \{(X^n(m),\tilde{Y}^n) \in \underline{\mathcal{T}}_\varepsilon^n(XY)\}\right) = 0.$$

Proof: From the union bound and the joint inf-typicality lemma (Lemma 10), we have

$$\begin{aligned}\mathbf{P}\left(\bigcup_{m\in\mathcal{M}_n} \{(X^n(m),\tilde{Y}^n) \in \underline{\mathcal{T}}_\varepsilon^n(XY)\}\right) &= \sum_{m\in\mathcal{M}_n} \mathbf{P}\{(X^n(m),\tilde{Y}^n) \in \underline{\mathcal{T}}_\varepsilon^n(XY)\} \\ &\le \sum_{m\in\mathcal{M}_n} \int_{\underline{\mathcal{T}}_\varepsilon^{Y^n|X^n}} \mathbf{P}\{(X^n(m),\tilde{y}^n) \in \underline{\mathcal{T}}_\varepsilon^n(XY)\} dP_{\tilde{Y}^n}(\tilde{y}^n) \\ &\le |\mathcal{M}_n| e^{-n(\underline{I}(\mathbf{X};\mathbf{Y})-\varepsilon)} \le e^{n(R-\underline{I}(\mathbf{X};\mathbf{Y})+\varepsilon)},\end{aligned}$$

which establishes the lemma. ■

Lemma 13 (Covering) *Given $P_{\mathbf{XY}}$, for all $n \in \mathbb{Z}^+$, let $X^n \sim P_X^n$, $Y^n(m) \sim P_Y^n, m \in \mathcal{M}_n$ with $|\mathcal{M}_n| = e^{nR}$, and X^n and $Y^n(m)$'s are independent of each other. If $R > \bar{I}(\mathbf{X};\mathbf{Y})$, then there exists an ε such that*

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcap_{m\in\mathcal{M}_n} \{(X^n,Y^n(m)) \notin \overline{\mathcal{T}}_\varepsilon^n(XY)\}\right) = 0.$$

Proof: From the joint sup-typicality lemma (Lemma 11) and the inequality $(1-y)^n \le e^{-yn}$ for $0 \le y \le 1$ and $n \ge 0$ [11, Lemma 10.5.3], we have

$$\begin{aligned}&\mathbf{P}\left(\bigcap_{m\in\mathcal{M}_n} \{(X^n,Y^n(m)) \notin \overline{\mathcal{T}}_\varepsilon^n(XY)\}\right) \\ &\le (1-e^{-n(\bar{I}(\mathbf{X};\mathbf{Y})+\varepsilon)})^{|\mathcal{M}_n|} \le \exp(-e^{n(R-\bar{I}(\mathbf{X};\mathbf{Y})-\varepsilon)}),\end{aligned}$$

which establishes the lemma. ■

Lemma 14 (Mutual Covering) *Given $P_{\mathbf{XY}}$, let $X^n(m_1) \sim P_X^n, m_1 \in \mathcal{M}_{1n}$ with $|\mathcal{M}_{1n}| = e^{nR_1}$, $Y^n(m_2) \sim P_X^n, m_2 \in \mathcal{M}_{2n}$ with $|\mathcal{M}_{2n}| = e^{nR_2}$, $X^n(m_1)$'s are pairwise independent, $Y^n(m_2)$'s are pairwise independent, and $\{X^n(m_1)\}_{m_1 \in \mathcal{M}_{1n}}$ is independent of $\{Y^n(m_2)\}_{m_2 \in \mathcal{M}_{2n}}$. If $R_1 + R_2 > I(X;Y)$, then there exists an $\varepsilon > 0$ such that*

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcap_{\substack{(m_1,m_2)\in \\ \mathcal{M}_{1n}\times\mathcal{M}_{2n}}} \{(X^n(m_1), Y^n(m_2)) \notin \overline{\mathcal{T}}_\varepsilon^n(XY)\} \right) = 0.$$

Proof: Let X_0 be a constant, $X_{\mathcal{S}} = X$ and $X_{\mathcal{S}^\complement} = Y$ in Lem. 7, then this conclusion is obtained. ■

4.5 Conclusion

In this chapter, we generalised the conventional weak typicality for a general multivariate probability measure of the general stochastic process with a general abstract alphabet. We have provided a collection of typicality lemmas based on our proposed typicality. We have first proposed a general asymptotic conclusion for the generic typical set sequence. Based on this conclusion, we obtained the conditional typicality lemma and the strong Markov lemma for our generalised typicality, filling the gap that there were no such types of lemmas for the weak typicality. We have also obtained several packing and covering lemmas. For the convenience of some applications, we have provided corresponding bivariate versions of the typicality lemmas.

Chapter 5

Applications of Generalised Typicality Lemmas in Coding Problems

In this chapter, we will show several applications of generalised multivariate typicality lemmas, in source and channel coding problems with general alphabets and general source/channel measures. These problems are classical ones when restricted in discrete or i.i.d. scenarios, while we will here show it is possible to directly resolve them without any discretisation-and-approximation technique. In general, we only focus on achievability proofs in this chapter.

5.1 Applications in Source Coding

5.1.1 Rate-Distortion Problem with a General Source

In [11, Sec. 10.6], the authors studied a rate distortion problem with a discrete memoryless source, based on conditional strong typicality lemma, and obtain a result stronger than many other works.

In this section, we will extend this problem to a general scenario. As in [61] and [4, Sec 5.5], we state the problem as follows.

Definition 31 (Rate Distortion) *Given a source* $\mathbf{X} = (X_n)_{n=1}^{\infty}$ *with state space* $(\mathcal{X}, \mathcal{B}_X)$ *and*

probability law $P_{\mathbf{X}}$, let $(x_n)_{n=1}^{\infty}$ be any realisation of $(X_n)_{n=1}^{\infty}$ and $(y_n)_{n=1}^{\infty}$ is a $\mathcal{Y}^n$-valued sequence which is assumed to recover $(x_n)_{n=1}^{\infty}$, then the distortion measure *is defined as a sequence of measurable functions $\{d^{(n)}\}_{n=1}^{\infty}$, where $d^{(n)}(x^n,y^n)$ is a positive measurable function on $\mathcal{X}^n \times \mathcal{Y}^n$ for all n. For all n, we set an encoding function $\varphi^{(n)} : \mathcal{X}^n \to \mathcal{M}^{(n)}$ and a decoding function $\psi^{(n)} : \mathcal{M}^{(n)} \to \mathcal{Y}^n$, where $\mathcal{M}^{(n)} = \{m_n\}_{n=1}^{M_n}$.*

Definition 32 *A pair (R,D) is called* achievable *if there exists a sequence of $\{(\varphi^{(n)}, \psi^{(n)})\}_{n=1}^{\infty}$ satisfying*

$$\mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} d^{(n)}(X, \psi(\varphi(X))) \leq D$$

and

$$\mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} M_n \leq R.$$

Definition 33 *The* rate-distortion function $R(D|\mathbf{X})$ *is defined as the supremum of R over all achievable (R,D).*

Theorem 14 *Given a general source $\mathbf{X}$ with a $P_{\mathbf{X}}$ and a distortion measure $\{d^{(n)}\}_{n=1}^{\infty}$, then a) the rate-distortion function*

$$R(D|\mathbf{X}) = \inf_{P_{\mathbf{Y}|\mathbf{X}}: \overline{D}(\mathbf{X},\mathbf{Y}) \leq D} \overline{I}(\mathbf{X},\mathbf{Y}).$$

where

$$\overline{D}(\mathbf{X},\mathbf{Y}) = \mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} d^{(n)}(X,Y);$$

and b) there exists a sequence of $\{\mathcal{A}^{(n)} \subset \mathcal{Y}^n\}$ satisfying that $\lim_{n\to\infty} P_X^n(\mathcal{A}) = 1$ and for each $\{x^n \in \mathcal{A}^{(n)}\}$, $d^{(n)}(x, \psi(\phi(x))) \leq D$.

Proof: The first conclusion has been proved in [4, Sec 5.5]. We will here prove the second conclusion, which is an extension of the proof in [11, Sec. 10.6].

Fix a $P_{\mathbf{Y}|\mathbf{X}}$. First we define

$$\mathcal{S}_{\delta}^{(n)} = \left\{ (x^n,y^n) \in X^n \times Y^n \mid \left| \frac{1}{n} d^{(n)}(x,y) - \overline{D}(\mathbf{X},\mathbf{Y}) \right| < \delta \right\}$$

and set $\mathcal{T}^{(n)} = \mathcal{T}_{\varepsilon}^n(XY) \cap \mathcal{S}_{\delta}^{(n)}$.

Random codebook generation For all n, generate e^{nR} many $Y^n(m_n)$'s according to P_{Y^n}.

Encoding For each x^n, we assign an index m_n if $(x^n, Y^n(m_n)) \in \mathcal{T}^{(n)}$.

Decoding Output $Y^n(m_n)$ if m_n is received.

Error probability analysis An error will occur when we can not pick out a typical pair for any x^n in the coding procedure, or when the decoding output provides a distortion larger than D. The latter implies that $(x^n, Y^n(m_n)) \notin \mathcal{T}^{(n)}$ for all m_n.

■

Remark 8 *Part b) of Thm. 14 is the counterpart of a strong conclusion obtained in [11, Sec. 10.6], for discrete memoryless rate distortion problem.*

Remark 9 *In [1, Sec. 3.8], a Gaussian rate-distortion problem with a quadratic distortion measure was resolved by employing the strong typicality and the discretisation-and-approximation-technique. However, it might be difficult to resolve a rate-distortion problem with general a source and a general distortion measure, by the same technique based on the strong typicality. Hence, our generalised typicality is more useful in the general rate-distortion problem.*

5.1.2 Multiple Description Problem with General Sources

The multiple description problem is an multi-terminal extension of the rate-distortion problem. The tightest inner bound of the rate-distortion region for the discrete memoryless case was obtained by El Gamal and Cover [1, Sec. 13.3]. In this section, we will derive an El-Gamal-Cover-type inner bound for the case where the source is general.

Definition 34 (Multiple Description) *Given a source* $\mathbf{X} = (X_n)_{n=1}^{\infty}$ *with state space* $(\mathcal{X}, \mathcal{B}_X)$ *and probability law* $P_{\mathbf{X}}$*, define distortion measures* $d_k^{(n)}(x^n, y^n), k = 0,1,2$*, encoding functions* $\varphi_k^{(n)} : \mathcal{X}^n \to \mathcal{M}_k^{(n)}, k = 1,2$ *and decoding functions* $\psi_0^{(n)} : \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)} \to \mathcal{Y}_0^n, \psi_k^{(n)} : \mathcal{M}_k^{(n)} \to \mathcal{Y}_k^n, k = 1,2$*, where* $\mathcal{M}_k^{(n)} = \{m_n\}_{n=1}^{M_n}, k = 0,1,2$*. The encoding outputs are called* descriptions.

Definition 35 *A tuplet (R_1,R_2,D_0,D_1,D_2) is called maximum-achievable if there exits a code with rate pair (R_1,R_2) satisfying*

$$\overline{D}(\mathbf{X},\mathbf{Y}_j)=\mathrm{p}-\limsup_{n\to\infty}\frac{1}{n}d_j^{(n)}(X,Y_j)\leq D_j, j=0,1,2.$$

Definition 36 *The maximum-rate-distortion region $\mathcal{R}(D_0,D_1,D_2)$ is defined as the closure of all maximum-achievable rate pairs with respect to (D_0,D_1,D_2).*

Theorem 15 *Given a general source $\mathbf{X}$ with a $P_{\mathbf{X}}$, then $\mathcal{R}(D_0,D_1,D_2)$ is the closure of all R_1,R_2 satisfying that*

$$\begin{aligned}R_1&>\bar{I}(\mathbf{X},\mathbf{Y}_1),\\ R_2&>\bar{I}(\mathbf{X},\mathbf{Y}_2),\\ R_1+R_2&>\bar{I}(\mathbf{X};\mathbf{Y}_0,\mathbf{Y}_1,\mathbf{Y}_2)+\bar{I}(\mathbf{Y}_1,\mathbf{Y}_2),\end{aligned}$$

with some input measure $P_{\mathbf{Y}_0\mathbf{Y}_1\mathbf{Y}_2|\mathbf{X}}$ satisfying that $\overline{D}(\mathbf{X},\mathbf{Y}_j)\leq D_j, j=0,1,2$.

Proof: Fix a $P_{\mathbf{Y}_0\mathbf{Y}_1\mathbf{Y}_2|\mathbf{X}}$. For $j=1,2$, we define

$$\mathcal{S}_{\delta,j}^{(n)}=\left\{(x^n,y_j^n)\in X^n\times Y_j^n \mid \left|\frac{1}{n}d_j^{(n)}(x,y_j)-\overline{D}(\mathbf{X},\mathbf{Y}_j)\right|<\delta\right\}$$

and set $\mathcal{T}^{(n)}=\mathcal{T}_\varepsilon^n(XY_j)\cap\mathcal{S}_{\delta,j}^{(n)}$; then divide M_j into M_{0j} at rate R_{0j} and M_{jj} at rate R_{jj}.

Random codebook generation The codebook will be generated randomly and (conditionally) independently in each step. For $j=1,2$, generate $e^{nR_{jj}}$ sequences $Y_{jj}^n(m_{jj})$'s according to $P_{X_j^n}$. Then for each (m_{11},m_{22}), generate e^{nR_0} $Y_{jj}^n(m_{11},m_{22},m_0)$'s according to $P_{Y_0|Y_1Y_2}^n$.

Encoding For a source sequence x^n, pick out any tuplet

$$(x^n,y_1^n(m_{11}),y_2^n(m_{22}),y_0^n(m_0,m_{11},m_{22}))$$

from $\mathcal{T}^{(n)}$. Then send $m_0 = (m_{01}, m_{02})$, $m_1 = (m_{01}, m_{11})$ and $m_2 = (m_{02}, m_{22})$.

Decoding Decoders output $y_j^n(m_{jj}), j = 1,2$ and $y_0^n(m_0, m_{11}, m_{22})$, respectively.

Error probability and distortion analysis First we consider the error in the encoding procedure. If $\mathcal{T}^{(n)}$ is empty, then an error occurs. By the general multivariate covering lemma and the fact that the rate of X^n is 0, the probability of this error approaches 0 if

$$\begin{aligned} R_{11} &> \bar{I}(\mathbf{X}, \mathbf{Y}_1), \\ R_{22} &> \bar{I}(\mathbf{X}, \mathbf{Y}_2), \\ R_0 + R_{11} + R_{22} &> \bar{I}(\mathbf{X}; \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Y}_2) + \bar{I}(\mathbf{Y}_1, \mathbf{Y}_2). \end{aligned}$$

The expression of the inner bound can be obtained by Fourier-Motzkin elimination.

Then we consider the maximum-distortion. For $j = 0,1,2$ and arbitrary $\gamma > 0$, we have

$$\mathbf{P}\{\frac{1}{n} d_j^{(n)}(X^n, Y_j^n(M_j)) > \overline{D}(\mathbf{X}; \mathbf{Y}_j) + \gamma\} < \mathbf{P}\{((X^n, Y_j^n(M_j)) \notin \mathcal{T}^{(n)}\},$$

where the RHS approaches 0 if $R_j > \bar{I}(\mathbf{X}, \mathbf{Y}_j)$. This implies that

$$\mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} d_j^{(n)}(X, Y_j) \le \overline{D}(\mathbf{X}; \mathbf{Y}_j) + \gamma \le D_j + \gamma.$$

Because γ can be arbitrarily small, there must exist a code satisfying

$$\mathrm{p}-\limsup_{n\to\infty} \frac{1}{n} d_j^{(n)}(X, Y_j) \le \overline{D}(\mathbf{X}; \mathbf{Y}_j) + \gamma \le D_j.$$

The theorem is proved. ■

As for the stationary memoryless case in [1, Sec. 13.3], we can provide an alternative proof through our general conclusion. The technique we employ here is similar as in [4, Sec. 7.8].

Corollary 4 *In Thm. 15, if the source X is a stationary and memoryless one with a P_X, then*

$\mathcal{R}(D_0,D_1,D_2)$ is the closure of all R_1,R_2 satisfying that

$$\begin{aligned} R_1 &> I(X,Y_1|Q), \\ R_2 &> I(X,Y_2|Q), \\ R_1+R_2 &> I(X;Y_0,Y_1,Y_2|Q)+I(Y_1,Y_2|Q), \end{aligned}$$

with some input measure $P_Q P_{Y_0Y_1Y_2|XQ}$ with $|\mathcal{Q}| \le 6$, satisfying that $D(X,Y_j)=\mathbf{E}(d_j(X,Y_j)) \le D_j$, $j=0,1,2$.

Proof: Fix a $P_Q P_{Y_0Y_1Y_2|XQ}$ with $\mathcal{Q}=\{1,2,3,4,5,6\}$, satisfying that

$$D(X,Y_j)=\mathbf{E}(d_j(X,Y_j)) \le D_j, j=0,1,2.$$

In $\mathcal{R}(D_0,D_1,D_2)$ of Thm. 15, for any n, let $P^n_{Y_0Y_1Y_2|X}$ follow $P_{Y_{0k}Y_{1k}Y_{2k}|X_k}=P_{Y_0Y_1Y_2|XQ}(\cdot|\cdot,q)$ if $n_{q-1}<k\le n_q$, where $n_0=0, n_k=\lceil n\sum_{l=1}^k P_Q(q)\rceil$ for $q=1,2,3,4,5,6$. Then we have

$$i^n(X;Y_1)=\sum_{q=1}^{6}\frac{n_q-n_{q-1}}{n}\sum_{k=n_{q-1}}^{n_q} i(X_k;Y_{1k}).$$

Because $\lim_{n\to\infty}\frac{n_q-n_{q-1}}{n}=P_Q(q)$, then following the LLN, we have

$$\begin{aligned} \bar{I}(\mathbf{X},\mathbf{Y}_1) &= \mathrm{p}-\lim_{n\to\infty}\sum_{q=1}^{6}\frac{n_q-n_{q-1}}{n}\sum_{k=n_{q-1}}^{n_q} i(X_k;Y_{1k}) \\ &= \sum_{q=1}^{6} P_Q(q) I(X;Y_1|Q=q) \\ &= I(X;Y_1|Q). \end{aligned}$$

Similarly, we can obtain the other two constraints. $\overline{D}(\mathbf{X},\mathbf{Y}_j)=D(X;Y_j)$ can also be obtained according to the LLN. ■

5.1.3 Memoryless Berger-Tung Inner Bound with General Alphabets

The Berger-Tung Inner bound is obtained for a multi-terminal lossy source coding problem. In [1, Sec. 12.1], it is proved based on a strong typicality based Markov lemma with finite alphabets. In this section we generalise it to a general alphabet scenario with a slight restriction.

First we provide a specified Markov lemma for this problem, which is also based on the proposed multivariate typicality.

Lemma 15 *With a given $P_{\mathbf{XYZ}}$ satisfying that $P^n_{XYZ}(x^n,y^n,z^n)=\prod_{k=1}^n P_{X|Y}(x_k|y_k)\times P_Y(y_k)\times P_{Z|Y}(z_k|y_k)$ for all n, let $(x^n,y^n)\in \underline{\mathfrak{T}}^n_\varepsilon(XY|Z)$. If $Q_{Z^n|Y^n}\le e^{n\varepsilon}P_{Z^n|Y^n}$, and for any $\delta>0$,*

$$\left[\frac{\mathrm{d}(P_{X|Y}\times P_{Z|Y})}{\mathrm{d}(P_X\times P_Z)}\right]^{1+\delta}$$

is $P_{X|Y}\times P_{Z|Y}$-integrable, then

$$\lim_{n\to\infty} Q_{Z^n|Y^n}(\underline{\mathfrak{T}}_\varepsilon^{Z^n|X^nY^n})=1.$$

Proof: Note that the spectral inf-relative entropy rate degrades to relative entropy rate in this case. Let

$$\begin{aligned} g(x,y,z)&=\frac{\mathrm{d}(P_{X|Y}\times P_Y\times P_{Z|Y})}{\mathrm{d}(P_X\times P_Y\times P_Z)}(x,y,z)\\ &=\frac{\mathrm{d}(P_{X|Y}\times P_{Z|Y})}{\mathrm{d}(P_X\times P_Z)}(x,y,z), \end{aligned}$$

which satisfies the log-exponential property in [17, Sec. VI], which makes this lemma as a special case of [17, Cor. VI.4]. ■

Because the random coding, typicality decoding and error analysis is trivial, we omit the proof and results. We only point out here that a closeness of two conditional probability in the form of $Q_{Z^n|Y^n}\le e^{n\varepsilon}P_{Z^n|Y^n}$ can be obtained following a similar way in [1, App. 12B].

5.2 Applications in Channel Coding

5.2.1 Channel Coding with Input Constraint

This problem is based on the point-to-point channel coding problem. We will first define the channel input constraint as follows.

Definition 37 *Given a channel $P_{\mathbf{Y}|\mathbf{X}}$, if for all n, we impose a constraint on the channel input that $\Gamma^{(n)}(x^n) \doteq \Gamma^{(n)}(x) < \gamma_n$, where $\gamma_n > 0$ and $\Gamma^{(n)}(\cdot)$ is a measurable function, then the sequence $\{\Gamma^{(n)}(x) < \gamma_n\}_{n=1}^{\infty}$ is called the* input constraint.

A Feinstein-type lemma with input constraint was considered in [7].

Theorem 16 *Given a general channel $P_{\mathbf{Y}|\mathbf{X}}$ with the input constraint $\{\Gamma^{(n)}(x) < \gamma_n\}_{n=1}^{\infty}$, then the capacity is*

$$C = \sup_{\mathcal{P}} \underline{I}(\mathbf{X};\mathbf{Y}),$$

where

$$\mathcal{P} = \{P_{\mathbf{X}} | \mathbf{E}(\Gamma^{(n)}(X) < \gamma_n, n = 1,2,\cdots\}.$$

Proof:

Random coding For fixed $P_{\mathbf{X}} \in \mathcal{P}$ and n, randomly generate e^{nR} many $x^n(m)$'s according to P_X^n.

Decoding Assume y^n is received. If $(x^n(\hat{m}), y^n) \in \mathcal{T}_\varepsilon^n(X,Y)$, then declare $\hat{m}$ is sent.

Error analysis $\mathcal{E} = \mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2$, where

$$\begin{aligned}
\mathcal{E}_0 &= \{\Gamma^{(n)}(X) \geq \gamma_n\}, \\
\mathcal{E}_1 &= \{(X^n(m), Y^n) \notin \mathcal{T}_\varepsilon^n(X,Y)\}, \\
\mathcal{E}_2 &= \{(X^n(\hat{m}), Y^n) \in \mathcal{T}_\varepsilon^n(X,Y)\}
\end{aligned}$$

for an $\hat{m} \neq m$. We set a

$$\mathcal{T}^n = \mathcal{T}^n_\varepsilon(X,Y) \cap \{(x^n, y^n) \mid \Gamma^{(n)}(x) < \gamma_n\}$$

for all n. When X^n and P_X^n satisfies $\mathbf{E}(\Gamma^{(n)}(X)) < \gamma_n$, $\{\mathcal{T}^n\}_{n=1}^\infty$ is a generic typical set sequence, and

$$\mathcal{E}_1 = \{(X^n(m), Y^n) \notin \mathcal{T}^n\},$$
$$\mathcal{E}_2 = \{(X^n(\hat{m}), Y^n) \in \mathcal{T}^n.\}$$

Hence, $\lim_{n\to\infty} \mathbf{P}\{\mathcal{E}_0\} = \lim_{n\to\infty} \mathbf{P}\{\mathcal{E}_1\} = \lim_{n\to\infty} \mathbf{P}\{\mathcal{E}_2\} = 0$ when $R < \underline{I}(\mathbf{X};\mathbf{Y})$ and $\mathbf{E}(\Gamma^{(n)}(X) < \gamma_n$ for all n.

■

5.2.2 Gelfand-Pinsker Coding

Gelfand-Pinsker (GP) coding problem [20] is a channel coding problem in which a channel state is noncausally available at the encoder. We restate the general GP coding problem [62] as follows. The channel is defined by the input $\mathbf{X}$, the output $\mathbf{Y}$, the general state $\mathbf{S} \sim P_{\mathbf{S}}$, and the transition probability $P_{\mathbf{Y}|\mathbf{SX}}$. For a fixed codelenth n, the encoder f is a mapping from $\mathcal{M} \times \mathcal{S}^n$ to $\mathcal{X}^n$ where $\mathcal{M}$ is the message set and $\mathcal{S}$ is the state space, and the decoder g is a mapping from $\mathcal{Y}^n$ to $\mathcal{M}$. The average error probability ε_n is the average probability of the event that $g(Y^n)$ is not equal to the sent message. In [62], Tan obtained the capacity of the generalised GP coding.

Theorem 17 (Gelfand-Pinsker-Tan) *The capacity of the general channel $P_{\mathbf{Y}|\mathbf{XS}}$ with general non-causal state $\mathbf{S}$ only available at the encoder is $C = \sup_{P_{\mathbf{UX}} \in \mathcal{P}} \underline{I}(\mathbf{U};\mathbf{Y}) - \bar{I}(\mathbf{U};\mathbf{S})$, where $\mathcal{P}$ is the set of all $P_{\mathbf{UX}}$'s satisfying that $\mathbf{U} \to (\mathbf{X},\mathbf{S}) \to \mathbf{Y}$ forms a Markov chain.*

In order to prove the achievability of the capacity, Tan employed a modified piggyback coding lemma (PBL) [63, Lemma 4.3] to get around a counterpart of the conditional typical

lemma in the case where typicality is defined by the information-spectral quantity. In the following, we will recover the achievability of the capacity of the general GP coding using our proposed conditional typicality lemma (Lemma 5), instead of a lemma analogous to PBL.

Analogous to the proof of [1, Theorem 7.3], we employ the random coding and the typicality decoding techniques to prove the achievability of Theorem 17. The difference is that we employ the general information-density-based definition of typicality in the decoding metrics and the error analysis.

Random codebook generation For fixed $P_{\mathbf{U}|\mathbf{S}}, P_{\mathbf{X}|\mathbf{US}}$ and let $P_{\mathbf{USXY}}$ be determined by $P_{\mathbf{S}}$ and the transition probability measures $P_{\mathbf{U}|\mathbf{S}}, P_{\mathbf{X}|\mathbf{US}}$ and $P_{\mathbf{Y}|\mathbf{USX}} = P_{\mathbf{Y}|\mathbf{SX}}$. For $\tilde{R} > R$, a fixed codelength n and each $m \in \mathcal{M}$, randomly and independently generate $e^{n(\tilde{R}-R)}$ $u^n(l)$'s according to P_{U^n}, where l's are indices of the sequences. For each $u^n(l)$ and s^n, randomly and independently generate an $x^n(s^n,l)$ according to $P_{X^n|U^nS^n}(\cdot|u^n(l),s^n)$.

Encoding Assume that a specified message M is sent. Choose a $u^n(L)$ from $u^n(l)$'s corresponding to M such that $(u^n(L),s^n) \in \overline{\mathcal{T}}_\varepsilon^{U^nS^n} \cap \underline{\mathcal{T}}_\varepsilon^{U^n|Y^n}$, where $P_{U^nY^n}$ is the marginal of $P_{U^nS^nX^nY^n}$, and then the index L is specified. Send the corresponding $x^n(s^n,L)$.

Decoding Assume that y^n is received. The decoding output returns that $\hat{m}$ is sent if there exists an $u^n(\hat{l})$ satisfying $(u^n(\hat{l}),y^n) \in \underline{\mathcal{T}}_\varepsilon^{U^nY^n}$, where $u^n(\hat{l})$ corresponds to $\hat{m}$.

Error probability analysis $\varepsilon_n \leq \mathbf{P}(\mathcal{E}_1) + \mathbf{P}(\mathcal{E}_1^\complement \cap \mathcal{E}_2) + \mathbf{P}(\mathcal{E}_3)$, where the error events are $\mathcal{E}_1$:$(U^n(l),S^n) \notin \overline{\mathcal{T}}_\varepsilon^{U^nS^n} \cap \underline{\mathcal{T}}_\varepsilon^{U^n|Y^n}$ for all $U^n(l)$ corresponding to M, $\mathcal{E}_2$:$(U^n(L),Y^n) \notin \underline{\mathcal{T}}_\varepsilon^{U^nY^n}$, $\mathcal{E}_3$:$(U^n(l),Y^n) \in \overline{\mathcal{T}}_\varepsilon^{U^nY^n}$ for some $U^n(l)$ corresponding to $m \neq M$.

From the covering lemma and the bivariate conditional typicality lemma, $\lim_{n\to\infty}\mathbf{P}(\mathcal{E}_1) = 0$ if $\tilde{R} - R > \bar{I}(\mathbf{U};\mathbf{S})$; from bivariate conditional typicality lemma, $\lim_{n\to\infty}\mathbf{P}(\mathcal{E}_1^\complement \cap \mathcal{E}_2) = 0$; and from the packing lemma, $\lim_{n\to\infty}\mathbf{P}(\mathcal{E}_3) = 0$ if $\tilde{R} < \underline{I}(\mathbf{U};\mathbf{Y})$. The achievability is established.

5.2.3 Multi-User BC with a Common Message

The capacity region of a general two-user broadcast channel (BC) was obtained in [64] by the information-spectrum approach. However, the single-letter expression of capacity region of the discrete memoryless BC (SMBC) is still an open problem. The tightest inner bound of this region is a Marton-type inner bound [1, 18, 19]. We will show that a Marton-type inner bound of the capacity region of a K-user memoryless BC (MBC) can be derived using the generalised multivariate typicality lemmas.

Theorem 18 *Let $\mathcal{K} = \{2, \cdots, K\}$. Given a K-receiver MBC $P_{Y_1 \cdots Y_K|X}$ with a common message M_0 for all receivers and a private massage M_1 for receiver 1, then the capacity region of the K-receiver MBC is the closure of the set of rate pair (R_0, R_1) satisfying*

$$\sum_{k\in\mathcal{S}} \tilde{R}_{1k} > D(P_{V_{\mathcal{S}}|U} || \prod_{k\in\mathcal{S}} P_{V_k|U} | P_U), \tag{5.1}$$

$$R_{11} + \sum_{k\in\mathcal{S}} \bar{R}_{1k} < I(V_{\mathcal{S}}, X; Y_1 | U, V_{\mathcal{S}}^{\complement}), \tag{5.2}$$

$$R_0 + R_{10} + R_{11} + \sum_{k\in\mathcal{S}} \bar{R}_{1k} < I(U, V_{\mathcal{S}}, X; Y_1 | V_{\mathcal{S}}^{\complement}), \tag{5.3}$$

$$R_0 + R_{10} < I(U; Y_k), \tag{5.4}$$

and $R_1 = R_{10} + \cdots + R_{1K}, \bar{R}_{1k} = R_{1k} + \tilde{R}_{1k}$, for some $P_U \times \prod_{k=2}^{K} P_{V_k|U} \times P_{X|V_{\mathcal{K}}}$ and the given $P_{Y_1 \cdots Y_K|X}$ and for each $k \in \mathcal{K}$ and $\mathcal{S} \subset \mathcal{K}$.

Proof:

Codebook Generation Fix an input measure $P_U \times \prod_{k=2}^{K} P_{V_k|U} \times P_{X|V_{\mathcal{K}}}$. Divide M_1 into $M_{10}, M_{11}, \cdots, M_{1K}$ and let. We suppose that the following random codebooks are all generated randomly and (conditionally) independently. First we generate $e^{n(R_0+R_{10})}$ sequences $u^n(m_0, m_{10})$, each according to $\prod_{l=1}^{n} P_U(u_l)$. For each $k \in \mathcal{K}$ and m_{1k}, we generate $e^{n\tilde{R}_{1k}}$ sequences $v_k^n(m_0, m_{10}, m_{1k}, l_{1k})$, where $l_{1k} = (m_{1k} - 1)e^{n\tilde{R}_{1k}} + 1, \cdots, m_{1k}e^{n\tilde{R}_{1k}}$, each according to $\prod_{l=1}^{n} P_{V_k|U}(v_{kl}|u_l)$. Then for each $\{m_{1k}\}_{k=2}^{K}$, we pick out a jointly typical tuple $\{v_k^n(m_0, m_{10}, m_{1k}, L_{1k})\}_{k=2}^{K}$. In the end, for each $(m_0, m_{10}, m_{11}, \cdots, m_{1K})$, we generate an

$x^n(m_0, m_{10}, m_{11}, \cdots, m_{1K})$, each according to $\prod_{l=1}^{n} P_{X|V_{\mathcal{K}}}(x_l|v_{\mathcal{K}l})$.

Encoding To send the message tuple $(m_0, m_{10}, m_{11}, \cdots, m_{1K})$, transmit $x^n(m_0, m_{10}, m_{11}, \cdots, m_{1K})$.

Deoding Receiver 1 performs the joint-typicality decoding on $(u^n, v_2^n, \cdots, v_K^n, x^n, y_1^n)$. Receiver k where $k \in \mathcal{K}$ performs the joint-typicality decoding on (u^n, v_k^n, y_k^n).

Error Probability Analysis In the codebook generation. if there exists no jointly typical tuple $\{v_k^n(m_0, m_{10}, m_{1k}, L_{1k})\}_{k=2}^{K}$, then an error occurs. According to the multivariate covering lemma, the probability of this sort of error event finally approaches 0 when $n \to \infty$, if for each $\mathcal{S} \subset \mathcal{K}$

$$\sum_{k \in \mathcal{S}} \tilde{R}_{1k} > D(P_{V_{\mathcal{S}|U}} || \prod_{k \in \mathcal{S}} P_{V_k|U} | P_U). \tag{5.5}$$

Let $\bar{R}_{1k} = R_{1k} + \tilde{R}_{1k}$. In the typicality decoding, if any decoder output an estimated message pair which is corresponding to a jointly typical tuple but is not equivalent to the one sent by the transmitter, then an error occurs. According to the packing lemma, the probability of this sort of error event finally approaches 0 when $n \to \infty$, if for each $k \in \mathcal{K}$ and $\mathcal{S} \subset \mathcal{K}$

$$R_{11} + \sum_{k \in \mathcal{S}} \bar{R}_{1k} < I(V_{\mathcal{S}}, X; Y_1 | U, V_{\mathcal{S}}^{\mathcal{C}}), \tag{5.6}$$

$$R_0 + R_{10} + R_{11} + \sum_{k \in \mathcal{S}} \bar{R}_{1k} < I(U, V_{\mathcal{S}}, X; Y_1 | V_{\mathcal{S}}^{\mathcal{C}}), \tag{5.7}$$

$$R_0 + R_{10} < I(U; Y_k). \tag{5.8}$$

According to the joint typicality lemma, the probability of the error event that the sent codeword $x^n(m_0, m_{10}, m_{11}, \cdots, m_{1K})$ is not jointly typical with any received sequence finally approaches 0 when $n \to \infty$, ■

5.3 Asymptotic Analysis on the Second-Order Coding Rate of the General MAC

5.3.1 System Model and Basic Definitions

Definition 38 *A general two-user multiple access channel (MAC) model is defined by the channel components and channel coding.*

Channel input and output are sets of random sequences (not necessarily stochastic processes according to [65]) and

$$\mathbf{Y} = \{Y^n = (Y_1^{(n)}, \cdots, Y_n^{(n)})\}_{n=1}^{\infty},$$

where $X_{1,i}^{(n)}, X_{2,i}^{(n)}$ and $Y_i^{(n)}$ are random variables on $\mathcal{X}_1, \mathcal{X}_2$ and $\mathcal{Y}$ respectively, for every i and n.

Channel transition probability is determined by the conditional probability measure $P_{\mathbf{Y}|\mathbf{X}_1\mathbf{X}_2}(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)$, or equivalently, an infinite set of consistent conditional probability measures $\left\{P_{Y^n|X_1^n X_2^n}(y^n|x_1^n, x_2^n)\right\}_{n=1}^{\infty}$.

An $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ channel code consists of the following elements.

Message sets *the sets of message indices $\mathcal{M}_1^{(n)} = \{1, 2, \cdots, N_1^{(n)}\}, \mathcal{M}_2^{(n)} = \{1, 2, \cdots, N_2^{(n)}\}$, where $N_1^{(n)}$ and $N_2^{(n)}$ equal to the cardinality of corresponding message set, respectively. The random message pair (M_1, M_2) is uniformly distributed on $\mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)}$.*

Encoding functions $\varphi_k^{(n)} : \mathcal{M}_k \to \mathcal{X}_k^n, k = 1, 2.$

Decoding function $\psi^{(n)} : \mathcal{Y}^n \to \mathcal{M}_1 \times \mathcal{M}_2.$

Decoding sets

$\{\mathcal{D}_{ij}\}_{(i,j)\in\mathcal{M}_1^{(n)}\times\mathcal{M}_2^{(n)}}$, *where* $\mathcal{D}_{ij} = \{y^n | \psi^{(n)}(y) = (i, j)\}$.

Average error probability is the probability of the event that the decision output of the receiver does not equal to the actually sent message pair, i.e. $\varepsilon^{(n)} = \mathbf{P}\{(\hat{M}_1, \hat{M}_2) \neq (M_1, M_2)\}$.

5.3.2 Upper and Lower Bounds on the Average Error Probability in the General MAC

In this section, we will propose an upper bound and a lower bound on the average error probability of the general MAC, in the non-asymptotic form.

Upper Bounds

First we will compare two previous results regarding the non-asymptotic upper bound.

Lemma 16 (Verdú's Bound [66]) *For any positive integer n and any positive real number* γ*, there exists a* $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ *code satisfying*

$$\begin{aligned}\varepsilon^{(n)} \leq\ & \mathbf{P}\{i^n(X_1;Y|X_2,U) \leq \log N_1^{(n)} + \gamma\} + \mathbf{P}\{i^n(X_2;Y|X_1,U) \leq \log N_2^{(n)} + \gamma\} \\ & + \mathbf{P}\{i^n(X_1,X_2;Y|U) \leq \log (N_1^{(n)} N_2^{(n)}) + \gamma\} + 3e^{-\gamma} \end{aligned} \tag{5.9}$$

for any input measure $P_{UX_1^nX_2^n}$ *satisfying that* X_1^n *and* X_2^n *are conditionally independent given* U.

Remark 10 *Verdú did not specify the blocklength n explicitly in his original bound [66, Theorem 4]. However, we learn from [66, Section I] that his bound can be expressed as* (5.9).

Lemma 17 (Han's Bound [65]) *For any positive integer n and any positive real number* γ*,*

there exists a $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ *code satisfying*

$$\varepsilon^{(n)} \leq \mathbf{P}\Big(\{\frac{1}{n} i^n(X_1;Y|X_2,U) \leq \frac{1}{n}\log N_1^{(n)} + \gamma\} \cup \{\frac{1}{n} i^n(X_2;Y|X_1,U) \leq \frac{1}{n}\log N_2^{(n)} + \gamma\} \\ \cup \{\frac{1}{n} i^n(X_1,X_2;Y|U) \leq \frac{1}{n}\log (N_1^{(n)} N_2^{(n)}) + \gamma\} \Big) + 3e^{-n\gamma} \tag{5.10}$$

for any input $P_{UX_1^nX_2^n}$ *satisfying that* X_1^n *and* X_2^n *are conditionally independent given* U.

Remark 11 *Actually, the time-sharing random variable* U *is introduced by us in Han's bound [65, Lemma 3] to compare with Verdú's bound. Han did not include* U *in his result because the auxiliary variable is not necessary to obtain the asymptotic capacity region.*

Verdu's bound was obtained by introducing the non-asymptotic packing lemma and covering lemma, while Han's bound was an intermediate result in the asymptotic analysis employing the standard analysis procedures of the information-spectrum method. The difference between inequalities (5.9) and (5.10) mainly lies in the union bound. If we substitute $n\gamma$ for γ in (5.9), the sum of the three probabilities are actually the union bound of the first term in (5.10). Hence, Han's bound is tighter in this scenario. On the other hand, if we circumvent the union bound in Verdú's proof, we can obtain a more general bound in Lemma 18, which is applicable in not only the first-order but also higher-order analysis of the coding rate.

Lemma 18 *For any positive integer n and any positive real number* γ*, there exists an* $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ *code satisfying*

$$\varepsilon^{(n)} \leq \mathbf{P}\Big(\{i^n(X_1;Y|X_2,U) \leq \log N_1^{(n)} + \gamma\} \cup \{i^n(X_2;Y|X_1,U) \leq \log N_2^{(n)} + \gamma\} \\ \cup \{i^n(X_1,X_2;Y|U) \leq \log (N_1^{(n)} N_2^{(n)}) + \gamma\} \Big) + 3e^{-\gamma}$$

for any input $P_{UX_1^nX_2^n}$ *satisfying that* X_1^n *and* X_2^n *are conditionally independent given* U.

The difference between our proposed upper bound and Verdu's bound is that we do not employ the union bound. We will show that this is a necessary improvement in Section 5.3.3,

becauce from our bound, we can deduce a tight inner region of the asymptotic second-order capacity region for the general MAC. Similar to the upper bound, we will also provide a lower bound of the average error probability.

A Lower Bound

Han has provided a non-asymptotic lower bound to the average error probability [65, Lemma 4]. Inspired by the preceding comparison, we obtain a more general expression for the lower bound. We will first introduce a lemma in probability theory, which will then be used in the proof of our proposed lower bound in Lemma 20.

Lemma 19 *Any probability measures P_1 and $Q_i, i \in \{1,2,\cdots,n\}$ on a measurable space $(\Omega,\mathcal{A})$ satisfy*

$$\begin{aligned}&\max_{A\in\mathcal{A}}[P_1(A)-\sum_{i=1}^{n}a_iQ_i(A)]\\&=P_1\big(\{\omega|P_1(\omega)-\sum_{i=1}^{n}a_iQ_i(\omega)\geq 0\}\big)-\sum_{i=1}^{n}a_iQ_i\big(\{\omega|P_1(\omega)-\sum_{i=1}^{n}a_iQ_i(\omega)\geq 0\}\big)\end{aligned}$$

for any positive real numbers $a_i, i \in \{1,2,\cdots,n\}$.

Remark 12 *The essence of Lemma 19 was implied in Han's proof [65, Lemma 4]. A specialized version of Lemma 19 was employed by Hayashi to prove inequality (65) in [67].*

The following lemma depicts our proposed lower bound.

Lemma 20 *All $(n,N_1^{(n)},N_2^{(n)},\varepsilon^{(n)})$ codes must satisfy*

$$\begin{aligned}\varepsilon^{(n)}\geq\mathbf{P}\Big(&\{i^n(X_1,Y|X_2)\leq\log N_1^{(n)}-\gamma\}\cup\{i^n(X_2,Y|X_1)\leq\log N_2^{(n)}-\gamma\}\\&\cup\{i^n(X_1,X_2,Y)\leq\log(N_1^{(n)}N_2^{(n)})-\gamma\}\Big)-3e^{-\frac{\gamma}{2}}\end{aligned}$$

for some input measure $P_{X_1}^n\times P_{X_2}^n$, where $P_{X_k}^n(x_k)=\frac{\chi_{\varphi_k^{(n)}(\mathcal{M}_k^{(n)})}(x_k^n)}{N_k^{(n)}}, k=1,2$, $\times$ denotes the product probability measure and $\chi_A(\cdot)$ denotes the indicator function of a set A.

Proof: First we set

$$\mathcal{L}^{(n)} = \{y^n | P_1^n(y) - N_1^{(n)} e^{-\frac{\gamma}{2}} Q_1(y^n) - N_2^{(n)} e^{-\frac{\gamma}{2}} Q_2(y^n) - N_1^{(n)} N_2^{(n)} e^{-\frac{\gamma}{2}} Q_{12}(y^n) \geq 0\}$$

for any probability measures P_1, Q_1, Q_2 and Q_{12} on $\mathcal{X}^n$.

From Lemma 19, we can obtain

$$\begin{aligned}
&P_1(\mathcal{D}_{ij}^{(n)}) - N_1^{(n)} e^{-\frac{\gamma}{2}} Q_1(\mathcal{D}_{ij}^{(n)}) - N_2^{(n)} e^{-\frac{\gamma}{2}} Q_2(\mathcal{D}_{ij}^{(n)}) - N_1^{(n)} N_2^{(n)} e^{-\frac{\gamma}{2}} Q_{12}(\mathcal{D}_{ij}^{(n)}) \\
&\leq P_1(\mathcal{L}^{(n)}) - N_1^{(n)} e^{-\frac{\gamma}{2}} Q_1(\mathcal{L}^{(n)}) - N_2^{(n)} e^{-\frac{\gamma}{2}} Q_2(\mathcal{L}^{(n)}) - N_1^{(n)} N_2^{(n)} e^{-\frac{\gamma}{2}} Q_{12}(\mathcal{L}^{(n)}) \\
&\leq P_1(\mathcal{L}^{(n)}) \\
&\leq P_1\Big(\{y^n | P_1^n(y) - N_1^{(n)} e^{-\frac{\gamma}{2}} Q_1(y^n) \geq 0\} \cap \{y^n | P_1^n(y) - N_2^{(n)} e^{-\frac{\gamma}{2}} Q_2(y^n) \geq 0\} \\
&\quad \cap \{y^n | P_1^n(y) - N_1^{(n)} N_2^{(n)} e^{-\frac{\gamma}{2}} Q_{12}(y^n) \geq 0\}\Big) \\
&= P_1\Big(\{y^n | \log \frac{P_1^n(y)}{Q_1(y^n)} \geq \log N_1^{(n)} - \frac{\gamma}{2}\} \cap \{y^n | \log \frac{P_1^n(y)}{Q_2(y^n)} \geq \log N_2^{(n)} - \frac{\gamma}{2}\} \\
&\quad \cap \{y^n | \log \frac{P_1^n(y)}{Q_{12}(y^n)} \geq \log N_1^{(n)} N_2^{(n)} - \frac{\gamma}{2}\}\Big). \qquad (5.11)
\end{aligned}$$

Because P_1, Q_1, Q_2 and Q_{12} are arbitrarily selected, we can obtain that for any

$(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ code and any positive real number γ

$$
\begin{aligned}
1-\varepsilon^{(n)} &= \frac{1}{N_1^{(n)}N_2^{(n)}} \sum_{i=1}^{N_1^{(n)}} \sum_{j=1}^{N_2^{(n)}} P_{Y^n|X_1^nX_2^n}(\mathcal{D}_{ij}^{(n)}|x_1^n(i), x_2^n(j)) \\
&\leq \frac{1}{N_1^{(n)}N_2^{(n)}} \sum_{i=1}^{N_1^{(n)}} \sum_{j=1}^{N_2^{(n)}} \Big(N_1^{(n)} e^{-\frac{\gamma}{2}} P_{Y^n|X_2^n}(\mathcal{D}_{ij}^{(n)}|x_2^n(j)) \\
&\quad - N_2^{(n)} e^{-\frac{\gamma}{2}} P_{Y^n|X_1^n}(\mathcal{D}_{ij}^{(n)}|x_1^n(i)) - N_1^{(n)}N_2^{(n)} e^{-\frac{\gamma}{2}} P_{Y^n}(\mathcal{D}_{ij}^{(n)}) \\
&\quad + P_{Y^n|X_1^nX_2^n}\Big(\{y^n | i(x_1^n(i); y^n|x_2^n(j)) \geq \log N_1^{(n)} - \frac{\gamma}{2}\} \\
&\quad\quad \cap \{y^n | i(x_2^n(j); y^n|x_1^n(i)) \geq \log N_2^{(n)} - \frac{\gamma}{2}\} \\
&\quad\quad \cap \{y^n | i(x_1^n(i), x_2^n(j); y^n) \geq \log N_1^{(n)}N_2^{(n)} - \frac{\gamma}{2}\}\Big) \qquad (5.12) \\
&\leq 3e^{-\frac{\gamma}{2}} + \mathbf{P}\Big(\{i^n(X_1, Y|X_2) > \log N_1^{(n)} - \gamma\} \cap \{i^n(X_2, Y|X_1) > \log N_2^{(n)} - \gamma\} \\
&\quad \cap \{i^n(X_1, X_2, Y) > \log (N_1^{(n)}N_2^{(n)}) - \gamma\}\Big),
\end{aligned}
$$

where (5.12) follows from (5.11). Thus, Lemma 20 is proved. ■

In the non-asymptotic bounds proposed in Lemmas 18 and 20, no n^α coefficients are included in $\log N^{(n)}$ terms, which is different from Han's bounds in [65]. Actually, our bounds are more general because their applicability is not limited to first-order analysis. Besides, it is possible to derive asymptotic results from the non-asymptotic bounds, as referred to in [66]. In the next section, we will perform an asymptotic second-order coding rate analysis for the general MAC, exactly based on the bounds in Lemmas 18 and 20.

5.3.3 An Asymptotic Second-Order Capacity Region of the General MAC

It was shown in [67] that the second-order coding rate is more accurate than conventional asymptotic results. In this section we will propose a second-order capacity region of the general MAC based on the information-spectrum method. As is pointed out in Section 5.3.2, our proposed non-aymptotic upper and lower bounds will play a fundamental role in

the derivation of the asymptotic second-order capacity region. First we define some basic notions in the second-order analysis for the MAC, as a counterpart of Hayashi's second-order notations for the general single-user channel [67, Section VII].

Definition 39 *A second-order coding rate pair $(R_1^{(2)}, R_2^{(2)})$ is (ε, R_1, R_2)-achievable if and only if there exists a sequence of $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ codes satisfying*

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}}(\log N_1^{(n)} - R_1 n) \geq R_1^{(2)},$$
$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}}(\log N_2^{(n)} - R_2 n) \geq R_2^{(2)},$$
$$\limsup_{n\to\infty} \varepsilon^{(n)} \leq \varepsilon.$$

The second-order (ε, R_1, R_2)-capacity region $\mathcal{C}(\varepsilon, R_1, R_2)$ is the set of all (ε, R_1, R_2)-achievable second-order rate pairs.

Definition 40 *Fixing the channel input $\mathbf{X}_1, \mathbf{X}_2$, we define a function of the first- and second-order coding rates as*

$$\begin{aligned} &J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \\ &= \limsup_{n\to\infty} \mathbf{P}\Big(\{\frac{1}{\sqrt{n}}(i^n(X_1; Y|X_2) - R_1 n) \leq R_1^{(2)}\} \cup \{\frac{1}{\sqrt{n}}(i^n(X_2; Y|X_1) - R_2 n) \leq R_2^{(2)}\} \\ &\quad \cup \{\frac{1}{\sqrt{n}}(i^n(X_1, X_2; Y) - R_1 n - R_2 n) \leq R_1^{(2)} + R_2^{(2)}\} \Big). \end{aligned}$$

Theorem 19 *The second-order (ε, R_1, R_2)-capacity region of the general MAC is given by*

$$\mathcal{C}(\varepsilon, R_1, R_2) = \mathrm{cl}\Big(\bigcup_{P_{\mathbf{X}_1} \times P_{\mathbf{X}_2}} \{(R_1^{(2)}, R_2^{(2)}) | J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \leq \varepsilon\} \Big),$$

where $\mathrm{cl}(\cdot)$ *is the closure operation of a set.*

Proof: The proof is composed of two parts.

Direct Part

We will prove the existence of a code satisfying the constraints in $\mathcal{C}(\varepsilon, R_1, R_2)$. By employing Lemma 18 and setting

$$N_1^{(n)} = e^{R_1 n + R_1^{(2)}\sqrt{n} - 2\delta\sqrt{n}},$$
$$N_2^{(n)} = e^{R_2 n + R_2^{(2)}\sqrt{n} - 2\delta\sqrt{n}},$$
$$\gamma = \delta\sqrt{n}, U = 0,$$

where δ is an arbitrary positive real number, we obtain

$$\begin{aligned}
\varepsilon^{(n)} &\leq 3e^{-\delta\sqrt{n}} + \mathbf{P}\Big(\{\frac{1}{\sqrt{n}}(i^n(X_1;Y|X_2) - R_1 n) \leq R_1^{(2)} - \delta\} \\
&\cup \{\frac{1}{\sqrt{n}}(i^n(X_2;Y|X_1) - R_2 n) \leq R_2^{(2)} - \delta\} \\
&\cup \{\frac{1}{\sqrt{n}}(i^n(X_1,X_2;Y) - R_1 n - R_2 n) \leq R_1^{(2)} + R_1^{(2)} - 3\delta\}\Big) \\
&\leq 3e^{-\delta\sqrt{n}} + \mathbf{P}\Big(\{\frac{1}{\sqrt{n}}(i^n(X_1;Y|X_2) - R_1 n) \leq R_1^{(2)}\} \cup \{\frac{1}{\sqrt{n}}(i^n(X_2;Y|X_1) - R_2 n) \leq R_2^{(2)}\} \\
&\quad \cup \{\frac{1}{\sqrt{n}}(i^n(X_1,X_2;Y) - R_1 n - R_2 n) \leq R_1^{(2)} + R_1^{(2)}\}\Big).
\end{aligned}$$

Then, from Definition 40, we have

$$\limsup_{n\to\infty} \varepsilon^{(n)} \leq J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \tag{5.13}$$

and

$$\mathcal{C}(\varepsilon, R_1, R_2) \supseteq \mathrm{cl}\Big(\bigcup_{P_{\mathbf{X}_1}\times P_{\mathbf{X}_2}} \{(R_1^{(2)}, R_2^{(2)}) | J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \leq \varepsilon\}\Big). \tag{5.14}$$

Remark 13 *Similar to [68], we can prove that the second-order capacity region of the general MAC is a convex and closed set. Hence we can include the closure operation.*

Converse Part For any second-order (ε, R_1, R_2)-achievable rate pair $(R_1^{(2)}, R_2^{(2)})$, there exists a $(n, N_1^{(n)}, N_2^{(n)}, \varepsilon^{(n)})$ code satisfying

$$\frac{1}{\sqrt{n}}(\log N_1^{(n)} - R_1 n) \geq R_1^{(2)} - \delta, \tag{5.15}$$

$$\frac{1}{\sqrt{n}}(\log N_2^{(n)} - R_2 n) \geq R_2^{(2)} - \delta, \tag{5.16}$$

$$\limsup_{n\to\infty} \varepsilon^{(n)} \leq \varepsilon. \tag{5.17}$$

Substitution of (5.15), (5.16) and $\gamma = \delta\sqrt{n}$ into Lemma 20 results in

$$\begin{aligned}
\varepsilon^{(n)} &\geq \mathbf{P}\Big(\{\frac{1}{\sqrt{n}}(i^n(X_1;Y|X_2) - R_1 n) \leq R_1^{(2)} - 2\delta\} \cup \{\frac{1}{\sqrt{n}}(i^n(X_2;Y|X_1) - R_2 n) \leq R_2^{(2)} - 2\delta\} \\
&\quad \cup \{\frac{1}{\sqrt{n}}(i^n(X_1,X_2;Y) - R_1 n - R_2 n) \leq R_1^{(2)} + R_2^{(2)} - 3\delta\}\Big) - 3e^{-\frac{\delta\sqrt{n}}{2}} \\
&\geq \mathbf{P}\Big(\{\frac{1}{\sqrt{n}}(i^n(X_1;Y|X_2) - R_1 n) \leq R_1^{(2)} - 2\delta\} \cup \{\frac{1}{\sqrt{n}}(i^n(X_2;Y|X_1) - R_2 n) \leq R_2^{(2)} - 2\delta\} \\
&\quad \cup \{\frac{1}{\sqrt{n}}(i^n(X_1,X_2;Y) - R_1 n - R_2 n) \leq R_1^{(2)} + R_2^{(2)} - 4\delta\}\Big) - 3e^{-\frac{\delta\sqrt{n}}{2}}.
\end{aligned} \tag{5.18}$$

By taking the limit of both sides in (5.18), we obtain

$$\varepsilon \geq \limsup_{n\to\infty} \varepsilon^{(n)} \geq J(R_1, R_2, R_1^{(2)} - 2\delta, R_2^{(2)} - 2\delta | \mathbf{X}_1, \mathbf{X}_2). \tag{5.19}$$

Assuming

$$(R_1^{(2)}, R_2^{(2)}) \notin \mathrm{cl}\Big(\bigcup_{P_{\mathbf{X}_1}\times P_{\mathbf{X}_2}} \{(R_1^{(2)}, R_2^{(2)}) | J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \leq \varepsilon\}\Big).$$

Then there exists $0 < \delta_0 < \min\{R_1^{(2)}, R_2^{(2)}\}$ such that

$$(R_1^{(2)} - 2\delta_0, R_2^{(2)} - 2\delta_0) \notin \mathrm{cl}\Big(\bigcup_{P_{\mathbf{X}_1}\times P_{\mathbf{X}_2}} \{(R_1^{(2)}, R_2^{(2)}) | J(R_1, R_2, R_1^{(2)}, R_2^{(2)} | \mathbf{X}_1, \mathbf{X}_2) \leq \varepsilon\}\Big).$$

Hence,

$$J(R_1,R_2,R_1^{(2)}-2\delta_0,R_2^{(2)}-2\delta_0|\mathbf{X}_1,\mathbf{X}_2)>\varepsilon,$$

which contradicts (5.19). Thus we obtain

$$\mathcal{C}(\varepsilon,R_1,R_2)\subseteq \mathrm{cl}\Big(\bigcup_{P_{\mathbf{X}_1}\times P_{\mathbf{X}_2}}\{(R_1^{(2)},R_2^{(2)})|J(R_1,R_2,R_1^{(2)},R_2^{(2)}|\mathbf{X}_1,\mathbf{X}_2)\leq\varepsilon\}\Big).$$

■

Remark 14 *From the proof, we can see the inner region given by* (5.14) *is exactly the capacity region. However, if we employ Verdú's upper bound of the average error probability in the direct proof, we will obtain a strictly smaller region in general. This illustrates the necessity to propose the modified upper bound in Lemma 18.*

Theorem 19 provides a second-order capacity region through the typical information-spectrum analysing procedure, similar to Han's first-order ε-capacity region analysis for the general MAC [65, Section VII]. In the above proof, by applying our proposed non-asymptotic bounds in Section 5.3.2, we obtain asymptotic inner and outer regions in (5.13) and (5.19), respecively.

5.4 Conclusion

In this chapter, we have applied our generalised typicality lemmas to various coding problems.

In several source coding problems, with a minor modification on the typical set according to the distortion, we have applied the conditional typicality lemma to the general rate-distortion problem, and the multivariate covering lemma to the multiple description problem. We have also applied the strong Markov lemma and the multivariate packing lemmas to the Berger-Tung problems.

In several channel coding problems. we have proved the achievability of the capacity of a general channel coding theorem with input constraint, with a minor modification on

the typical set according to the constraint. We have then applied the conditional typicality lemma to the Gelfand-Pinsker problem. We have also applied the multivariate covering lemma to the general BC coding problem.

Besides those, we have also modified our generalised typicality in a second-order fashion, and then applied it to a second-order analysis on the MAC coding problem.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

In this thesis, we have proposed and studied a generalised typicality. Specifically, we have done the following works.

In Chap. 3, we have made a summary on different approaches to prove channel coding problems. We have then pointed out that it is sufficient to study the typicality for various cases.

In Chap. 4, We have proposed a generalised definition of weak typicality for general multivariate alphabets and general measures on product spaces. We have then obtained several typicality lemmas, including conditional and joint typicality lemmas, packing and covering lemmas, as well as the strong Markov lemma, based our generalised typicality.

In Chap. 5, we have applied the typicality lemmas to some source and channel coding problems with general sources or channels, using the generalised typicality lemmas in Chap. 4. We have specified that in most cases it will be simpler to prove the coding theorems by using the generalised typicality, than using strong typicality with the discretisation-and-approximation technique.

From Chaps. 3-5, we see that our proposed generalised typicality lemmas are useful in information theoretical problems. We recognise that the discretisation-and-approximation technique have succeeded in many problems. However, there exists no axiom or theorem as-

suring the existence of a general coding theorem if there is a discrete one. The discretisation-and-approximation technique is no more than a probability theoretical trick that treat a general random variable as a limit of its quantised version. The definition and properties of integration and several fundamental theorems (ex. multiplication theorem of expectation) can be obtained in this way, and others can be based on them, but it would be not graceful if the probabilistic theorists always resort to a quantised version before they prove a general conclusion. As a comparison, our generalised typicality lemmas have provided a unified approach to coding problems with discrete, continuous or more general settings, hence we consider proofs of coding theorems will benefit from our approach,

6.2 Future Works

In this thesis, we have not considered the joint measure defined for the stochastic process with an arbitrary index set. However, inspired by some previous works on coding problems with continuous-time settings, we can extend our typicality definition and typicality lemmas to more general cases. Non-asymptotic case is another possible topic, which is related to second-order analysis.

6.2.1 Continuous-Time Case

The distribution on a product space is equivalent to one of a discrete time series. Furthermore, it is possible to define a typicality for distributions of more general stochastic processes, which are not necessarily discrete time series. Historically, this was first mentioned in Goldman's book [69, Chap. III], as a notion of "typical function". In [70], the authors used a set which was essentially an continuous time extension of the weak typicality set. Inspired by previous works, we expect to extend our proposed typicality to the more general stochastic process.

More specifically, given a stochastic process $(\mathbf{X},\mathbf{Y}) = \{(X_t,Y_t)|t \in \mathbf{R}\}$ with state space $(\mathcal{X}\times\mathcal{Y},\mathcal{B}_X\times\mathcal{B}_Y)$, and the probability law $P_{\mathbf{XY}}$ as well as induced marginals $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$, then we can define the mutual information between $\{X_t|t\in\mathbf{R}\}$ and $\{Y_t|t\in\mathbf{R}\}$ as $I(\mathbf{X};\mathbf{Y}) =$

$\mathbf{E}[i(\mathbf{X};\mathbf{Y})]$, where the mutual information density

$$i(x;y) = \log \frac{\mathrm{d}P_{\mathbf{XY}}}{\mathrm{d}P_{\mathbf{X}} \times P_{\mathbf{Y}}}(\mathbf{x},\mathbf{y}).$$

We denote $\{X_t | t \in (-\infty, t]\}$ by $\mathbf{X}^t$, then the set sequence $\mathscr{A}^{(t)}$ defined by

$$\mathscr{A}^{(t)} = \left\{ (x^t, y^t) \mid \left| \frac{1}{t} i(x^t, y^t) - I(\mathbf{X};\mathbf{Y}) \right| \leq \varepsilon \right\}$$

for all $t \in \mathbf{R}$ is a generic set sequence in the meaning of

$$\lim_{t\to\infty} P^{(t)}(\mathcal{A}^{(t)}) = 1.$$

Hence, it is possible to extend our study on the generalised typicality to the continuous-time case.

6.2.2 Non-Asymptotic Case

The typicality lemmas we have studied are all in the asymptotic regime. There have also been researches on non-asymptotic typicality lemmas. Recently, Verdú, Liu, and Cuff has proved non-asymptotic typicality, covering and mutual covering lemmas [26, 27, 66] and showed some applications in multi-terminal problems. [71] also provided a non-asymptotic mutual covering lemma. The alternative proof of Lem. 3 in Chap. 4 shows the possibility to relate asymptotic typicality lemmas with their non-asymptotic counterparts.

Besides, as mentioned in Sec. 3.2 and studied in Sec. 5.3, we can also make second-order analysis on general source and channel coding problems.

References

[1] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[2] A. Feinstein, *Foundations of Information Theory*, ser. McGraw-Hill Electr. Electron. Eng. McGraw-Hill, 1958.

[3] R. Ahlswede, *Transmitting and Gaining Data*, ser. Found. Signal Process Commun. Netw., A. Ahlswede, I. Althöfer, C. Deppe, and U. Tamm, Eds. Springer Int. Publ., 2015, no. 11.

[4] T. S. Han, *Information-Spectrum Methods in Information Theory*, ser. Stoch. Model. Appl. Probab. Springer-Verlag, 2003, no. 50, originally published in Japanese in 1998, translated by H. Koga.

[5] H. Joe, "Relative entropy measures of multivariate dependence," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 157–164, 1989.

[6] E. Çinlar, *Probability and Stochastics*, ser. Grad. Texts Math. Springer, 2011, no. 261.

[7] R. L. Dobrushin, "A general formulation of the fundamental theorem of Shannon in the theory of information," *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, Nov.-Dec. 1959.

[8] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed., ser. Ergeb. Math. Grenzgeb. Springer-Verlag, 1978, no. 31.

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[10] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[12] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4181–4206, Sept. 2010.

[13] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, ser. Holden-Day Ser. Time Ser. Anal. Holden-Day, 1964, originally published in Russian in 1960, translated and edited by A. Feinstein.

[14] R. W. Yeung, *Information Theory and Network Coding*, ser. Inf. Technol. Transm. Process. Storage. Springer, 2008.

[15] P. Mitran, "On a Markov lemma and typical sequences for Polish alphabets," *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5342–5356, Oct. 2015.

[16] M. Raginsky, "Empirical processes, typical sequences, and coordinated actions in standard Borel spaces," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1288–1301, Mar. 2013.

[17] J. Jeon, "A generalized typicality for abstract alphabets," *CoRR*, vol. abs/1401.6728, 2014. [Online]. Available: http://arxiv.org/abs/1401.6728

[18] A. El Gamal and E. C. van der Meulen, "A proof of Marton's coding theorem for the discrete memoryless broadcast channel (corresp.)," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 120–122, Jan. 1981.

[19] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 3, pp. 306–311, May 1979.

[20] S. I. Gelfand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Control Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[21] R. Ahlswede, *Storing and Transmitting Data*, ser. Found. Signal Process Commun. Netw., A. Ahlswede, I. Althöfer, C. Deppe, and U. Tamm, Eds. Springer Int. Publ., 2014, no. 10.

[22] P. Noorzad, M. Effros, and M. Langberg, "The multivariate covering lemma and its converse," *CoRR*, vol. abs/1508.03349, Aug. 2015. [Online]. Available: http://arxiv.org/abs/1508.03349

[23] K. B. Viswanatha, E. Akyol, and K. Rose, "Subset typicality lemmas and improved achievable regions in multiterminal source coding," *CoRR*, vol. abs/1205.1173, 2012. [Online]. Available: http://arxiv.org/abs/1205.1173

[24] S.-W. Ho, "Markov lemma for countable alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, Texas, U.S.A.,, June 2010, pp. 1448–1452.

[25] P. Piantanida, L. R. Vega, and A. O. Hero III, "A proof of the generalized Markov lemma with countable infinite sources," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, June 2014, pp. 591–595.

[26] S. Verdú, "Non-asymptotic covering lemmas," in *Proc. Inf. Theory Workshop*, Jeruzalem, Israel, Apr. 2015.

[27] J. Liu, P. W. Cuff, and S. Verdú, "One-shot mutual covering lemma and Marton's inner bound with a common message," *CoRR*, vol. abs/1504.04092, 2015. [Online]. Available: http://arxiv.org/abs/1504.04092

[28] W. Rudin, *Real and Complex Analysis*, 3rd ed. McGraw-Hill, 1987.

[29] A. Klenke, *Probability Theory*, 2nd ed., ser. Universitext. Springer, 2014.

[30] S. Kotz, "Recent results in information theory," *J. Appl. Probab.*, vol. 3, no. 1, pp. 1–93, June 1966.

[31] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," 2017.

[32] S. Verdú and T. S. Han, "The role of the asymptotic equipartition property in noiseless source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 847–857, May 1997.

[33] B. McMillan, "The basic theorems of information theory," *Ann. Math. Stat.*, vol. 24, no. 2, pp. 196–219, 6 1953.

[34] R. B. Ash, *Information Theory*, ser. Intersci. Tracts Pure Appl. Math. Wiley-Interscience, 1965, no. 19.

[35] R. E. Blahut, *Principles and Practice of Information Theory*, ser. Addison-Wesley Electr. Comput. Eng. Addison-Wesley, 1987.

[36] A. N. Shiryaev, *Probability*, 2nd ed., ser. Grad. Texts Math. Springer-Verlag, 1996, no. 95, originally published in Russian in 1989, translated by R. P. Boas.

[37] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3-4, pp. 379–423, 623–656, July-Oct. 1948.

[38] S. Verdú, "Fifty years of Shannon theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2057–2078, Oct. 1998.

[39] A. El Gamal and T. M. Cover, "Multiple user information theory," *Proc. IEEE*, vol. 68, no. 12, pp. 1466–1483, Dec. 1980.

[40] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, p. 49–60, Jan. 1981.

[41] G. Kramer, "Topics in multi-user information theory," *Foundations and Trends® in Communications and Information Theory*, vol. 4, no. 4-5, pp. 265–444, 2008.

[42] R. G. Gallager, *Principles of Digital Communication*. Cambridge University Press, 2008.

[43] M. M. Wilde, *Quantum Information Theory*. Cambridge University Press, 2013.

[44] S. Verdú, "Teaching it," *IEEE Inf. Theory Soc. Newsl.*, vol. 57, no. 4, pp. 1,8–9, Dec. 2007.

[45] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory,*, vol. 4, no. 4, pp. 2–22, Sept. 1954.

[46] A. Y. Khinchin, *Mathematical Foundations of Information Theory*, ser. Dover Books Math. Dover Publ., 1957, originally published in Russian as two separate papers in 1953 and 1956, translated by R. A. Silverman and M. D. Friedman.

[47] K. Takano, "On the basic theorems of information theory," *Ann. Inst. Statist. Math.*, vol. 9, no. 1, pp. 53–77, Dec. 1957.

[48] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Stat.*, vol. 29, no. 4, pp. 1209–1220, Dec. 1958.

[49] S. Guiaşu, *Information Theory with Applications*, ser. Adv. Book Program. McGraw-Hill Books Company, 1977.

[50] R. M. Gray, *Entropy and Information Theory*, 2nd ed. Springer, 2011.

[51] C. E. Shannon, "Certain results in coding theory for noisy channels," *Inf. Contr.*, vol. 1, no. 1, pp. 6 – 25, Sept. 1957.

[52] M. Rosenblatt-Roth, "The concept of entropy in probability theory and its application in the theory of information transmission through communication channels," *Theory Prob. Its Appl.*, vol. 9, no. 2, pp. 212–235, 1964.

[53] Y. Kakihara, *Abstract Methods in Information Theory*, 2nd ed., ser. Multivariate Anal. World Sci. Publ., 2016, no. 10.

[54] A. R. Barron, "The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem," *Ann. Probab.*, vol. 13, pp. 1292–1303, 1985.

[55] M. S. Pinsker, "Some mathematical questions of theory of information transmission," *Probl. Inf. Transm.*, vol. 43, no. 4, pp. 380–392, 2007.

[56] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[57] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[58] T. Tao, *Topics in Random Matrix Theory*, ser. Grad. Stud. Math. Amer. Math. Soc., 2012, vol. 132.

[59] A. Somekh-Baruch, "A general formula for the mismatch capacity," *CoRR*, vol. abs/1309.7964, 2013. [Online]. Available: http://arxiv.org/abs/1309.7964

[60] W. Liu, X. Chu, and J. Zhang, "On a generalised typicality with respect to general probability distributions," in *Proc. IEEE 14th Canadian Workshop Inf. Theory*, July 2015, pp. 165–169.

[61] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 63–86, Jan. 1996.

[62] V. Y. F. Tan, "A formula for the capacity of the general Gel'fand-Pinsker channel," *IEEE Trans. Commun.*, vol. 62, no. 6, pp. 1857–1870, June 2014.

[63] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

[64] K. Iwata and Y. Oohama, "Information-spectrum characterization of broadcast channel with general source," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E88-A, no. 10, pp. 2808–2818, Oct. 2005.

[65] T. S. Han, "An information-spectrum approach to capacity theorems for the general multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2773–2795, Nov. 1998.

[66] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *Proc. 50th Annu. Allerton Conf. Commun. Contr. Comput.*, 2012, pp. 1–8.

[67] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[68] T. S. Han, "The capacity region of general multiple-access channel with certain correlated sources," *Inf. Contr.*, vol. 40, no. 1, pp. 37–60, Jan. 1979.

[69] S. Goldman, *Information Theory*, ser. Prentice-Hall Electr. Eng. Prentice-Hall, 1953.

[70] T. Kadota and A. D. Wyner, "Coding theorem for stationary, asymptotically memoryless, continuous-time channels," *Ann. Math. Stat.*, pp. 1603–1611, 1972.

[71] J. Radhakrishnan, P. Sen, and N. Warsi, "One-shot Marton inner bound for classical-quantum broadcast channel," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2836–2848, May 2016.