

Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities

Vincent Y. F. Tan

Department of Electrical and Computer Engineering

Department of Mathematics

National University of Singapore

Singapore 119077

Email: vtan@nus.edu.sg

now

the essence of knowledge

Boston — Delft



Documents

Document Sharing Hub

Foundations and Trends[®] in Communications and Information Theory

Published, sold and distributed by:

now Publishers Inc.

PO Box 1024

Hanover, MA 02339

United States

Tel. +1-781-985-4510

www.nowpublishers.com

sales@nowpublishers.com

Outside North America:

now Publishers Inc.

PO Box 179

2600 AD Delft

The Netherlands

Tel. +31-6-51115274

The preferred citation for this publication is

V. Y. F. Tan. *Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities*. Foundations and Trends[®] in Communications and Information Theory, vol. 11, no. 1-2, pp. 1–184, 2014.

This Foundations and Trends[®] issue was typeset in L^AT_EX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-853-9

© 2014 V. Y. F. Tan

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Foundations and Trends® in Communications and Information Theory

Volume 11, Issue 1-2, 2014

Editorial Board

Editor-in-Chief

Sergio Verdú
Princeton University
United States

Editors

Venkat Anantharam <i>UC Berkeley</i>	Tara Javidi <i>UC San Diego</i>	Shlomo Shamai <i>Technion</i>
Helmut Bölcskei <i>ETH Zurich</i>	Ioannis Kontoyiannis <i>Athens University of Economy and Business</i>	Amin Shokrollahi <i>EPF Lausanne</i>
Giuseppe Caire <i>USC</i>	Gerhard Kramer <i>TU Munich</i>	Yossef Steinberg <i>Technion</i>
Daniel Costello <i>University of Notre Dame</i>	Sanjeev Kulkarni <i>Princeton University</i>	Wojciech Szpankowski <i>Purdue University</i>
Anthony Ephremides <i>University of Maryland</i>	Amos Lapidoth <i>ETH Zurich</i>	David Tse <i>UC Berkeley</i>
Alex Grant <i>University of South Australia</i>	Bob McEliece <i>Caltech</i>	Antonia Tulino <i>Alcatel-Lucent Bell Labs</i>
Andrea Goldsmith <i>Stanford University</i>	Muriel Medard <i>MIT</i>	Rüdiger Urbanke <i>EPF Lausanne</i>
Albert Guillen i Fabregas <i>Pompeu Fabra University</i>	Neri Merhav <i>Technion</i>	Emanuele Viterbo <i>Monash University</i>
Dongning Guo <i>Northwestern University</i>	David Neuhoff <i>University of Michigan</i>	Tsachy Weissman <i>Stanford University</i>
Dave Forney <i>MIT</i>	Alon Orlitsky <i>UC San Diego</i>	Frans Willems <i>TU Eindhoven</i>
Te Sun Han <i>University of Tokyo</i>	Yury Polyanskiy <i>MIT</i>	Raymond Yeung <i>CUHK</i>
Babak Hassibi <i>Caltech</i>	Vincent Poor <i>Princeton University</i>	Bin Yu <i>UC Berkeley</i>
Michael Honig <i>Northwestern University</i>	Maxim Raginsky <i>UIUC</i>	
Johannes Huber <i>University of Erlangen</i>	Kannan Ramchandran <i>UC Berkeley</i>	



Documents
Document Sharing Hub

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2014, Volume 11, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.

Foundations and Trends® in Communications and
Information Theory
Vol. 11, No. 1-2 (2014) 1–184
© 2014 V. Y. F. Tan
DOI: 10.1561/01000000086



Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities

Vincent Y. F. Tan
Department of Electrical and Computer Engineering
Department of Mathematics
National University of Singapore
Singapore 119077
Email: vtan@nus.edu.sg

Contents

I	Fundamentals	2
1	Introduction	3
1.1	Motivation for this Monograph	5
1.2	Preview of this Monograph	7
1.3	Fundamentals of Information Theory	10
1.4	The Method of Types	13
1.5	Probability Bounds	15
2	Binary Hypothesis Testing	22
2.1	Non-Asymptotic Quantities and Their Properties	23
2.2	Asymptotic Expansions	28
II	Point-To-Point Communication	33
3	Source Coding	34
3.1	Lossless Source Coding: Non-Asymptotic Bounds	35
3.2	Lossless Source Coding: Asymptotic Expansions	37
3.3	Second-Order Asymptotics of Lossless Source Coding via the Method of Types	39

3.4	Lossy Source Coding: Non-Asymptotic Bounds	42
3.5	Lossy Source Coding: Asymptotic Expansions	46
3.6	Second-Order Asymptotics of Lossy Source Coding via the Method of Types	48
4	Channel Coding	52
4.1	Definitions and Non-Asymptotic Bounds	53
4.2	Asymptotic Expansions for Discrete Memoryless Channels .	58
4.3	Asymptotic Expansions for Gaussian Channels	71
4.4	A Digression: Third-Order Asymptotics vs Error Exponent Prefactors	77
4.5	Joint Source-Channel Coding	78
III	Network Information Theory	83
5	Channels with Random State	84
5.1	Random State at the Decoder	85
5.2	Random State at the Encoder and Decoder	87
5.3	Writing on Dirty Paper	90
5.4	Mixed Channels	96
5.5	Quasi-Static Fading Channels	101
6	Distributed Lossless Source Coding	105
6.1	Definitions and Non-Asymptotic Bounds	106
6.2	Second-Order Asymptotics	107
6.3	Second-Order Asymptotics of Slepian-Wolf Coding via the Method of Types	114
6.4	Other Fixed Error Asymptotic Notions	117
7	A Special Class of Gaussian Interference Channels	120
7.1	Definitions and Non-Asymptotic Bounds	123
7.2	Second-Order Asymptotics	126
7.3	Proof Sketch of the Main Result	129
8	A Special Class of Gaussian Multiple Access Channels	136
8.1	Definitions and Non-Asymptotic Bounds	139

8.2	Second-Order Asymptotics	141
8.3	Proof Sketches of the Main Results	148
8.4	Difficulties in the Fixed Error Analysis for the MAC	156
9	Summary, Other Results, Open Problems	158
9.1	Summary and Other Results	158
9.2	Open Problems and Challenges Ahead	163
	Acknowledgements	167
	References	169

Abstract

This monograph presents a unified treatment of single- and multi-user problems in Shannon's information theory where we depart from the requirement that the error probability decays asymptotically in the blocklength. Instead, the error probabilities for various problems are bounded above by a non-vanishing constant and the spotlight is shone on achievable coding rates as functions of the growing blocklengths. This represents the study of *asymptotic estimates with non-vanishing error probabilities*.

In Part I, after reviewing the fundamentals of information theory, we discuss Strassen's seminal result for binary hypothesis testing where the type-I error probability is non-vanishing and the rate of decay of the type-II error probability with growing number of independent observations is characterized. In Part II, we use this basic hypothesis testing result to develop second- and sometimes, even third-order asymptotic expansions for point-to-point communication. Finally in Part III, we consider network information theory problems for which the second-order asymptotics are known. These problems include some classes of channels with random state, the multiple-encoder distributed lossless source coding (Slepian-Wolf) problem and special cases of the Gaussian interference and multiple-access channels. Finally, we discuss avenues for further research.

V. Y. F. Tan. *Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities*. Foundations and Trends® in Communications and Information Theory, vol. 11, no. 1-2, pp. 1–184, 2014.

DOI: 10.1561/01000000086.

Part I

Fundamentals

1

Introduction

Claude E. Shannon's epochal "*A Mathematical Theory of Communication*" [141] marks the dawn of the digital age. In his seminal paper, Shannon laid the theoretical and mathematical foundations for the basis of all communication systems today. It is not an exaggeration to say that his work has had a tremendous impact in communications engineering and beyond, in fields as diverse as statistics, economics, biology and cryptography, just to name a few.

It has been more than 65 years since Shannon's landmark work was published. Along with impressive research advances in the field of *information theory*, numerous excellent books on various aspects of the subject have been written. The author's favorites include Cover and Thomas [33], Gallager [56], Csiszár and Körner [39], Han [67], Yeung [189] and El Gamal and Kim [49]. Is there sufficient motivation to consolidate and present another aspect of information theory systematically? It is the author's hope that the answer is in the affirmative.

To motivate why this is so, let us recapitulate two of Shannon's major contributions in his 1948 paper. First, Shannon showed that to *reliably* compress a discrete memoryless source (DMS) $X^n = (X_1, \dots, X_n)$ where each X_i has the same distribution as a common random vari-

able X , it is sufficient to use $H(X)$ bits per source symbol in the limit of large blocklengths n , where $H(X)$ is the Shannon entropy of the source. By *reliable*, it is meant that the probability of incorrect decoding of the source sequence tends to zero as the blocklength n grows. Second, Shannon showed that it is possible to *reliably* transmit a message $M \in \{1, \dots, 2^{nR}\}$ over a discrete memoryless channel (DMC) W as long as the message rate R is smaller than the capacity of the channel $C(W)$. Similarly to the source compression scenario, by *reliable*, one means that the probability of incorrectly decoding M tends to zero as n grows.

There is, however, substantial motivation to revisit the criterion of having error probabilities vanish asymptotically. To state Shannon's source compression result more formally, let us define $M^*(P^n, \varepsilon)$ to be the minimum code size for which the length- n DMS P^n is compressible to within an error probability $\varepsilon \in (0, 1)$. Then, Theorem 3 of Shannon's paper [141], together with the strong converse for lossless source coding [49, Ex. 3.15], states that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M^*(P^n, \varepsilon) = H(X), \quad \text{bits per source symbol.} \quad (1.1)$$

Similarly, denoting $M_{\text{ave}}^*(W^n, \varepsilon)$ as the maximum code size for which it is possible to communicate over a DMC W^n such that the average error probability is no larger than ε , Theorem 11 of Shannon's paper [141], together with the strong converse for channel coding [180, Thm. 2], states that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M_{\text{ave}}^*(W^n, \varepsilon) = C(W), \quad \text{bits per channel use.} \quad (1.2)$$

In many practical communication settings, one does not have the luxury of being able to design an arbitrarily long code, so one must settle for a non-vanishing, and hence finite, error probability ε . In this *finite blocklength* and *non-vanishing error probability* setting, how close can one hope to get to the asymptotic limits $H(X)$ and $C(W)$? This is, in general a difficult question because exact evaluations of $\log M^*(P^n, \varepsilon)$ and $\log M_{\text{ave}}^*(W^n, \varepsilon)$ are intractable, apart from a few special sources and channels.

In the early years of information theory, Dobrushin [45], Kemperman [91] and, most prominently, Strassen [152] studied approxima-

tions to $\log M^*(P^n, \varepsilon)$ and $\log M_{\text{ave}}^*(W^n, \varepsilon)$. These beautiful works were largely forgotten until recently, when interest in so-called *Gaussian approximations* were revived by Hayashi [75, 76] and Polyanskiy-Poor-Verdú [122, 123].¹ Strassen showed that the limiting statement in (1.1) may be refined to yield the *asymptotic expansion*

$$\log M^*(P^n, \varepsilon) = nH(X) - \sqrt{nV(X)}\Phi^{-1}(\varepsilon) - \frac{1}{2}\log n + O(1), \quad (1.3)$$

where $V(X)$ is known as the *source dispersion* or the *varentropy*, terms introduced by Kostina-Verdú [97] and Kontoyiannis-Verdú [95]. In (1.3), Φ^{-1} is the inverse of the Gaussian cumulative distribution function. Observe that the first-order term in the asymptotic expansion above, namely $H(X)$, coincides with the (first-order) fundamental limit shown by Shannon. From this expansion, one sees that if the error probability is fixed to $\varepsilon < \frac{1}{2}$, the extra rate above the entropy we have to pay for operating at finite blocklength n with admissible error probability ε is approximately $\sqrt{V(X)/n}\Phi^{-1}(1 - \varepsilon)$. Thus, the quantity $V(X)$, which is a function of P just like the entropy $H(X)$, quantifies how fast the rates of optimal source codes converge to $H(X)$. Similarly, for well-behaved DMCs, under mild conditions, Strassen showed that the limiting statement in (1.2) may be refined to

$$\log M_{\text{ave}}^*(W^n, \varepsilon) = nC(W) + \sqrt{nV_\varepsilon(W)}\Phi^{-1}(\varepsilon) + O(\log n) \quad (1.4)$$

and $V_\varepsilon(W)$ is a channel parameter known as the ε -*channel dispersion*, a term introduced by Polyanskiy-Poor-Verdú [123]. Thus the backoff from capacity at finite blocklengths n and average error probability ε is approximately $\sqrt{V_\varepsilon(W)/n}\Phi^{-1}(1 - \varepsilon)$.

1.1 Motivation for this Monograph

It turns out that Gaussian approximations (first two terms of (1.3) and (1.4)) are *good proxies* to the true non-asymptotic fundamental limits ($\log M^*(P^n, \varepsilon)$ and $\log M_{\text{ave}}^*(W^n, \varepsilon)$) at moderate blocklengths and

¹Some of the results in [122, 123] were already announced by S. Verdú in his Shannon lecture at the 2007 International Symposium on Information Theory (ISIT) in Nice, France.

moderate error probabilities for some channels and sources as shown by Polyanskiy-Poor-Verdú [123] and Kostina-Verdú [97]. For error probabilities that are not too small (e.g., $\varepsilon \in [10^{-6}, 10^{-3}]$), the Gaussian approximation is often better than that provided by traditional *error exponent* or *reliability function* analysis [39, 56], where the code rate is fixed (below the first-order fundamental limit) and the exponential decay of the error probability is analyzed. Recent refinements to error exponent analysis using exact asymptotics [10, 11, 135] or saddlepoint approximations [137] are alternative proxies to the non-asymptotic fundamental limits. The accuracy of the Gaussian approximation in *practical* regimes of errors and finite blocklengths gives us motivation to study refinements to the first-order fundamental limits of other single- and multi-user problems in Shannon theory.

The study of *asymptotic estimates with non-vanishing error probabilities*—or more succinctly, *fixed error asymptotics*—also uncovers several interesting phenomena that are not observable from studies of first-order fundamental limits in single- and multi-user information theory [33, 49]. This analysis may give engineers deeper insight into the design of practical communication systems. A non-exhaustive list includes:

1. Shannon showed that *separating* the tasks of source and channel coding is optimal rate-wise [141]. As we see in Section 4.5.2 (and similarly to the case of error exponents [35]), this is not the case when the probability of excess distortion of the source is allowed to be non-vanishing.
2. Shannon showed that feedback does not increase the capacity of a DMC [142]. It is known, however, that variable-length feedback [125] and full output feedback [8] improve on the fixed error asymptotics of DMCs.
3. It is known that the entropy can be achieved *universally* for fixed-to-variable length almost lossless source coding of a DMS [192], i.e., the source statistics do not have to be known. The redundancy has also been studied for prefix-free codes [27]. In the fixed error setting (a setting complementary to [27]), it was shown by

Kosut and Sankar [100, 101] that universality imposes a penalty in the *third-order* term of the asymptotic expansion in (1.3).

4. Han showed that the output from any source encoder at the optimal coding rate with asymptotically vanishing error appears almost completely random [68]. This is the so-called *folklore theorem*. Hayashi [75] showed that the analogue of the folklore theorem does not hold when we consider the second-order terms in asymptotic expansions (i.e., the second-order asymptotics).
5. Slepian and Wolf showed that separate encoding of two correlated sources incurs no loss rate-wise compared to the situation where side information is also available at all encoders [151]. As we shall see in Chapter 6, the fixed error asymptotics in the vicinity of a corner point of the polygonal Slepian-Wolf region suggests that side-information at the encoders may be beneficial.

None of the aforementioned books [33, 39, 49, 56, 67, 189] focus exclusively on the situation where the error probabilities of various Shannon-theoretic problems are upper bounded by $\varepsilon \in (0, 1)$ and asymptotic expansions or second-order terms are sought. This is what this monograph attempts to do.

1.2 Preview of this Monograph

This monograph is organized as follows: In the remaining parts of this chapter, we recap some quantities in information theory and results in the *method of types* [37, 39, 74], a particularly useful tool for the study of discrete memoryless systems. We also mention some probability bounds that will be used throughout the monograph. Most of these bounds are based on refinements of the central limit theorem, and are collectively known as *Berry-Esseen theorems* [17, 52]. In Chapter 2, our study of asymptotic expansions of the form (1.3) and (1.4) begins in earnest by revisiting Strassen's work [152] on binary hypothesis testing where the probability of false alarm is constrained to not exceed a positive constant. We find it useful to revisit the fundamentals of hypothesis testing as many information-theoretic problems such as source

and channel coding are intimately related to hypothesis testing.

Part II of this monograph begins our study of information-theoretic problems starting with lossless and lossy compression in Chapter 3. We emphasize, in the first part of this chapter, that (fixed-to-fixed length) lossless source coding and binary hypothesis testing are, in fact, the same problem, and so the asymptotic expansions developed in Chapter 2 may be directly employed for the purpose of lossless source coding. Lossy source coding, however, is more involved. We review the recent works in [86] and [97], where the authors independently derived asymptotic expansions for the logarithm of the minimum size of a source code that reproduces symbols up to a certain distortion, with some admissible probability of excess distortion. Channel coding is discussed in Chapter 4. In particular, we study the approximation in (1.4) for both discrete memoryless and Gaussian channels. We make it a point here to be precise about the third-order $O(\log n)$ term. We state conditions on the channel under which the coefficient of the $O(\log n)$ term can be determined exactly. This leads to some new insights concerning optimum codes for the channel coding problem. Finally, we marry source and channel coding in the study of source-channel transmission where the probability of excess distortion in reproducing the source is non-vanishing.

Part III of this monograph contains a sparse sampling of fixed error asymptotic results in network information theory. The problems we discuss here have conclusive second-order asymptotic characterizations (analogous to the second terms in the asymptotic expansions in (1.3) and (1.4)). They include some channels with random state (Chapter 5), such as Costa's writing on dirty paper [30], mixed DMCs [67, Sec. 3.3], and quasi-static single-input-multiple-output (SIMO) fading channels [18]. Under the fixed error setup, we also consider the second-order asymptotics of the Slepian-Wolf [151] distributed lossless source coding problem (Chapter 6), the Gaussian interference channel (IC) in the strictly very strong interference regime [22] (Chapter 7), and the Gaussian multiple access channel (MAC) with degraded message sets (Chapter 8). The MAC with degraded message sets is also known as the *cognitive* [44] or *asymmetric* [72, 167, 128] MAC (A-MAC). Chapter 9

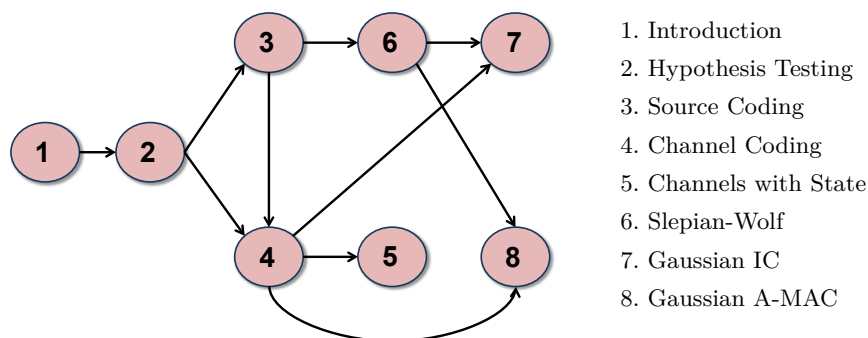


Figure 1.1: Dependence graph of the chapters in this monograph. An arrow from node s to t means that results and techniques in Chapter s are required to understand the material in Chapter t .

concludes with a brief summary of other results, together with open problems in this area of research. A dependence graph of the chapters in the monograph is shown in Fig. 1.1.

This area of information theory—*fixed error asymptotics*—is vast and, at the same time, rapidly expanding. The results described herein are not meant to be exhaustive and were somewhat dependent on the author’s understanding of the subject and his preferences at the time of writing. However, the author has made it a point to ensure that results herein are *conclusive* in nature. This means that the problem is *solved* in the information-theoretic sense in that an operational quantity is *equated* to an information quantity. In terms of asymptotic expansions such as (1.3) and (1.4), by *solved*, we mean that either the second-order term is known or, better still, both the second- and third-order terms are known. Having articulated this, the author confesses that there are many relevant information-theoretic problems that can be considered solved in the fixed error setting, but have not found their way into this monograph either due to space constraints or because it was difficult to meld them seamlessly with the rest of the story.

1.3 Fundamentals of Information Theory

In this section, we review some basic information-theoretic quantities. As with every article published in the *Foundations and Trends in Communications and Information Theory*, the reader is expected to have some background in information theory. Nevertheless, the only prerequisite required to appreciate this monograph is information theory at the level of Cover and Thomas [33]. We will also make extensive use of the method of types, for which excellent expositions can be found in [37, 39, 74]. The measure-theoretic foundations of probability will not be needed to keep the exposition accessible to as wide an audience as possible.

1.3.1 Notation

The notation we use is reasonably standard and generally follows the books by Csiszár-Körner [39] and Han [67]. Random variables (e.g., X) and their realizations (e.g., x) are in upper and lower case respectively. Random variables that take on finitely many values have alphabets (support) that are denoted by calligraphic font (e.g., $\mathcal{X}$). The cardinality of the finite set $\mathcal{X}$ is denoted as $|\mathcal{X}|$. Let the random vector X^n be the vector of random variables $(X_1, \dots, X_n)$. We use bold face $\mathbf{x} = (x_1, \dots, x_n)$ to denote a realization of X^n . The set of all distributions (probability mass functions) supported on alphabet $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. The set of all conditional distributions (i.e., channels) with the input alphabet $\mathcal{X}$ and the output alphabet $\mathcal{Y}$ is denoted by $\mathcal{P}(\mathcal{Y}|\mathcal{X})$. The joint distribution induced by a marginal distribution $P \in \mathcal{P}(\mathcal{X})$ and a channel $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ is denoted as $P \times V$, i.e.,

$$(P \times V)(x, y) := P(x)V(y|x). \quad (1.5)$$

The marginal output distribution induced by P and V is denoted as PV , i.e.,

$$PV(y) := \sum_{x \in \mathcal{X}} P(x)V(y|x). \quad (1.6)$$

If X has distribution P , we sometimes write this as $X \sim P$.

Vectors are indicated in lower case bold face (e.g., $\mathbf{a}$) and matrices in upper case bold face (e.g., $\mathbf{A}$). If we write $\mathbf{a} \geq \mathbf{b}$ for two vectors $\mathbf{a}$

and $\mathbf{b}$ of the same length, we mean that $a_j \geq b_j$ for every coordinate j . The transpose of $\mathbf{A}$ is denoted as $\mathbf{A}'$. The vector of all zeros and the identity matrix are denoted as $\mathbf{0}$ and $\mathbf{I}$ respectively. We sometimes make the lengths and sizes explicit. The ℓ_q -norm (for $q \geq 1$) of a vector $\mathbf{v} = (v_1, \dots, v_k)$ is denoted as $\|\mathbf{v}\|_q := (\sum_{i=1}^k |v_i|^q)^{1/q}$.

We use standard asymptotic notation [29] in this monograph: $a_n \in O(b_n)$ if and only if (iff) $\limsup_{n \rightarrow \infty} |a_n/b_n| < \infty$; $a_n \in \Omega(b_n)$ iff $b_n \in O(a_n)$; $a_n \in \Theta(b_n)$ iff $a_n \in O(b_n) \cap \Omega(b_n)$; $a_n \in o(b_n)$ iff $\limsup_{n \rightarrow \infty} |a_n/b_n| = 0$; and $a_n \in \omega(b_n)$ iff $\liminf_{n \rightarrow \infty} |a_n/b_n| = \infty$. Finally, $a_n \sim b_n$ iff $\lim_{n \rightarrow \infty} a_n/b_n = 1$.

1.3.2 Information-Theoretic Quantities

Information-theoretic quantities are denoted in the usual way [39, 49]. All logarithms and exponential functions are to the base 2. The *entropy* of a discrete random variable X with probability distribution $P \in \mathcal{P}(\mathcal{X})$ is denoted as

$$H(X) = H(P) := - \sum_{x \in \mathcal{X}} P(x) \log P(x). \quad (1.7)$$

For the sake of clarity, we will sometimes make the dependence on the distribution P explicit. Similarly given a pair of random variables (X, Y) with joint distribution $P \times V \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, the *conditional entropy* of Y given X is written as

$$H(Y|X) = H(V|P) := - \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} V(y|x) \log V(y|x). \quad (1.8)$$

The *joint entropy* is denoted as

$$H(X, Y) := H(X) + H(Y|X), \quad \text{or} \quad (1.9)$$

$$H(P \times V) := H(P) + H(V|P). \quad (1.10)$$

The *mutual information* is a measure of the correlation or dependence between random variables X and Y . It is interchangeably denoted as

$$I(X; Y) := H(Y) - H(Y|X), \quad \text{or} \quad (1.11)$$

$$I(P, V) := H(PV) - H(V|P). \quad (1.12)$$

Given three random variables (X, Y, Z) with joint distribution $P \times V \times W$ where $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and $W \in \mathcal{P}(\mathcal{Z}|\mathcal{X} \times \mathcal{Y})$, the *conditional mutual information* is

$$I(Y; Z|X) := H(Z|X) - H(Z|XY), \quad \text{or} \quad (1.13)$$

$$I(V, W|P) := \sum_{x \in \mathcal{X}} P(x) I(V(\cdot|x), W(\cdot|x, \cdot)). \quad (1.14)$$

A particularly important quantity is the *relative entropy* (or *Kullback-Leibler divergence* [102]) between P and Q which are distributions on the same finite support set $\mathcal{X}$. It is defined as the expectation with respect to P of the log-likelihood ratio $\log \frac{P(x)}{Q(x)}$, i.e.,

$$D(P\|Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \quad (1.15)$$

Note that if there exists an $x \in \mathcal{X}$ for which $Q(x) = 0$ while $P(x) > 0$, then the relative entropy $D(P\|Q) = \infty$. If for every $x \in \mathcal{X}$, if $Q(x) = 0$ then $P(x) = 0$, we say that P is absolutely continuous with respect to Q and denote this relation by $P \ll Q$. In this case, the relative entropy is finite. It is well known that $D(P\|Q) \geq 0$ and equality holds if and only if $P = Q$. Additionally, the *conditional relative entropy* between $V, W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ given $P \in \mathcal{P}(\mathcal{X})$ is defined as

$$D(V\|W|P) := \sum_{x \in \mathcal{X}} P(x) D(V(\cdot|x)\|W(\cdot|x)). \quad (1.16)$$

The mutual information is a special case of the relative entropy. In particular, we have

$$I(P, V) = D(P \times V\|P \times PV) = D(V\|PV|P). \quad (1.17)$$

Furthermore, if $U_{\mathcal{X}}$ is the uniform distribution on $\mathcal{X}$, i.e., $U_{\mathcal{X}}(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$, we have

$$D(P\|U_{\mathcal{X}}) = -H(P) + \log |\mathcal{X}|. \quad (1.18)$$

The definition of relative entropy $D(P\|Q)$ can be extended to the case where Q is not necessarily a probability measure. In this case non-negativity does not hold in general. An important property we exploit is the following: If μ denotes the counting measure (i.e., $\mu(\mathcal{A}) = |\mathcal{A}|$ for $\mathcal{A} \subset \mathcal{X}$), then similarly to (1.18)

$$D(P\|\mu) = -H(P). \quad (1.19)$$

1.4 The Method of Types

For finite alphabets, a particularly convenient tool in information theory is the *method of types* [37, 39, 74]. For a sequence $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ in which $|\mathcal{X}|$ is finite, its *type* or *empirical distribution* is the probability mass function

$$P_{\mathbf{x}}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{x_i = x\}, \quad \forall x \in \mathcal{X}. \quad (1.20)$$

Throughout, we use the notation $\mathbb{1}\{\text{clause}\}$ to mean the *indicator function*, i.e., this function equals 1 if “clause” is true and 0 otherwise. The set of types formed from n -length sequences in $\mathcal{X}$ is denoted as $\mathcal{P}_n(\mathcal{X})$. This is clearly a subset of $\mathcal{P}(\mathcal{X})$. The *type class* of P , denoted as $\mathcal{T}_P$, is the set of all sequences of length n for which their type is P , i.e.,

$$\mathcal{T}_P := \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\}. \quad (1.21)$$

It is customary to indicate the dependence of $\mathcal{T}_P$ on the *blocklength* n but we suppress this dependence for the sake of conciseness throughout. For a sequence $\mathbf{x} \in \mathcal{T}_P$, the set of all sequences $\mathbf{y} \in \mathcal{Y}^n$ such that $(\mathbf{x}, \mathbf{y})$ has joint type $P \times V$ is the *V-shell*, denoted as $\mathcal{T}_V(\mathbf{x})$. In other words,

$$\mathcal{T}_V(\mathbf{x}) := \{\mathbf{y} \in \mathcal{Y}^n : P_{\mathbf{x}, \mathbf{y}} = P \times V\}. \quad (1.22)$$

The conditional distribution V is also known as the *conditional type* of $\mathbf{y}$ given $\mathbf{x}$. Let $\mathcal{V}_n(\mathcal{Y}; P)$ be the set of all $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ for which the V -shell of a sequence of type P is non-empty.

We will often times find it useful to consider information-theoretic quantities of empirical distributions. All such quantities are denoted using hats. So for example, the *empirical entropy* of a sequence $\mathbf{x} \in \mathcal{X}^n$ is denoted as

$$\hat{H}(\mathbf{x}) := H(P_{\mathbf{x}}). \quad (1.23)$$

The *empirical conditional entropy* of $\mathbf{y} \in \mathcal{Y}^n$ given $\mathbf{x} \in \mathcal{X}^n$ where $\mathbf{y} \in \mathcal{T}_V(\mathbf{x})$ is denoted as

$$\hat{H}(\mathbf{y}|\mathbf{x}) := H(V|P_{\mathbf{x}}). \quad (1.24)$$

The *empirical mutual information* of a pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ with joint type $P_{\mathbf{x}, \mathbf{y}} = P_{\mathbf{x}} \times V$ is denoted as

$$\hat{I}(\mathbf{x} \wedge \mathbf{y}) := I(P_{\mathbf{x}}, V). \quad (1.25)$$

The following lemmas form the basis of the method of types. The proofs can be found in [37, 39].

Lemma 1.1 (Type Counting). The sets $\mathcal{P}_n(\mathcal{X})$ and $\mathcal{V}_n(\mathcal{Y}; P)$ for $P \in \mathcal{P}_n(\mathcal{X})$ satisfy

$$|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}, \quad \text{and} \quad |\mathcal{V}_n(\mathcal{Y}; P)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}. \quad (1.26)$$

In fact, it is easy to check that $|\mathcal{P}_n(\mathcal{X})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$ but (1.26) or its slightly stronger version

$$|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|-1} \quad (1.27)$$

usually suffices for our purposes in this monograph. This key property says that the number of types is polynomial in the blocklength n .

Lemma 1.2 (Size of Type Class). For a type $P \in \mathcal{P}_n(\mathcal{X})$, the type class $\mathcal{T}_P \subset \mathcal{X}^n$ satisfies

$$|\mathcal{P}_n(\mathcal{X})|^{-1} \exp(nH(P)) \leq |\mathcal{T}_P| \leq \exp(nH(P)). \quad (1.28)$$

For a conditional type $V \in \mathcal{V}_n(\mathcal{Y}; P)$ and a sequence $\mathbf{x} \in \mathcal{T}_P$, the V -shell $\mathcal{T}_V(\mathbf{x}) \subset \mathcal{Y}^n$ satisfies

$$|\mathcal{V}_n(\mathcal{Y}; P)|^{-1} \exp(nH(V|P)) \leq |\mathcal{T}_V(\mathbf{x})| \leq \exp(nH(V|P)). \quad (1.29)$$

This lemma says that, on the exponential scale,

$$|\mathcal{T}_P| \cong \exp(nH(P)), \quad \text{and} \quad |\mathcal{T}_V(\mathbf{x})| \cong \exp(nH(V|P)), \quad (1.30)$$

where we used the notation $a_n \cong b_n$ to mean equality up to a polynomial, i.e., there exists polynomials p_n and q_n such that $a_n/p_n \leq b_n \leq q_n a_n$. We now consider probabilities of sequences. Throughout, for a distribution $Q \in \mathcal{P}(\mathcal{X})$, we let $Q^n(\mathbf{x})$ be the product distribution, i.e.,

$$Q^n(\mathbf{x}) = \prod_{i=1}^n Q(x_i), \quad \forall \mathbf{x} \in \mathcal{X}^n. \quad (1.31)$$

Lemma 1.3 (Probability of Sequences). If $\mathbf{x} \in \mathcal{T}_P$ and $\mathbf{y} \in \mathcal{T}_V(\mathbf{x})$,

$$Q^n(\mathbf{x}) = \exp(-nD(P||Q) - nH(P)) \quad \text{and} \quad (1.32)$$

$$W^n(\mathbf{y}|\mathbf{x}) = \exp(-nD(V||W|P) - nH(V|P)). \quad (1.33)$$

This, together with Lemma 1.2, leads immediately to the final lemma in this section.

Lemma 1.4 (Probability of Type Classes). For a type $P \in \mathcal{P}_n(\mathcal{X})$,

$$|\mathcal{P}_n(\mathcal{X})|^{-1} \exp(-nD(P\|Q)) \leq Q^n(\mathcal{T}_P) \leq \exp(-nD(P\|Q)). \quad (1.34)$$

For a conditional type $V \in \mathcal{V}_n(\mathcal{Y}; P)$ and a sequence $\mathbf{x} \in \mathcal{T}_P$, we have

$$\begin{aligned} |\mathcal{V}_n(\mathcal{Y}; P)|^{-1} \exp(-nD(V\|W|P)) &\leq W^n(\mathcal{T}_V(\mathbf{x})|\mathbf{x}) \\ &\leq \exp(-nD(V\|W|P)). \end{aligned} \quad (1.35)$$

The interpretation of this lemma is that the probability that a random i.i.d. (independently and identically distributed) sequence X^n generated from Q^n belongs to the type class $\mathcal{T}_P$ is exponentially small with exponent $D(P\|Q)$, i.e.,

$$Q^n(\mathcal{T}_P) \cong \exp(-nD(P\|Q)). \quad (1.36)$$

The bounds in (1.35) can be interpreted similarly.

1.5 Probability Bounds

In this section, we summarize some bounds on probabilities that we use extensively in the sequel. For a random variable X , we let $\mathbb{E}[X]$ and $\text{Var}(X)$ be its expectation and variance respectively. To emphasize that the expectation is taken with respect to a random variable X with distribution P , we sometimes make this explicit by using a subscript, i.e., $\mathbb{E}_X$ or $\mathbb{E}_P$.

1.5.1 Basic Bounds

We start with the familiar Markov and Chebyshev inequalities.

Proposition 1.1 (Markov's inequality). Let X be a real-valued non-negative random variable. Then for any $a > 0$, we have

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \quad (1.37)$$

If we let X above be the non-negative random variable $(X - \mathbb{E}[X])^2$, we obtain Chebyshev's inequality.

Proposition 1.2 (Chebyshev's inequality). Let X be a real-valued random variable with mean μ and variance σ^2 . Then for any $b > 0$, we have

$$\Pr(|X - \mu| \geq b\sigma) \leq \frac{1}{b^2}. \quad (1.38)$$

We now consider a collection of real-valued random variables that are i.i.d. In particular, let $X^n = (X_1, \dots, X_n)$ be a collection of independent random variables where each X_i has distribution P with zero mean and finite variance σ^2 .

Proposition 1.3 (Weak Law of Large Numbers). For every $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i\right| > \epsilon\right) = 0. \quad (1.39)$$

Consequently, the average $\frac{1}{n} \sum_{i=1}^n X_i$ converges to 0 in probability.

This follows by applying Chebyshev's inequality to the random variable $\frac{1}{n} \sum_{i=1}^n X_i$. In fact, under mild conditions, the convergence to zero in (1.39) occurs exponentially fast. See, for example, Cramer's theorem in [43, Thm. 2.2.3].

1.5.2 Central Limit-Type Bounds

In preparation for the next result, we denote the *probability density function* (pdf) of a univariate Gaussian as

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}. \quad (1.40)$$

We will also denote this as $\mathcal{N}(\mu, \sigma^2)$ if the argument x is unnecessary. A *standard Gaussian distribution* is one in which the mean $\mu = 0$ and the standard deviation $\sigma = 1$. In the multivariate case, the pdf is

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad (1.41)$$

where $\mathbf{x} \in \mathbb{R}^k$. A *standard multivariate Gaussian distribution* is one in which the mean is $\mathbf{0}_k$ and the covariance is the identity matrix $\mathbf{I}_{k \times k}$.

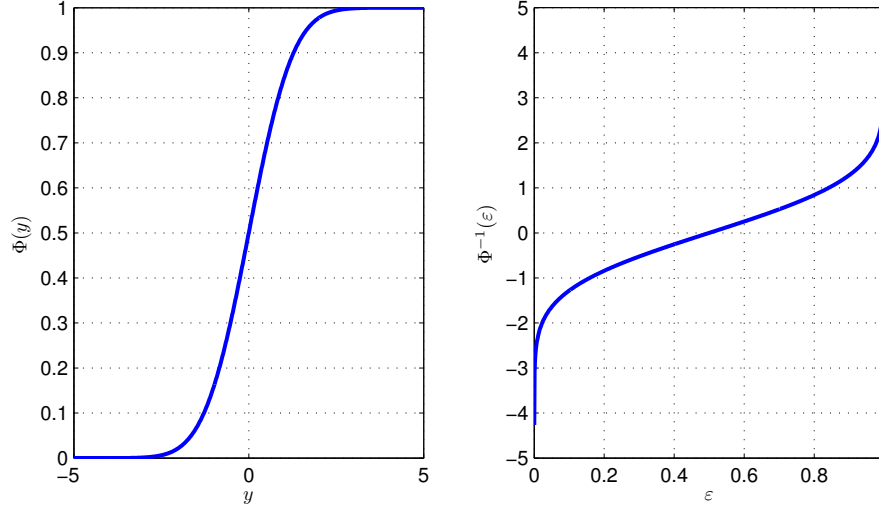


Figure 1.2: Plots of $\Phi(y)$ and $\Phi^{-1}(\varepsilon)$

For the univariate case, the *cumulative distribution function* (cdf) of the standard Gaussian is denoted as

$$\Phi(y) := \int_{-\infty}^y \mathcal{N}(x; 0, 1) dx. \quad (1.42)$$

We also find it convenient to introduce the inverse of Φ as

$$\Phi^{-1}(\varepsilon) := \sup \{y \in \mathbb{R} : \Phi(y) \leq \varepsilon\} \quad (1.43)$$

which evaluates to the usual inverse for $\varepsilon \in (0, 1)$ and extends continuously to take values $\pm\infty$ for ε outside $(0, 1)$. These monotonically increasing functions are shown in Fig. 1.2.

If the scaling in front of the sum in the statement of the law of large numbers in (1.39) is $\frac{1}{\sqrt{n}}$ instead of $\frac{1}{n}$, the resultant random variable $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ converges in distribution to a Gaussian random variable. As in Proposition 1.3, let X^n be a collection of i.i.d. random variables where each X_i has zero mean and finite variance σ^2 .

Proposition 1.4 (Central Limit Theorem). For any $a \in \mathbb{R}$, we have

$$\lim_{n \rightarrow \infty} \Pr \left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i < a \right) = \Phi(a). \quad (1.44)$$

In other words,

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} Z \quad (1.45)$$

where $\xrightarrow{d}$ means convergence in distribution and Z is the standard Gaussian random variable.

Throughout the monograph, in the evaluation of the non-asymptotic bounds, we will use a more quantitative version of the central limit theorem known as the Berry-Esseen theorem [17, 52]. See Feller [54, Sec. XVI.5] for a proof.

Theorem 1.5 (Berry-Esseen Theorem (i.i.d. Version)). Assume that the third absolute moment is finite, i.e., $T := \mathbb{E}[|X_1|^3] < \infty$. For every $n \in \mathbb{N}$, we have

$$\sup_{a \in \mathbb{R}} \left| \Pr \left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i < a \right) - \Phi(a) \right| \leq \frac{T}{\sigma^3\sqrt{n}}. \quad (1.46)$$

Remarkably, the Berry-Esseen theorem says that the convergence in the central limit theorem in (1.44) is uniform in $a \in \mathbb{R}$. Furthermore, the convergence of the distribution function of $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ to the Gaussian cdf occurs at a rate of $O(\frac{1}{\sqrt{n}})$. The constant of proportionality in the $O(\cdot)$ -notation depends *only* on the variance and the third absolute moment and not on any other statistics of the random variables.

There are many generalizations of the Berry-Esseen theorem. One which we will need is the relaxation of the assumption that the random variables are identically distributed. Let $X^n = (X_1, \dots, X_n)$ be a collection of independent random variables where each random variable has zero mean, variance $\sigma_i^2 := \mathbb{E}[X_i^2] > 0$ and third absolute moment $T_i := \mathbb{E}[|X_i|^3] < \infty$. We respectively define the average variance and average third absolute moment as

$$\sigma^2 := \frac{1}{n} \sum_{i=1}^n \sigma_i^2, \quad \text{and} \quad T := \frac{1}{n} \sum_{i=1}^n T_i. \quad (1.47)$$

Theorem 1.6 (Berry-Esseen Theorem (General Version)). For every $n \in \mathbb{N}$, we have

$$\sup_{a \in \mathbb{R}} \left| \Pr \left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i < a \right) - \Phi(a) \right| \leq \frac{6T}{\sigma^3\sqrt{n}}. \quad (1.48)$$

Observe that as with the i.i.d. version of the Berry-Esseen theorem, the remainder term scales as $O(\frac{1}{\sqrt{n}})$.

The proof of the following theorem uses the Berry-Esseen theorem (among other techniques). This theorem is proved in Polyanskiy-Poor-Verdú [123, Lem. 47]. Together with its variants, this theorem is useful for obtaining third-order asymptotics for binary hypothesis testing and other coding problems with non-vanishing error probabilities.

Theorem 1.7. Assume the same setup as in Theorem 1.6. For any $\gamma \geq 0$, we have

$$\mathbb{E} \left[\exp \left(- \sum_{i=1}^n X_i \right) \mathbb{1} \left\{ \sum_{i=1}^n X_i > \gamma \right\} \right] \leq 2 \left(\frac{\log 2}{\sqrt{2\pi}} + \frac{12T}{\sigma^2} \right) \frac{\exp(-\gamma)}{\sigma\sqrt{n}}. \quad (1.49)$$

It is trivial to see that the expectation in (1.49) is upper bounded by $\exp(-\gamma)$. The additional factor of $(\sigma\sqrt{n})^{-1}$ is crucial in proving coding theorems with better third-order terms. Readers familiar with strong large deviation theorems or exact asymptotics (see, e.g., [23, Thms. 3.3 and 3.5] or [43, Thm. 3.7.4]) will notice that (1.49) is in the same spirit as the theorem by Bahadur and Ranga-Rao [13]. There are two advantages of (1.49) compared to strong large deviation theorems. First, the bound is purely in terms of σ^2 and T , and second, one does not have to differentiate between lattice and non-lattice random variables. The disadvantage of (1.49) is that the constant is worse but this will not concern us as we focus on asymptotic results in this monograph, hence constants do not affect the main results.

For multi-terminal problems that we encounter in the latter parts of this monograph, we will require vector (or multidimensional) versions of the Berry-Esseen theorem. The following is due to Götze [63].

Theorem 1.8 (Vector Berry-Esseen Theorem I). Let $X_1^k, \dots, X_n^k$ be independent $\mathbb{R}^k$ -valued random vectors with zero mean. Let

$$S_n^k = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^k. \quad (1.50)$$

Assume that S_n^k has the following statistics

$$\text{Cov}(S_n^k) = \mathbb{E}[S_n^k (S_n^k)'] = \mathbf{I}_{k \times k}, \quad \text{and} \quad \xi := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|X_i^k\|_2^3]. \quad (1.51)$$

Let Z^k be a standard Gaussian random vector, i.e., its distribution is $\mathcal{N}(0^k, \mathbf{I}_{k \times k})$. Then, for all $n \in \mathbb{N}$, we have

$$\sup_{\mathcal{C} \in \mathfrak{C}_k} \left| \Pr(S_n^k \in \mathcal{C}) - \Pr(Z^k \in \mathcal{C}) \right| \leq \frac{c_k \xi}{\sqrt{n}}, \quad (1.52)$$

where $\mathfrak{C}_k$ is the family of all convex subsets of $\mathbb{R}^k$, and where c_k is a constant that depends only on the dimension k .

Theorem 1.8 can be applied for random vectors that are independent but not necessarily identically distributed. The constant c_k can be upper bounded by $400 k^{1/4}$ if the random vectors are i.i.d., a result by Bentkus [15]. However, its precise value will not be of concern to us in this monograph. Observe that the scalar versions of the Berry-Esseen theorems (in Theorems 1.5 and 1.6) are special cases (apart from the constant) of the vector version in which the family of convex subsets is restricted to the family of semi-infinite intervals $(-\infty, a)$.

We will frequently encounter random vectors with non-identity covariance matrices. The following modification of Theorem 1.8 is due to Watanabe-Kuzuoka-Tan [177, Cor. 29].

Corollary 1.9 (Vector Berry-Esseen Theorem II). Assume the same setup as in Theorem 1.8, except that $\text{Cov}(S_n^k) = \mathbf{V}$, a positive definite matrix. Then, for all $n \in \mathbb{N}$, we have

$$\sup_{\mathcal{C} \in \mathfrak{C}_k} \left| \Pr(S_n^k \in \mathcal{C}) - \Pr(Z^k \in \mathcal{C}) \right| \leq \frac{c_k \xi}{\lambda_{\min}(\mathbf{V})^{3/2} \sqrt{n}}, \quad (1.53)$$

where $\lambda_{\min}(\mathbf{V}) > 0$ is the smallest eigenvalue of $\mathbf{V}$.

The final probability bound is a quantitative version of the so-called *multivariate delta method* [174, Thm. 5.15]. Numerous similar statements of varying generalities have appeared in the statistics literature (e.g., [24, 175]). The simple version we present was shown by Molavian-Jazi and Laneman [112] who extended ideas in Hoeffding and Robbins'

paper [81, Thm. 4] to provide rates of convergence to Gaussianity under appropriate technical conditions. This result essentially says that a differentiable function of a normalized sum of independent random vectors also satisfies a Berry-Esseen-type result.

Theorem 1.10 (Berry-Esseen Theorem for Functions of i.i.d. Random Vectors). Assume that $X_1^k, \dots, X_n^k$ are $\mathbb{R}^k$ -valued, zero-mean, i.i.d. random vectors with positive definite covariance $\text{Cov}(X_1^k)$ and finite third absolute moment $\xi := \mathbb{E}[\|X_1^k\|_2^3]$. Let $\mathbf{f}(\mathbf{x})$ be a vector-valued function from $\mathbb{R}^k$ to $\mathbb{R}^l$ that is also twice continuously differentiable in a neighborhood of $\mathbf{x} = \mathbf{0}$. Let $\mathbf{J} \in \mathbb{R}^{l \times k}$ be the Jacobian matrix of $\mathbf{f}(\mathbf{x})$ evaluated at $\mathbf{x} = \mathbf{0}$, i.e., its elements are

$$J_{ij} = \left. \frac{\partial f_i(\mathbf{x})}{\partial x_j} \right|_{\mathbf{x}=\mathbf{0}}, \quad (1.54)$$

where $i = 1, \dots, l$ and $j = 1, \dots, k$. Then, for every $n \in \mathbb{N}$, we have

$$\sup_{\mathcal{C} \in \mathcal{C}_l} \left| \Pr \left(\mathbf{f} \left(\frac{1}{n} \sum_{i=1}^n X_i^k \right) \in \mathcal{C} \right) - \Pr (Z^l \in \mathcal{C}) \right| \leq \frac{c}{\sqrt{n}} \quad (1.55)$$

where $c > 0$ is a finite constant, and Z^l is a Gaussian random vector in $\mathbb{R}^l$ with mean vector and covariance matrix respectively given as

$$\mathbb{E}[Z^l] = \mathbf{f}(\mathbf{0}), \quad \text{and} \quad \text{Cov}(Z^l) = \frac{\mathbf{J} \text{Cov}(X_1^k) \mathbf{J}'}{n}. \quad (1.56)$$

In particular, the inequality in (1.55) implies that

$$\sqrt{n} \left(\mathbf{f} \left(\frac{1}{n} \sum_{i=1}^n X_i^k \right) - \mathbf{f}(\mathbf{0}) \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{J} \text{Cov}(X_1^k) \mathbf{J}'), \quad (1.57)$$

which is a canonical statement in the study of the multivariate delta method [174, Thm. 5.15].

Finally, we remark that Ingber-Wang-Kochman [87] used a result similar to that of Theorem 1.10 to derive second-order asymptotic results for various Shannon-theoretic problems. However, they analyzed the behavior of functions of *distributions* instead of functions of *random vectors* as in Theorem 1.10.

2

Binary Hypothesis Testing

In this chapter, we review asymptotic expansions in simple (non-composite) binary hypothesis testing when one of the two error probabilities is non-vanishing. We find this useful, as many coding theorems we encounter in subsequent chapters can be stated in terms of quantities related to binary hypothesis testing. For example, as pointed out in Csiszár and Körner [39, Ch. 1], fixed-to-fixed length lossless source coding and binary hypothesis testing are intimately connected through the relation between relative entropy and entropy in (1.18). Another example is in point-to-point channel coding, where a powerful non-asymptotic converse theorem [152, Eq. (4.29)] [123, Sec. III-E] [164, Prop. 6] can be stated in terms of the so-called ε -*hypothesis testing divergence* and the ε -*information spectrum divergence* (cf. Proposition 4.4). The properties of these two quantities, as well as the relation between them are discussed. Using various probabilistic limit theorems, we also evaluate these quantities in the asymptotic setting for product distributions. A corollary of the results presented is the familiar Chernoff-Stein lemma [39, Thm. 1.2], which asserts that the exponent with growing number of observations of the type-II error for a non-vanishing type-I error in a binary hypothesis test of P against Q

is the relative entropy $D(P\|Q)$.

The material in this chapter is based largely on the seminal work by Strassen [152, Thm. 3.1]. The exposition is based on the more recent works by Polyanskiy-Poor-Verdú [123, App. C], Tomamichel-Tan [164, Sec. III] and Tomamichel-Hayashi [163, Lem. 12].

2.1 Non-Asymptotic Quantities and Their Properties

Consider the simple (non-composite) binary hypothesis test:

$$\begin{aligned} H_0 : Z &\sim P \\ H_1 : Z &\sim Q \end{aligned} \quad (2.1)$$

where P and Q are two probability distributions on the same space $\mathcal{Z}$. We assume that the space $\mathcal{Z}$ is finite to keep the subsequent exposition simple. The notation in (2.1) means that under the null hypothesis H_0 , the random variable Z is distributed as $P \in \mathcal{P}(\mathcal{Z})$ while under the alternative hypothesis H_1 , it is distributed according to a different distribution $Q \in \mathcal{P}(\mathcal{Z})$. We would like to study the optimal performance of a hypothesis test in terms of the distributions P and Q .

There are several ways to measure the performance of a hypothesis test which, in precise terms, is a mapping δ from the observation space $\mathcal{Z}$ to $[0, 1]$. If the observation z is such that $\delta(z) \approx 0$, this means the test favors the null hypothesis H_0 . Conversely, $\delta(z) \approx 1$ means that the test favors the alternative hypothesis H_1 (or alternatively, rejects the null hypothesis H_0). If $\delta(z) \in \{0, 1\}$, the test is called *deterministic*, otherwise it is called *randomized*. Traditionally, there are three quantities that are of interest for a given test δ . The first is the *probability of false alarm*

$$P_{\text{FA}} := \sum_{z \in \mathcal{Z}} \delta(z) P(z) = \mathbb{E}_P[\delta(Z)]. \quad (2.2)$$

The second is the *probability of missed detection*

$$P_{\text{MD}} := \sum_{z \in \mathcal{Z}} (1 - \delta(z)) Q(z) = \mathbb{E}_Q[1 - \delta(Z)]. \quad (2.3)$$

The third is the *probability of detection*, which is one minus the prob-

ability of missed detection, i.e.,

$$P_D := \sum_{z \in \mathcal{Z}} \delta(z) Q(z) = \mathbb{E}_Q[\delta(Z)]. \quad (2.4)$$

The probability of false alarm and miss detection are traditionally called the *type-I* and *type-II errors* respectively in the statistics literature. The probability of detection and the probability of false alarm are also called the *power* and the *significance level* respectively. The “holy grail” is, of course, to design a test such that $P_{FA} = 0$ while $P_D = 1$ but this is clearly impossible unless P and Q are mutually singular measures.

Since misses are usually more costly than false alarms, let us fix a number $\varepsilon \in (0, 1)$ that represents a tolerable probability of false alarm (type-II error). Then define the smallest type-II error in the binary hypothesis test (2.1) with type-I error not exceeding ε , i.e.,

$$\beta_{1-\varepsilon}(P, Q) := \inf_{\delta: \mathcal{Z} \rightarrow [0,1]} \left\{ \mathbb{E}_Q[1 - \delta(Z)] : \mathbb{E}_P[\delta(Z)] \leq \varepsilon \right\}. \quad (2.5)$$

Observe that $\mathbb{E}_P[\delta(Z)] \leq \varepsilon$ constrains the probability of false alarm to be no greater than ε . Thus, we are searching over all tests δ satisfying $\mathbb{E}_P[\delta(Z)] \leq \varepsilon$ such that the probability of missed detection is minimized. Intuitively, $\beta_{1-\varepsilon}(P, Q)$ quantifies, in a non-asymptotic fashion, the performance of an optimal hypothesis test between P and Q .

A related quantity is the ε -hypothesis testing divergence

$$D_h^\varepsilon(P\|Q) := -\log \frac{\beta_{1-\varepsilon}(P, Q)}{1 - \varepsilon}. \quad (2.6)$$

This is a measure of the *distinguishability* of P from Q . As can be seen from (2.6), $\beta_{1-\varepsilon}(P, Q)$ and $D_h^\varepsilon(P\|Q)$ are simple functions of each other. We prefer to express the results in this monograph mostly in terms of $D_h^\varepsilon(P\|Q)$ because it shares similar properties with the usual relative entropy $D(P\|Q)$, as is evidenced from the following lemma.

Lemma 2.1 (Properties of D_h^ε). The ε -hypothesis testing divergence satisfies the *positive definiteness* condition [48, Prop. 3.2], i.e.,

$$D_h^\varepsilon(P\|Q) \geq 0. \quad (2.7)$$

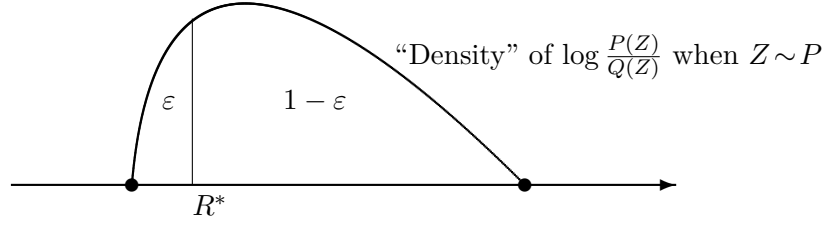


Figure 2.1: Illustration of the ε -information spectrum divergence $D_s^\varepsilon(P\|Q)$ which is the largest point R^* for which the probability mass to the left is no larger than ε .

Equality holds if and only if $P = Q$. In addition, it also satisfies the *data processing inequality* [173, Lem. 1], i.e., for any channel W ,

$$D_h^\varepsilon(PW\|QW) \leq D_h^\varepsilon(P\|Q). \quad (2.8)$$

While the ε -hypothesis testing divergence occurs naturally and frequently in coding problems, it is usually hard to analyze directly. Thus, we now introduce an equally important quantity. Define the ε -*information spectrum divergence* $D_s^\varepsilon(P\|Q)$ as

$$D_s^\varepsilon(P\|Q) := \sup \left\{ R \in \mathbb{R} : P \left(\left\{ z \in \mathcal{Z} : \log \frac{P(z)}{Q(z)} \leq R \right\} \right) \leq \varepsilon \right\}. \quad (2.9)$$

Just as in information spectrum analysis [67], this quantity places the *distribution* of the log-likelihood ratio $\log \frac{P(Z)}{Q(Z)}$ (where $Z \sim P$), and not just its expectation, in the most prominent role. See Fig. 2.1 for an interpretation of the definition in (2.9).

As we will see, the ε -information spectrum divergence is intimately related to the ε -hypothesis testing divergence (cf. Lemma 2.4). The former is, however, easier to compute. Note that if P and Q are product measures, then by virtue of the fact that $\log \frac{P(Z)}{Q(Z)}$ is a sum of independent random variables, one can estimate the probability in (2.9) using various probability tail bounds. This we do in the following section.

We now state two useful properties of $D_s^\varepsilon(P\|Q)$. The proofs of these lemmas are straightforward and can be found in [164, Sec. III.A].

Lemma 2.2 (Sifting from a convex combination). Let $P \in \mathcal{P}(\mathcal{Z})$ and $Q = \sum_k \alpha_k Q_k$ be an at most countable convex combination of distributions $Q_k \in \mathcal{P}(\mathcal{Z})$ with non-negative weights α_k summing to one, i.e.,

$\sum_k \alpha_k = 1$. Then,

$$D_s^\varepsilon(P\|Q) \leq \inf_k \left\{ D_s^\varepsilon(P\|Q_k) + \log \frac{1}{\alpha_k} \right\}. \quad (2.10)$$

In particular, Lemma 2.2 tells us that if there exists some $\gamma > 0$ such that $\tilde{Q}(z) \leq \gamma Q(z)$ for all $z \in \mathcal{Z}$ then,

$$D_s^\varepsilon(P\|\tilde{Q}) \geq D_s^\varepsilon(P\|Q) - \log \gamma. \quad (2.11)$$

Lemma 2.3 (“Symbol-wise” relaxation of D_s^ε). Let W and V be two channels from $\mathcal{X}$ to $\mathcal{Y}$ and let $P \in \mathcal{P}(\mathcal{X})$. Then,

$$D_s^\varepsilon(P \times W\|P \times V) \leq \sup_{x \in \mathcal{X}} D_s^\varepsilon(W(\cdot|x)\|V(\cdot|x)). \quad (2.12)$$

One can readily toggle between the ε -hypothesis testing divergence and the ε -information spectrum divergence because they satisfy the bounds in the following lemma. The proof of this lemma mimics that of [163, Lem. 12].

Lemma 2.4 (Relation between divergences). For every $\varepsilon \in (0, 1)$ and every $\eta \in (0, 1 - \varepsilon)$, we have

$$D_s^\varepsilon(P\|Q) - \log \frac{1}{1 - \varepsilon} \leq D_h^\varepsilon(P\|Q) \quad (2.13)$$

$$\leq D_s^{\varepsilon+\eta}(P\|Q) + \log \frac{1 - \varepsilon}{\eta}. \quad (2.14)$$

Proof. The following proof is based on that for [163, Lem. 12]. For the lower bound in (2.13), consider the likelihood ratio test

$$\delta(z) := \mathbb{1} \left\{ \log \frac{P(z)}{Q(z)} \leq \gamma \right\}, \quad \text{where } \gamma := D_s^\varepsilon(P\|Q) - \xi \quad (2.15)$$

for some small $\xi > 0$. This test clearly satisfies $\mathbf{E}_P[\delta(Z)] \leq \varepsilon$ by the definition of the ε -information spectrum divergence. On the other hand,

$$\begin{aligned} & \mathbf{E}_Q[1 - \delta(Z)] \\ &= \sum_{z \in \mathcal{Z}} Q(z) \mathbb{1} \left\{ \log \frac{P(z)}{Q(z)} > \gamma \right\} \end{aligned} \quad (2.16)$$

$$\leq \sum_{z \in \mathcal{Z}} P(z) \exp(-\gamma) \mathbb{1} \left\{ \log \frac{P(z)}{Q(z)} > \gamma \right\} \quad (2.17)$$

$$\leq \sum_{z \in \mathcal{Z}} P(z) \exp(-\gamma) \quad (2.18)$$

$$\leq \exp(-\gamma) \quad (2.19)$$

As a result, by the definition of $D_h^\varepsilon(P\|Q)$, we have

$$D_h^\varepsilon(P\|Q) \geq \gamma - \log \frac{1}{1-\varepsilon} = D_s^\varepsilon(P\|Q) - \xi - \log \frac{1}{1-\varepsilon}. \quad (2.20)$$

Finally, take $\xi \downarrow 0$ to complete the proof of (2.13).

For the upper bound in (2.14), we may assume $D_h^\varepsilon(P\|Q)$ is finite; otherwise there is nothing to prove as P is not absolutely continuous with respect to Q and so $D_s^{\varepsilon+\eta}(P\|Q)$ is infinite. According to the definition of $D_h^\varepsilon(P\|Q)$, for any $\gamma \geq 0$, there exists a test δ satisfying $\mathbb{E}_P[\delta(Z)] \leq \varepsilon$ such that

$$(1-\varepsilon) \exp(-D_h^\varepsilon(P\|Q)) \geq \mathbb{E}_Q[1-\delta(Z)] \quad (2.21)$$

$$\geq \sum_{z: P(z) \leq \gamma Q(z)} Q(z)(1-\delta(z)) \quad (2.22)$$

$$\geq \frac{1}{\gamma} \sum_{z: P(z) \leq \gamma Q(z)} P(z)(1-\delta(z)) \quad (2.23)$$

$$\geq \frac{1}{\gamma} \left[\sum_z P(z)(1-\delta(z)) - \sum_{z: P(z) > \gamma Q(z)} P(z) \right] \quad (2.24)$$

$$\geq \frac{1}{\gamma} \left[1-\varepsilon - P\left(\left\{z : \frac{P(z)}{Q(z)} > \gamma\right\}\right) \right] \quad (2.25)$$

where (2.25) follows because $\mathbb{E}_P[\delta(Z)] \leq \varepsilon$. Now fix a small $\xi > 0$ and choose

$$\gamma = \exp(D_s^{\varepsilon+\eta}(P\|Q) + \xi). \quad (2.26)$$

Consequently, from (2.25), we have

$$D_h^\varepsilon(P\|Q) \leq D_s^{\varepsilon+\eta}(P\|Q) + \xi - \log \left(1 - \frac{P(\{z : \log \frac{P(z)}{Q(z)} > D_s^{\varepsilon+\eta}(P\|Q) + \xi\})}{1-\varepsilon} \right) \quad (2.27)$$

By the definition of $D_s^{\varepsilon+\eta}(P\|Q)$, the probability within the logarithm is upper bounded by $1-\varepsilon-\eta$. Taking $\xi \downarrow 0$ completes the proof of (2.14) and hence, the lemma. $\square$

2.2 Asymptotic Expansions

In this section, we consider the asymptotic expansions of $D_h^\varepsilon(P^{(n)}\|Q^{(n)})$ and $D_s^\varepsilon(P^{(n)}\|Q^{(n)})$ when $P^{(n)}$ and $Q^{(n)}$ are product distributions, i.e.,

$$P^{(n)}(\mathbf{z}) := \prod_{i=1}^n P_i(z_i), \quad \text{and} \quad Q^{(n)}(\mathbf{z}) = \prod_{i=1}^n Q_i(z_i), \quad (2.28)$$

for all $\mathbf{z} = (z_1, \dots, z_n) \in \mathcal{Z}^n$. The *component distributions* $\{(P_i, Q_i)\}_{i=1}^n$ are not necessarily the same for each i . However, we do assume for the sake of simplicity that for each i , $P_i \ll Q_i$ so $D(P_i\|Q_i) < \infty$. Let $V(P\|Q)$ be the variance of the log-likelihood ratio between P and Q , i.e.,

$$V(P\|Q) := \sum_{z \in \mathcal{Z}} P(z) \left[\log \frac{P(z)}{Q(z)} - D(P\|Q) \right]^2. \quad (2.29)$$

This is also known as the *relative entropy variance*. Let the third absolute moment of the log-likelihood ratio between P and Q be

$$T(P\|Q) := \sum_{z \in \mathcal{Z}} P(z) \left| \log \frac{P(z)}{Q(z)} - D(P\|Q) \right|^3. \quad (2.30)$$

Also define the following quantities:

$$D_n := \frac{1}{n} \sum_{i=1}^n D(P_i\|Q_i), \quad (2.31)$$

$$V_n := \frac{1}{n} \sum_{i=1}^n V(P_i\|Q_i), \quad \text{and} \quad (2.32)$$

$$T_n := \frac{1}{n} \sum_{i=1}^n T(P_i\|Q_i). \quad (2.33)$$

The first result in this section is the following:

Proposition 2.1 (Berry-Esseen bounds for D_s^ε). Assume there exists a constant $V_- > 0$ such that $V_n \geq V_-$. We have

$$\Phi^{-1}\left(\varepsilon - \frac{6 T_n}{\sqrt{n V_-^3}}\right) \leq \frac{D_s^\varepsilon(P^{(n)}\|Q^{(n)}) - n D_n}{\sqrt{n V_n}} \leq \Phi^{-1}\left(\varepsilon + \frac{6 T_n}{\sqrt{n V_-^3}}\right). \quad (2.34)$$

Proof. Let Z^n be distributed according to $P^{(n)}$. By using the product structure of $P^{(n)}$ and $Q^{(n)}$ in (2.28),

$$\Pr \left(\log \frac{P^{(n)}(Z^n)}{Q^{(n)}(Z^n)} \leq R \right) = \Pr \left(\sum_{i=1}^n \log \frac{P_i(Z_i)}{Q_i(Z_i)} \leq R \right). \quad (2.35)$$

By the Berry-Esseen theorem in Theorem 1.6, we have

$$\left| \Pr \left(\sum_{i=1}^n \log \frac{P_i(Z_i)}{Q_i(Z_i)} \leq R \right) - \Phi \left(\frac{R - nD_n}{\sqrt{nV_n}} \right) \right| \leq \frac{6T_n}{\sqrt{nV_n^3}}. \quad (2.36)$$

The result immediately follows by using the bound $V_n \geq V_-$. $\square$

A special case of the bound above occurs when $P_i = P$ and $Q_i = Q$ for all $i = 1, \dots, n$. In this case, we write P^n for $P^{(n)}$ and similarly, Q^n for $Q^{(n)}$. One has:

Corollary 2.5 (Asymptotics of D_s^ε). If $V(P\|Q) > 0$, then

$$D_s^\varepsilon(P^n\|Q^n) = nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}(\varepsilon) + O(1). \quad (2.37)$$

Proof. Since $V(P\|Q) > 0$ and $T(P\|Q) < \infty$ (because $P \ll Q$), the term $6T_n/\sqrt{nV_n^3}$ in (2.34) is equal to $\frac{c}{\sqrt{n}}$ for some finite $c > 0$. By Taylor expansions,

$$\Phi^{-1} \left(\varepsilon \pm \frac{c}{\sqrt{n}} \right) = \Phi^{-1}(\varepsilon) + O \left(\frac{1}{\sqrt{n}} \right), \quad (2.38)$$

which completes the proof. $\square$

In some applications, it is not possible to guarantee that V_n is uniformly bounded away from zero (per Proposition 2.1). In this case, to obtain an upper bound on D_s^ε , we employ Chebyshev's inequality instead of the Berry-Esseen theorem. In the following proposition, which is usually good enough to establish *strong converses*, we do not assume that the component distributions are the same.

Proposition 2.2 (Chebyshev bound for D_s^ε). We have

$$D_s^\varepsilon(P^{(n)}\|Q^{(n)}) \leq nD_n + \sqrt{\frac{nV_n}{1-\varepsilon}}. \quad (2.39)$$

Proof. By the definition of the ε -information spectrum divergence, we have

$$D_s^\varepsilon(P^{(n)}\|Q^{(n)}) = \max\{D^-, D^+\} \quad (2.40)$$

where D^- and D^+ are defined as

$$D^- := \sup \left\{ R \leq nD_n : P\left(\left\{z \in \mathcal{Z} : \log \frac{P(z)}{Q(z)} \leq R\right\}\right) \leq \varepsilon \right\}, \quad (2.41)$$

$$D^+ := \sup \left\{ R > nD_n : P\left(\left\{z \in \mathcal{Z} : \log \frac{P(z)}{Q(z)} \leq R\right\}\right) \leq \varepsilon \right\}. \quad (2.42)$$

Clearly, $D^- \leq nD_n$ so it remains to upper bound D^+ . Let $R > nD_n$ be fixed. By Chebyshev's inequality,

$$\Pr\left(\sum_{i=1}^n \log \frac{P_i(Z_i)}{Q_i(Z_i)} \leq R\right) \geq 1 - \frac{nV_n}{(R - nD_n)^2}. \quad (2.43)$$

Hence, we have

$$D^+ \leq \sup \left\{ R > nD_n : 1 - \frac{nV_n}{(R - nD_n)^2} \leq \varepsilon \right\} \quad (2.44)$$

$$= nD_n + \sqrt{\frac{nV_n}{1 - \varepsilon}}. \quad (2.45)$$

Thus, we see that the bound on D^+ dominates. This yields (2.39) as desired. $\square$

Now we would like an expansion for D_h^ε similar to that for D_s^ε in Corollary 2.5. The following was shown by Strassen [152, Thm. 3.1].

Proposition 2.3 (Asymptotics of D_h^ε). Assume the conditions in Corollary 2.5. The following holds:

$$D_h^\varepsilon(P^n\|Q^n) = nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1). \quad (2.46)$$

As a result, in the asymptotic setting for identical product distributions, $D_h^\varepsilon(P^n\|Q^n)$ exceeds $D_s^\varepsilon(P^n\|Q^n)$ by $\frac{1}{2}\log n$ ignoring constant terms, i.e.,

$$D_h^\varepsilon(P^n\|Q^n) = D_s^\varepsilon(P^n\|Q^n) + \frac{1}{2}\log n + O(1). \quad (2.47)$$

Proof. Let us first verify the upper bound. Let η in the upper bound of Lemma 2.4 be chosen to be $\frac{1}{\sqrt{n}}$. Now, for n large enough (so $\frac{1}{\sqrt{n}} < 1 - \varepsilon$), combine this upper bound with Corollary 2.5 to obtain that

$$D_h^\varepsilon(P^n \| Q^n) \leq D_s^{\varepsilon + \frac{1}{\sqrt{n}}}(P^n \| Q^n) + \frac{1}{2} \log n + \log(1 - \varepsilon) \quad (2.48)$$

$$= nD(P \| Q) + \sqrt{nV(P \| Q)}\Phi^{-1}\left(\varepsilon + \frac{1}{\sqrt{n}}\right) + \frac{1}{2} \log n + O(1) \quad (2.49)$$

Applying a Taylor expansion to the last step and noting that $V(P \| Q) < \infty$ because $P \ll Q$ yields the upper bound in (2.46).

The proof of the lower bound in (2.46) is slightly more involved. Observe that if we naïvely employed (2.13) to lower bound $D_h^\varepsilon(P^n \| Q^n)$ with $D_s^\varepsilon(P^n \| Q^n) - \log \frac{1}{1 - \varepsilon}$, the third-order term would be $O(1)$ instead of the better $\frac{1}{2} \log n + O(1)$. The idea is to propose an appropriate test for D_h^ε and to use Theorem 1.7. Consider the likelihood ratio test

$$\delta(\mathbf{z}) := \mathbb{1}\left\{\log \frac{P^n(\mathbf{z})}{Q^n(\mathbf{z})} \leq \gamma\right\} \quad (2.50)$$

Define $\sigma^2 := V(P \| Q)$ and $T := T(P \| Q)$. Also define the i.i.d. random variables $U_i := \log P(Z_i) - \log Q(Z_i)$, $1 \leq i \leq n$, each having variance σ^2 and third absolute moment T . Consider, the expectation of $1 - \delta(Z^n)$ under the distribution Q^n :

$$\begin{aligned} \mathbb{E}_{Q^n}[1 - \delta(Z^n)] &= \sum_{\mathbf{z}} Q^n(\mathbf{z}) \mathbb{1}\left\{\log \frac{P^n(\mathbf{z})}{Q^n(\mathbf{z})} > \gamma\right\} \end{aligned} \quad (2.51)$$

$$= \sum_{\mathbf{z}} P^n(\mathbf{z}) \exp\left(-\log \frac{P^n(\mathbf{z})}{Q^n(\mathbf{z})}\right) \mathbb{1}\left\{\log \frac{P^n(\mathbf{z})}{Q^n(\mathbf{z})} > \gamma\right\} \quad (2.52)$$

$$= \mathbb{E}_{P^n}\left[\exp\left(-\sum_{i=1}^n U_i\right) \mathbb{1}\left\{\sum_{i=1}^n U_i > \gamma\right\}\right] \quad (2.53)$$

$$\leq 2\left(\frac{\log 2}{\sqrt{2\pi}} + \frac{12T}{\sigma^2}\right) \frac{\exp(-\gamma)}{\sigma\sqrt{n}} \quad (2.54)$$

where (2.54) is an application of Theorem 1.7. Now put

$$\gamma := nD(P \| Q) + \sqrt{nV(P \| Q)}\Phi^{-1}\left(\varepsilon - \frac{6T(P \| Q)}{\sqrt{nV(P \| Q)^3}}\right). \quad (2.55)$$

An application of the Berry-Esseen theorem yields

$$\mathbb{E}_P[\delta(Z^n)] = P\left(\left\{\mathbf{z} : \sum_{i=1}^n \log \frac{P(z_i)}{Q(z_i)} \leq \gamma\right\}\right) \leq \varepsilon. \quad (2.56)$$

From (2.54), (2.56) and the definition of D_h^ε , we have

$$D_h^\varepsilon(P^n \| Q^n) \geq \gamma + \log(\sigma\sqrt{n}) + O(1) = \gamma + \frac{1}{2} \log n + O(1). \quad (2.57)$$

The proof is concluded by plugging (2.55) into (2.57) and Taylor expanding $\Phi^{-1}(\cdot)$ around ε . $\square$

We remark that the lower bound in Proposition 2.3 can be achieved using *deterministic* tests, i.e., δ can be chosen to be an indicator function as in (2.50). Randomization is thus unnecessary. Also, one can relax the assumption that Q^n is a product probability measure; it can be an arbitrary product measure. These realizations are important to make the connection between hypothesis testing and lossless source coding which we discuss in the next chapter.

A corollary of Proposition 2.3 is the Chernoff-Stein lemma [25] quantifying the error exponent of the probability of missed detection keeping the probability of false alarm bounded above by ε .

Corollary 2.6 (Chernoff-Stein lemma). Assume the conditions in Corollary 2.5 and recall the definition of $\beta_{1-\varepsilon}$ in (2.5). For every $\varepsilon \in (0, 1)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\beta_{1-\varepsilon}(P^n, Q^n)} = \lim_{n \rightarrow \infty} \frac{D_h^\varepsilon(P^n \| Q^n)}{n} = D(P \| Q). \quad (2.58)$$

Part II

**Point-To-Point
Communication**

3

Source Coding

In this chapter, we revisit the fundamental problem of fixed-to-fixed length lossless and lossy source compression. Shannon, in his original paper [141] that launched the field of information theory, showed that the fundamental limit of compression of a discrete memoryless source (DMS) P is the entropy $H(P)$. For the case of continuous sources, lossless compression is not possible and some distortion must be allowed. Shannon showed in [144] that the corresponding fundamental limit of compression of memoryless source P up to distortion $\Delta \geq 0$, assuming a separable distortion measure d , is the *rate-distortion function*

$$R(P, \Delta) := \min_{W \in \mathcal{P}(\hat{\mathcal{X}}|\mathcal{X}) : \mathbb{E}_{P \times W}[d(X, \hat{X})] \leq \Delta} I(P, W). \quad (3.1)$$

These first-order fundamental limits are attained as the number of realizations of the source (i.e., the blocklength of the source) P tends to infinity. The strong converse for rate-distortion is also known and shown, for example, in [39, Ch. 7]. In the following, we present known non-asymptotic bounds for lossless and lossy source coding. We then fix the permissible error probability in the lossless case and the excess distortion probability in the lossy case at some non-vanishing $\varepsilon \in (0, 1)$. The non-asymptotic bounds are evaluated as n becomes large so as to

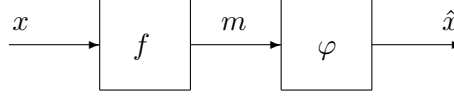


Figure 3.1: Illustration of the fixed-to-fixed length source coding problem.

obtain asymptotic expansions of the logarithm of the smallest achievable code size. These refined results provide an approximation of the extra code rate (beyond $H(P)$ or $R(P, \Delta)$) one must incur when operating in the finite blocklength regime. Finally, for both the lossless and lossy compression problems, we provide alternative proof techniques based on the method of types that are partially universal.

The material in this chapter concerning lossless source coding is based on the seminal work by Strassen [152, Thm. 1.1]. The material on lossy source coding is based on more recent work by Ingber-Kochman [86] and Kostina-Verdú [97].

3.1 Lossless Source Coding: Non-Asymptotic Bounds

We now set up the almost lossless source coding problem formally. As mentioned, we only consider *fixed-to-fixed length* source coding in this monograph. A *source* is simply a probability mass function P on some finite alphabet $\mathcal{X}$ or the associated random variable X with distribution P . See Fig. 3.1 for an illustration of the setup.

An (M, ε) -code for the source $P \in \mathcal{P}(\mathcal{X})$ consists of a pair of maps that includes an *encoder* $f : \mathcal{X} \rightarrow \{1, \dots, M\}$ and a *decoder* $\varphi : \{1, \dots, M\} \rightarrow \mathcal{X}$ such that the *error probability*

$$P(\{x \in \mathcal{X} : \varphi(f(x)) \neq x\}) \leq \varepsilon. \quad (3.2)$$

The number M is called the *size* of the code (f, φ) .

Given a source P , we define the *almost lossless source coding non-asymptotic fundamental limit* as

$$M^*(P, \varepsilon) := \min \{M \in \mathbb{N} : \exists \text{ an } (M, \varepsilon)\text{-code for } P\}. \quad (3.3)$$

Obviously for an arbitrary source, the exact evaluation of the minimum code size $M^*(P, \varepsilon)$ is challenging. In the following, we assume that it is

a discrete memoryless source (DMS), i.e., the distribution P^n consists of n copies of an underlying distribution P . With this assumption, we can find an asymptotic expansion of $\log M^*(P^n, \varepsilon)$.

The agenda for this and subsequent chapters will largely be standard. We first establish “good” bounds on non-asymptotic quantities like $M^*(P, \varepsilon)$. Subsequently, we replace the source or channel with n independent copies of it. Finally, we use an appropriate limit theorem (e.g., those in Section 1.5) to evaluate the non-asymptotic bounds in the large n limit.

3.1.1 An Achievability Bound

One of the take-home messages that we would like to convey in this section is that fixed-to-fixed length lossless source coding is nothing but binary hypothesis testing where the measures P and Q are chosen appropriately. In fact, a reasonable coding scheme for the lossless source coding would simply be to encode a “typical” set of source symbols $\mathcal{T} \subset \mathcal{X}$, ignore the rest, and declare an error if the realized symbol from the source is not in $\mathcal{T}$. In this way, one sees that

$$M^*(P, \varepsilon) \leq \min_{\mathcal{T} \subset \mathcal{X}: P(\mathcal{X} \setminus \mathcal{T}) \leq \varepsilon} |\mathcal{T}| \quad (3.4)$$

This bound can be stated in terms of $\beta_{1-\varepsilon}(P, Q)$ or, equivalently, the ε -hypothesis testing divergence $D_h^\varepsilon(P \| Q)$ if we restrict the tests that define these quantities to be deterministic, and also allow Q to be an arbitrary measure (not necessarily a probability measure). Let μ be the counting measure, i.e.,

$$\mu(\mathcal{A}) := |\mathcal{A}|, \quad \forall \mathcal{A} \subset \mathcal{X}. \quad (3.5)$$

Proposition 3.1 (Source coding as hypothesis testing: Achievability). Let $\varepsilon \in (0, 1)$ and P be any source with countable alphabet $\mathcal{X}$. We have

$$M^*(P, \varepsilon) \leq \beta_{1-\varepsilon}(P, \mu), \quad (3.6)$$

or in terms of the ε -hypothesis testing divergence (cf. (2.6)),

$$\log M^*(P, \varepsilon) \leq -D_h^\varepsilon(P \| \mu) - \log \frac{1}{1 - \varepsilon}. \quad (3.7)$$

3.1.2 A Converse Bound

The converse bound we evaluate is also intimately connected to a divergence we introduced in the previous chapter, namely the ε -information spectrum divergence where the distribution in the alternate hypothesis Q is chosen to be the counting measure.

Proposition 3.2 (Source coding as hypothesis testing: Converse). Let $\varepsilon \in (0, 1)$ and P be any source with countable alphabet $\mathcal{X}$. For any $\eta \in (0, 1 - \varepsilon)$, we have

$$\log M^*(P, \varepsilon) \geq -D_s^{\varepsilon+\eta}(P\|\mu) - \log \frac{1}{\eta}. \quad (3.8)$$

This statement is exactly Lemma 1.3.2 in Han's book [67]. Since the proof is short, we provide it for completeness.

Proof. By the definition of the ε -information spectrum divergence, it is enough to establish that every (M, ε) -code for P must satisfy

$$\varepsilon + \eta \geq P\left(\left\{x : P(x) \leq \frac{\eta}{M}\right\}\right), \quad (3.9)$$

for any $\eta \in (0, 1 - \varepsilon)$. Let $\mathcal{T} := \{x : P(x) \leq \frac{\eta}{M}\}$ and let $\mathcal{S} := \{x : \varphi(f(x)) = x\}$. Clearly, $|\mathcal{S}| \leq M$. We have

$$P(\mathcal{T}) \leq P(\mathcal{X} \setminus \mathcal{S}) + P(\mathcal{T} \cap \mathcal{S}) \leq \varepsilon + P(\mathcal{T} \cap \mathcal{S}). \quad (3.10)$$

Furthermore,

$$P(\mathcal{T} \cap \mathcal{S}) = \sum_{x \in \mathcal{T} \cap \mathcal{S}} P(x) \leq \sum_{x \in \mathcal{T} \cap \mathcal{S}} \frac{\eta}{M} \leq |\mathcal{S}| \frac{\eta}{M} \leq \eta. \quad (3.11)$$

Uniting (3.10) and (3.11) gives (3.9) as desired. $\square$

Observe the similarity of this proof to proof of the upper bound of D_h^ε in terms of D_s^ε in Lemma 2.4.

3.2 Lossless Source Coding: Asymptotic Expansions

Now we assume that the source P^n is stationary and memoryless, i.e., a DMS. More precisely,

$$P^n(\mathbf{x}) = \prod_{i=1}^n P(x_i), \quad \forall \mathbf{x} \in \mathcal{X}^n. \quad (3.12)$$

We assume throughout that $P(x) > 0$ for all $x \in \mathcal{X}$. Shannon [141] showed that the minimum rate to achieve almost lossless compression of a DMS P is the entropy $H(P)$. In this section as well as the next one, we derive finer evaluations of the fundamental compression limit by considering the asymptotic expansion of $\log M^*(P^n, \varepsilon)$. To do so, we need to define another important quantity related to the source P .

Let the *source dispersion* of P be the variance of the *self-information* random variable $-\log P(X)$, i.e.,

$$V(P) := \text{Var} \left[\log \frac{1}{P(X)} \right] = \sum_{x \in \mathcal{X}} P(x) \left[\log \frac{1}{P(x)} - H(P) \right]^2. \quad (3.13)$$

Note that the expectation of the self-information is the entropy $H(P)$. In Kontoyannis-Verdú [95], $V(P)$ is called the *varentropy*. If $V(P) = 0$ this means that the source is either deterministic or uniform.

The two non-asymptotic theorems in the preceding section combine to give the following asymptotic expansion of the minimum code size $M^*(P^n, \varepsilon)$.

Theorem 3.1. If the source $P \in \mathcal{P}(\mathcal{X})$ satisfies $V(P) > 0$, then

$$\log M^*(P^n, \varepsilon) = nH(P) - \sqrt{nV(P)}\Phi^{-1}(\varepsilon) - \frac{1}{2} \log n + O(1). \quad (3.14)$$

Otherwise, we have

$$\log M^*(P^n, \varepsilon) = nH(P) + O(1). \quad (3.15)$$

Proof. For the direct part of (3.14) (upper bound), note that the term $-\log \frac{1}{1-\varepsilon}$ in Proposition 3.1 is a constant, so we simply have to evaluate $D_h^\varepsilon(P^n \| \mu^n)$.¹ From Corollary 2.5 and its remark that the lower bound on $D_h^\varepsilon(P^n \| Q^n)$ can be achieved using deterministic tests, we have

$$D_h^\varepsilon(P^n \| \mu^n) = nD(P \| \mu) + \sqrt{nV(P \| \mu)}\Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1). \quad (3.16)$$

It can easily be verified (cf. (1.19)) that

$$D(P \| \mu) = -H(P), \quad \text{and} \quad V(P \| \mu) = V(P). \quad (3.17)$$

¹Just to be pedantic, for any $\mathcal{A} \subset \mathcal{X}^n$, the measure $\mu^n(\mathcal{A})$ is defined as $\sum_{\mathbf{x} \in \mathcal{A}} \mu^n(\mathbf{x}) = |\mathcal{A}|$ and $\mu^n(\mathbf{x}) = 1$ for each $\mathbf{x} \in \mathcal{X}^n$. Hence, μ^n has the required product structure for the application of Corollary 2.5, for which the second argument of $D_h^\varepsilon(P^n \| Q^n)$ is not restricted to product probability measures.

This concludes the proof of the direct part in light of Proposition 3.1.

For the converse part of (3.14) (lower bound), choose $\eta = \frac{1}{\sqrt{n}}$ so the term $-\log \frac{1}{\eta}$ in Proposition 3.2 gives us the $-\frac{1}{2} \log n$ term. Furthermore, by Proposition 2.1 and the simplifications in (3.17),

$$D_s^{\varepsilon + \frac{1}{\sqrt{n}}}(P^n \| \mu^n) = -nH(P) + \sqrt{nV(P)}\Phi^{-1}\left(\varepsilon + \frac{1}{\sqrt{n}}\right) + O(1). \quad (3.18)$$

A Taylor expansion of $\Phi^{-1}(\cdot)$ completes the proof of (3.14).

For (3.15), note that the self-information $-\log P(X)$ takes on the value $H(P)$ with probability one. In other words,

$$P^n(\log P^n(X^n) \leq R) = \mathbb{1}\{R \geq -nH(P)\}. \quad (3.19)$$

The bounds on $\log M^*(P^n, \varepsilon)$ in Propositions 3.1 and 3.2 and the relaxation to the ε -information spectrum divergence (Lemma 2.4) yields (3.15). $\square$

The expansion in (3.14) in Theorem 3.1 appeared in early works by Yushkevich [190] (with $o(\sqrt{n})$ in place of $-\frac{1}{2} \log n$ but for Markov chains) and Strassen [152, Thm. 1.1] (in the form stated). It has since appeared in various other forms and levels of generality in Kontoyannis [93], Hayashi [75], Kostina-Verdú [97], Nomura-Han [117] and Kontoyannis-Verdú [95] among others.

As can be seen from the non-asymptotic bounds and the asymptotic evaluation, fixed-to-fixed length lossless source coding and binary hypothesis testing are virtually the same problem. Asymptotic expansions for D_h^ε and D_s^ε can be used directly to estimate the minimum code size $M^*(P^n, \varepsilon)$ for an ε -reliable lossless source code.

3.3 Second-Order Asymptotics of Lossless Source Coding via the Method of Types

Clearly, the coding scheme described in (3.4) is *non-universal*, i.e., the code depends on knowledge of the source distribution. In many applications, the exact source distribution is unknown, and hence has to be estimated *a priori*, or one has to design a source code that works well for any source distribution. It is a well-known application of the

method of types that universal source codes achieve the lossless source coding error exponent [39, Thm. 2.15]. One then wonders whether there is any degradation in the asymptotic expansion of $\log M^*(P^n, \varepsilon)$ if the encoder and decoder know less about the source. It turns out that the source dispersion term can be achieved only with the knowledge of $H(P)$ and $V(P)$. However, one has to work much harder to determine the third-order term. For conclusive results on the third-order term for fixed-to-variable length source coding, the reader is referred to the elegant work by Kosut and Sankar [100, 101]. The technique outlined in this section involves the method of types.

Let $M_u^*(P, \varepsilon)$ be the almost lossless source coding non-asymptotic fundamental limit where the source code (f, φ) is ignorant of the probability distribution of the source P , except for the entropy $H(P)$ and the varentropy $V(P)$.

Theorem 3.2. If the source $P \in \mathcal{P}(\mathcal{X})$ satisfies $V(P) > 0$, then

$$\log M_u^*(P^n, \varepsilon) \leq nH(P) - \sqrt{nV(P)}\Phi^{-1}(\varepsilon) + (|\mathcal{X}| - 1)\log n + O(1). \quad (3.20)$$

The proof we present here results in a third-order term that is likely to be far from optimal but we present this proof to demonstrate the similarity to the classical proof of the fixed-to-fixed length source coding error exponent using the method of types [39, Thm. 2.15].

Proof of Theorem 3.2. Let $\mathcal{X} = \{a_1, \dots, a_d\}$ without loss of generality. Set M , the size of the code, to be the smallest integer exceeding

$$\exp \left[nH(P) - \sqrt{nV(P)}\Phi^{-1} \left(\varepsilon - \frac{c}{\sqrt{n}} \right) + (d-1)\log(n+1) \right], \quad (3.21)$$

for some finite constant $c > 0$ (given in Theorem 1.10). Let $\mathcal{K}$ be the set of sequences in $\mathcal{X}^n$ whose empirical entropy is no larger than

$$\gamma := \frac{1}{n} \log M - \frac{(d-1)\log(n+1)}{n}. \quad (3.22)$$

In other words,

$$\mathcal{K} := \bigcup_{Q \in \mathcal{P}_n(\mathcal{X}): H(Q) \leq \gamma} \mathcal{T}_Q. \quad (3.23)$$

Encode all sequences in $\mathcal{K}$ in a one-to-one way and sequences not in $\mathcal{K}$ arbitrarily. By the type counting lemma in (1.27) and Lemma 1.2 (size of type class), we have

$$|\mathcal{K}| \leq \sum_{Q \in \mathcal{P}_n(\mathcal{X}): H(Q) \leq \gamma} \exp(nH(Q)) \leq (n+1)^{d-1} \exp(n\gamma) \leq M \quad (3.24)$$

so the number of sequences that can be reconstructed without error is at most M as required. An error occurs if and only if the source sequence has empirical entropy exceeding γ , i.e., the error probability is

$$\mathbf{p} := \Pr(H(P_{X^n}) > \gamma) \quad (3.25)$$

where $P_{X^n} \in \mathcal{P}_n(\mathcal{X})$ is the random type of $X^n \in \mathcal{X}^n$. This probability can be written as

$$\mathbf{p} = \Pr(f(P_{X^n} - P) > \gamma), \quad (3.26)$$

where the function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as

$$f(\mathbf{v}) = H(\mathbf{v} + P), \quad (3.27)$$

In (3.26) and (3.27), we regarded the type P_{X^n} and the true distribution P as vectors of length $d = |\mathcal{X}|$, and $H(\mathbf{w}) = -\sum_j w_j \log w_j$ is the entropy. Note that the argument of $f(\cdot)$ in (3.26) can be written as

$$P_{X^n} - P = \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \mathbb{1}\{X_i = a_1\} - P(a_1) \\ \vdots \\ \mathbb{1}\{X_i = a_d\} - P(a_d) \end{bmatrix} =: \frac{1}{n} \sum_{i=1}^n U_i^d. \quad (3.28)$$

Since $U_i^d := [\mathbb{1}\{X_i = a_1\} - P(a_1), \dots, \mathbb{1}\{X_i = a_d\} - P(a_d)]'$ for $i = 1, \dots, n$ are zero-mean, i.i.d. random vectors, we may appeal to the Berry-Esseen theorem for functions of i.i.d. random vectors in Theorem 1.10. Indeed, we note that $f(\mathbf{0}) = H(P)$, the Jacobian of f evaluated at $\mathbf{v} = \mathbf{0}$ is

$$\mathbf{J} = \left[\log\left(\frac{1}{eP(a_1)}\right) \quad \dots \quad \log\left(\frac{1}{eP(a_d)}\right) \right], \quad (3.29)$$

and the $(s, t) \in \{1, \dots, d\}^2$ element of the covariance matrix of U_1^d is

$$[\text{Cov}(U_1^d)]_{st} = \begin{cases} P(a_s)(1 - P(a_s)) & s = t \\ -P(a_s)P(a_t) & s \neq t \end{cases}. \quad (3.30)$$

As such, by a routine multiplication of matrices,

$$\mathbf{J} \text{Cov}(U_1^d) \mathbf{J}' = V(P), \quad (3.31)$$

the varentropy of the source. We deduce from Theorem 1.10 that

$$\mathfrak{p} \leq \Phi\left(\frac{H(P) - \gamma}{\sqrt{V(P)/n}}\right) + \frac{c}{\sqrt{n}} \quad (3.32)$$

where c is a finite positive constant (depending on P). By the choice of M and γ in (3.21)–(3.22), we see that $\mathfrak{p}$ is no larger than ε . $\square$

This coding scheme is partially universal in the sense that $H(P)$ and $V(P)$ need to be known to be used to determine and the threshold γ in (3.22), but otherwise no other characteristic of the source P is required to be known. This achievability proof strategy is rather general and can be applied to rate-distortion (cf. Section 3.6), channel coding, joint source-channel coding [87, 170], as well as multi-terminal problems [157] (cf. Section 6.3).

The point we would like to emphasize in this section is the following: In large deviations (error exponent) analysis of almost lossless source coding, the probability of error in (3.25) is evaluated using, for example, Sanov's theorem [39, Ex. 2.12], or refined versions of it [39, Ex. 2.7(c)]. In the above proof, the probability of error is instead estimated using the Berry-Esseen theorem (Theorem 1.10) since the deviation of the code rate from the first-order fundamental limit $H(P)$ is of the order $\Theta(\frac{1}{\sqrt{n}})$ instead of a constant. Essentially, the proof of Theorem 3.2 hinges on the fact that for a random vector X^n with distribution P^n , the entropy of the type P_{X^n} , namely the empirical entropy $\hat{H}(X^n)$, satisfies the following central limit relation:

$$\sqrt{n}(\hat{H}(X^n) - H(P)) \xrightarrow{d} \mathcal{N}(0, V(P)). \quad (3.33)$$

Finally, we note that the technique to bound the probability in (3.26) is similar to that suggested by Kosut and Sankar [101, Lem. 1].

3.4 Lossy Source Coding: Non-Asymptotic Bounds

In the second half of this chapter, we consider the lossy source coding problem where the source $P \in \mathcal{P}(\mathcal{X})$ does not have to be discrete. The