Network Information Theory

This comprehensive treatment of network information theory and its applications provides the first unified coverage of both classical and recent results. With an approach that balances the introduction of new models and new coding techniques, readers are guided through Shannon's point-to-point information theory, single-hop networks, multihop networks, and extensions to distributed computing, secrecy, wireless communication, and networking. Elementary mathematical tools and techniques are used throughout, requiring only basic knowledge of probability, whilst unified proofs of coding theorems are based on a few simple lemmas, making the text accessible to newcomers. Key topics covered include successive cancellation and superposition coding, MIMO wireless communication, network coding, and cooperative relaying. Also covered are feedback and interactive communication, capacity approximations and scaling laws, and asynchronous and random access channels. This book is ideal for use in the classroom, for self-study, and as a reference for researchers and engineers in industry and academia.

Abbas El Gamal is the Hitachi America Chaired Professor in the School of Engineering and the Director of the Information Systems Laboratory in the Department of Electrical Engineering at Stanford University. In the field of network information theory, he is best known for his seminal contributions to the relay, broadcast, and interference channels; multiple description coding; coding for noisy networks; and energy-efficient packet scheduling and throughput–delay tradeoffs in wireless networks. He is a Fellow of IEEE and the winner of the 2012 Claude E. Shannon Award, the highest honor in the field of information theory.

Young-Han Kim is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of California, San Diego. His research focuses on information theory and statistical signal processing. He is a recipient of the 2008 NSF Faculty Early Career Development (CAREER) Award and the 2009 US–Israel Binational Science Foundation Bergmann Memorial Award.

NETWORK INFORMATION THEORY

Abbas El Gamal
Stanford University

Young-Han Kim
University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107008731

First published 2011
3rd printing 2014

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-00873-1 Hardback

Additional resources for this publication at www.cambridge.org/9781107008731

To our families
whose love and support
made this book possible

Contents

Part III Multihop Networks

Appendices

Preface

Network information theory aims to establish the fundamental limits on information flow in networks and the optimal coding schemes that achieve these limits. It extends Shannon's fundamental theorems on point-to-point communication and the Ford–Fulkerson max-flow min-cut theorem for graphical unicast networks to general networks with multiple sources and destinations and shared resources. Although the theory is far from complete, many elegant results and techniques have been developed over the past forty years with potential applications in real-world networks. This book presents these results in a coherent and simplified manner that should make the subject accessible to graduate students and researchers in electrical engineering, computer science, statistics, and related fields, as well as to researchers and practitioners in industry.

The first paper on network information theory was on the two-way channel by Shannon (1961). This was followed a decade later by seminal papers on the broadcast channel by Cover (1972), the multiple access channel by Ahlswede (1971, 1974) and Liao (1972), and distributed lossless compression by Slepian and Wolf (1973a). These results spurred a flurry of research on network information theory from the mid 1970s to the early 1980s with many new results and techniques developed; see the survey papers by van der Meulen (1977) and El Gamal and Cover (1980), and the seminal book by Csiszár and Körner (1981b). However, many problems, including Shannon's two-way channel, remained open and there was little interest in these results from communication theorists or practitioners. The period from the mid 1980s to the mid 1990s represents a "lost decade" for network information theory during which very few papers were published and many researchers shifted their focus to other areas. The advent of the Internet and wireless communication, fueled by advances in semiconductor technology, compression and error correction coding, signal processing, and computer science, revived the interest in this subject and there has been an explosion of activities in the field since the mid 1990s. In addition to progress on old open problems, recent work has dealt with new network models, new approaches to coding for networks, capacity approximations and scaling laws, and topics at the intersection of networking and information theory. Some of the techniques developed in network information theory, such as successive cancellation decoding, multiple description coding, successive refinement of information, and network coding, are being implemented in real-world networks.

Development of the Book

The idea of writing this book started a long time ago when Tom Cover and the first author considered writing a monograph based on their aforementioned 1980 survey paper. The first author then put together a set of handwritten lecture notes and used them to teach a course on multiple user information theory at Stanford University from 1982 to 1984. In response to high demand from graduate students in communication and information theory, he resumed teaching the course in 2002 and updated the early lecture notes with recent results. These updated lecture notes were used also in a course at EPFL in the summer of 2003. In 2007 the second author, who was in the 2002 class, started teaching a similar course at UC San Diego and the authors decided to collaborate on expanding the lecture notes into a textbook. Various versions of the lecture notes have been used since then in courses at Stanford University, UC San Diego, the Chinese University of Hong Kong, UC Berkeley, Tsinghua University, Seoul National University, University of Notre Dame, and McGill University among others. The lecture notes were posted on arXiv in January 2010. This book is based on these notes. Although we have made an effort to provide a broad coverage of the results in the field, we do not claim to be all-inclusive. The explosion in the number of papers on the subject in recent years makes it almost impossible to provide a complete coverage in a single textbook.

Organization of the Book

We considered several high-level organizations of the material in the book, from source coding to channel coding or vise versa, from graphical networks to general networks, or along historical lines. We decided on a pedagogical approach that balances the introduction of new network models and new coding techniques. We first discuss single-hop networks and then multihop networks. Within each type of network, we first study channel coding settings, followed by their source coding counterparts, and then joint source–channel coding. There were several important topics that did not fit neatly into this organization, which we grouped under Extensions. The book deals mainly with discrete memoryless and Gaussian network models because little is known about the limits on information flow for more complex models. Focusing on these models also helps us present the coding schemes and proof techniques in their simplest possible forms.

The first chapter provides a preview of network information theory using selected examples from the book. The rest of the material is divided into four parts and a set of appendices.

Part I. Background (Chapters 2 and 3). We present the needed basic information theory background, introduce the notion of typicality and related lemmas used throughout the book, and review Shannon's point-to-point communication coding theorems.

Part II. Single-hop networks (Chapters 4 through 14). We discuss networks with single-round one-way communication. Here each node is either a sender or a receiver. The material is divided into three types of communication settings.

- *Independent messages over noisy channels (Chapters 4 through 9).* We discuss noisy

single-hop network building blocks, beginning with multiple access channels (many-to-one communication) in Chapter 4, followed by broadcast channels (one-to-many communication) in Chapters 5 and 8, and interference channels (multiple one-to-one communications) in Chapter 6. We split the discussion on broadcast channels for a pedagogical reason—the study of general broadcast channels in Chapter 8 requires techniques that are introduced more simply through the discussion of channels with state in Chapter 7. In Chapter 9, we study Gaussian vector channels, which model multiple-antenna (multiple-input multiple-output/MIMO) communication systems.

- *Correlated sources over noiseless links (Chapters 10 through 13).* We discuss the source coding counterparts of the noisy single-hop network building blocks, beginning with distributed lossless compression in Chapter 10, followed by lossy compression with side information in Chapter 11, distributed lossy compression in Chapter 12, and multiple description coding in Chapter 13. Again we spread the discussion on distributed compression over three chapters to help develop new ideas gradually.
- *Correlated sources over noisy channels (Chapter 14).* We discuss the general setting of sending uncompressed sources over noisy single-hop networks.

Part III. Multihop networks (Chapters 15 through 20). We discuss networks with relaying and multiple rounds of communication. Here some of the nodes can act as both sender and receiver. In an organization parallel to Part II, the material is divided into three types of settings.

- *Independent messages over graphical networks (Chapter 15).* We discuss coding for networks modeled by graphs beyond simple routing.
- *Independent messages over noisy networks (Chapters 16 through 19).* In Chapter 16, we discuss the relay channel, which is a simple two-hop network with a sender, a receiver, and a relay. We then discuss channels with feedback and the two-way channel in Chapter 17. We extend results on the relay channel and the two-way channel to general noisy networks in Chapter 18. We further discuss approximations and scaling laws for the capacity of large wireless networks in Chapter 19.
- *Correlated sources over graphical networks (Chapter 20).* We discuss source coding counterparts of the channel coding problems in Chapters 15 through 18.

Part IV. Extensions (Chapters 21 through 24). We study extensions of the theory discussed in the first three parts of the book to communication for computing in Chapter 21, communication with secrecy constraints in Chapter 22, wireless fading channels in Chapter 23, and to problems at the intersection of networking and information theory in Chapter 24.

Appendices. To make the book as self-contained as possible, Appendices A, B, and E provide brief reviews of the necessary background on convex sets and functions, probability and estimation, and convex optimization, respectively. Appendices C and D describe techniques for bounding the cardinality of auxiliary random variables appearing in many

capacity and rate region characterizations, and the Fourier–Motzkin elimination procedure, respectively.

Presentation of the Material

Each chapter typically contains both teaching material and advanced topics. Starred sections contain topics that are either too technical to be discussed in detail or are not essential to the main flow of the material. The chapter ends with a bulleted summary of key points and open problems, bibliographic notes, and problems on missing proof steps in the text followed by exercises around the key ideas. Some of the more technical and less central proofs are delegated to appendices at the end of each chapter in order to help the reader focus on the main ideas and techniques.

The book follows the adage "a picture is worth a thousand words." We use illustrations and examples to provide intuitive explanations of models and concepts. The proofs follow the principle of making everything as simple as possible but not simpler. We use elementary tools and techniques, requiring only basic knowledge of probability and some level of mathematical maturity, for example, at the level of a first course on information theory. The achievability proofs are based on joint typicality, which was introduced by Shannon in his 1948 paper and further developed in the 1970s by Forney and Cover. We take this approach one step further by developing a set of simple lemmas to reduce the repetitiveness in the proofs. We show how the proofs for discrete memoryless networks can be extended to their Gaussian counterparts by using a discretization procedure and taking appropriate limits. Some of the proofs in the book are new and most of them are simplified—and in some cases more rigorous—versions of published proofs.

Use of the Book in Courses

As mentioned earlier, the material in this book has been used in courses on network information theory at several universities over many years. We hope that the publication of the book will help make such a course more widely adopted. One of our main motivations for writing the book, however, is to broaden the audience for network information theory. Current education of communication and networking engineers encompasses primarily point-to-point communication and wired networks. At the same time, many of the innovations in modern communication and networked systems concern more efficient use of shared resources, which is the focus of network information theory. We believe that the next generation of communication and networking engineers can benefit greatly from having a working knowledge of network information theory. We have made every effort to present some of the most relevant material to this audience as simply and clearly as possible. In particular, the material on Gaussian channels, wireless fading channels, and Gaussian networks can be readily integrated into an advanced course on wireless communication.

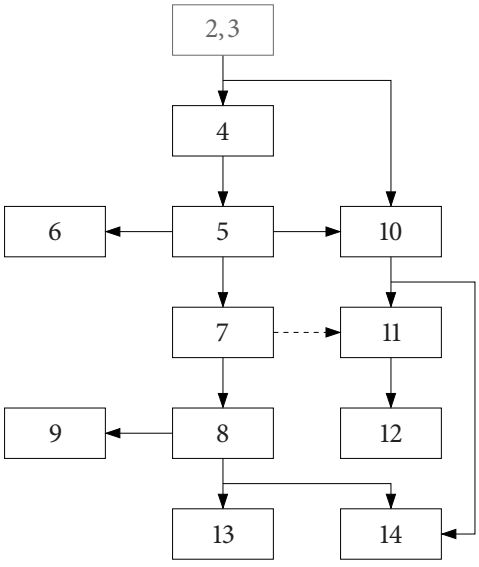
The book can be used as a main text in a one-quarter/semester first course on information theory with emphasis on communication or a one-quarter second course on information theory, or as a supplementary text in courses on communication, networking,

computer science, and statistics. Most of the teaching material in the book can be covered in a two-quarter course sequence. Slides for such courses are posted at http://arxiv.org/abs/1001.3404/.

Dependence Graphs

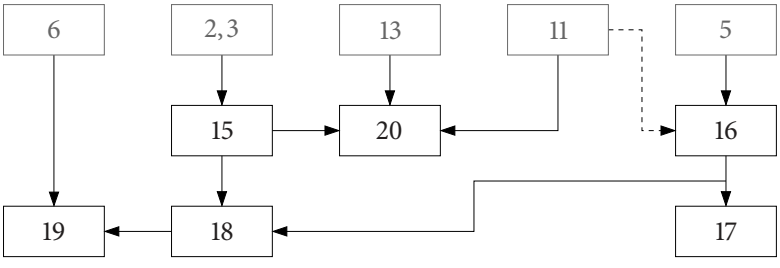
The following graphs depict the dependence of each chapter on its preceding chapters. Each box contains the chapter number and lighter boxes represent dependence on previous parts. Solid edges represent required reading and dashed edges represent recommended reading.

Part II.



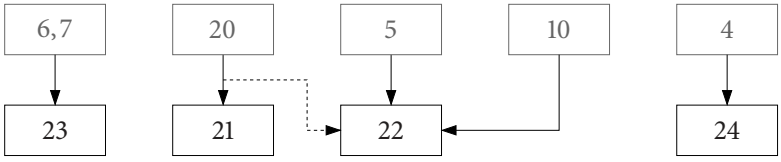
2 Information measures and typicality
3 Point-to-point information theory
4 Multiple access channels
5 Degraded broadcast channels
6 Interference channels
7 Channels with state
8 General broadcast channels
9 Gaussian vector channels
10 Distributed lossless compression
11 Lossy compression with side information
12 Distributed lossy compression
13 Multiple description coding
14 Joint source–channel coding

Part III.



15 Graphical networks
16 Relay channels
17 Interactive channel coding
18 Discrete memoryless networks
19 Gaussian networks
20 Compression over graphical networks

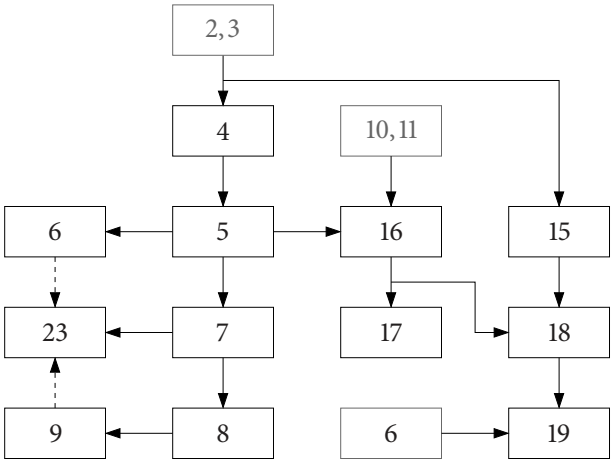
Part IV.



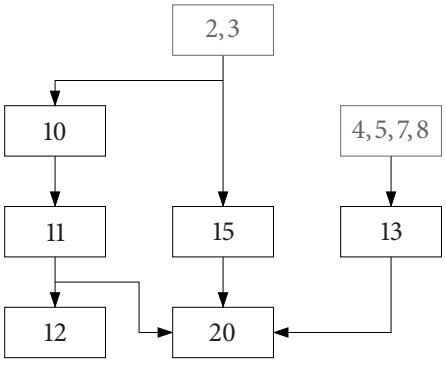
21 Communication for computing
22 Information theoretic secrecy
23 Wireless fading channels
24 Networking and information theory

In addition to the dependence graphs for each part, we provide below some interest-based dependence graphs.

Communication.



Data compression.



Abbas El Gamal
Young-Han Kim

Palo Alto, California
La Jolla, California
July 2011

Acknowledgments

The development of this book was truly a community effort. Many colleagues, teaching assistants of our courses on network information theory, and our postdocs and PhD students provided invaluable input on the content, organization, and exposition of the book, and proofread earlier drafts.

First and foremost, we are indebted to Tom Cover. He taught us everything we know about information theory, encouraged us to write this book, and provided several insightful comments. We are also indebted to our teaching assistants—Ehsan Ardestanizadeh, Chiao-Yi Chen, Yeow-Khiang Chia, Shirin Jalali, Paolo Minero, Haim Permuter, Han-I Su, Sina Zahedi, and Lei Zhao—for their invaluable contributions to the development of this book. In particular, we thank Sina Zahedi for helping with the first set of lecture notes that ultimately led to this book. We thank Han-I Su for his contributions to the chapters on quadratic Gaussian source coding and distributed computing and his thorough proofreading of the entire draft. Yeow-Khiang Chia made invaluable contributions to the chapters on information theoretic secrecy and source coding over graphical networks, contributed several problems, and proofread many parts of the book. Paolo Minero helped with some of the material in the chapter on information theory and networking.

We are also grateful to our PhD students. Bernd Bandemer contributed to the chapter on interference channels and proofread several parts of the book. Sung Hoon Lim contributed to the chapters on discrete memoryless and Gaussian networks. James Mammen helped with the first draft of the lecture notes on scaling laws. Lele Wang and Yu Xiang also provided helpful comments on many parts of the book.

We benefited greatly from discussions with several colleagues. Chandra Nair contributed many of the results and problems in the chapters on broadcast channels. David Tse helped with the organization of the chapters on fading and interference channels. Mehdi Mohseni helped with key proofs in the chapter on Gaussian vector channels. Amin Gohari helped with the organization and several results in the chapter on information theoretic secrecy. Olivier Lévêque helped with some of the proofs in the chapter on Gaussian networks. We often resorted to John Gill for stylistic and editorial advice. Jun Chen, Sae-Young Chung, Amos Lapidoth, Prakash Narayan, Bobak Nazer, Alon Orlitsky, Ofer Shayevitz, Yossi Steinberg, Aslan Tchamkerten, Dimitris Toumpakaris, Sergio Verdú, Mai Vu, Michèle Wigger, Ram Zamir, and Ken Zeger provided helpful input during the writing of this book. We would also like to thank Venkat Anantharam, François Baccelli, Stephen Boyd, Max Costa, Paul Cuff, Suhas Diggavi, Massimo Franceschetti, Michael Gastpar,

Andrea Goldsmith, Bob Gray, Te Sun Han, Tara Javidi, Ashish Khisti, Gerhard Kramer, Mohammad Maddah-Ali, Andrea Montanari, Balaji Prabhakar, Bixio Rimoldi, Anant Sahai, Anand Sarwate, Devavrat Shah, Shlomo Shamai, Emre Telatar, Alex Vardy, Tsachy Weissman, and Lin Zhang.

This book would not have been written without the enthusiasm, inquisitiveness, and numerous contributions of the students who took our courses, some of whom we have already mentioned. In addition, we would like to acknowledge Ekine Akuiyibo, Lorenzo Coviello, Chan-Soo Hwang, Yashodhan Kanoria, Tae Min Kim, Gowtham Kumar, and Moshe Malkin for contributions to some of the material. Himanshu Asnani, Yuxin Chen, Aakanksha Chowdhery, Mohammad Naghshvar, Ryan Peng, Nish Sinha, and Hao Zou provided many corrections to earlier drafts. Several graduate students from UC Berkeley, MIT, Tsinghua, University of Maryland, Tel Aviv University, and KAIST also provided valuable feedback.

We would like to thank our editor Phil Meyler and the rest of the Cambridge staff for their exceptional support during the publication stage of this book. We also thank Kelly Yilmaz for her wonderful administrative support. Finally, we acknowledge partial support for the work in this book from the DARPA ITMANET and the National Science Foundation.

Notation

We introduce the notation and terminology used throughout the book.

Sets, Scalars, and Vectors

We use lowercase letters x, y, ... to denote constants and values of random variables. We use $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$ to denote an $(j - i + 1)$-sequence/column vector for $1 \le i \le j$. When $i = 1$, we always drop the subscript, i.e., $x^j = (x_1, x_2, \ldots, x_j)$. Sometimes we write $\mathbf{x}, \mathbf{y}, \ldots$ for constant vectors with specified dimensions and x_j for the j-th component of $\mathbf{x}$. Let $\mathbf{x}(i)$ be a vector indexed by time i and $x_j(i)$ be the j-th component of $\mathbf{x}(i)$. The sequence of these vectors is denoted by $\mathbf{x}^n = (\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n))$. An all-one column vector $(1, \ldots, 1)$ with a specified dimension is denoted by $\mathbf{1}$.

Let $\alpha, \beta \in [0, 1]$. Then $\bar{\alpha} = (1 - \alpha)$ and $\alpha * \beta = \alpha\bar{\beta} + \beta\bar{\alpha}$.

Let $x^n, y^n \in \{0, 1\}^n$ be binary n-vectors. Then $x^n \oplus y^n$ is the componentwise modulo-2 sum of the two vectors.

Calligraphic letters $\mathcal{X}, \mathcal{Y}, \ldots$ are used exclusively for finite sets and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. The following notation is used for common sets:

- $\mathbb{R}$ is the real line and $\mathbb{R}^d$ is the d-dimensional real Euclidean space.
- $\mathbb{F}_q$ is the finite field GF(q) and $\mathbb{F}_q^d$ is the d-dimensional vector space over GF(q).

Script letters $\mathscr{C}, \mathscr{R}, \mathscr{P}, \ldots$ are used for subsets of $\mathbb{R}^d$.

For a pair of integers $i \le j$, we define the discrete interval $[i : j] = \{i, i + 1, \ldots, j\}$. More generally, for $a \ge 0$ and integer $i \le 2^a$, we define

- $[i : 2^a) = \{i, i + 1, \ldots, 2^{\lfloor a \rfloor}\}$, where $\lfloor a \rfloor$ is the integer part of a, and
- $[i : 2^a] = \{i, i + 1, \ldots, 2^{\lceil a \rceil}\}$, where $\lceil a \rceil$ is the smallest integer $\ge a$.

Probability and Random Variables

The probability of an event $\mathcal{A}$ is denoted by $\mathsf{P}(\mathcal{A})$ and the conditional probability of $\mathcal{A}$ given $\mathcal{B}$ is denoted by $\mathsf{P}(\mathcal{A} \,|\, \mathcal{B})$. We use uppercase letters $X, Y, \ldots$ to denote random variables. The random variables may take values from finite sets $\mathcal{X}, \mathcal{Y}, \ldots$ or from the real line $\mathbb{R}$. By convention, $X = \emptyset$ means that X is a degenerate random variable (unspecified constant) regardless of its support. The probability of the event $\{X \in \mathcal{A}\}$ is denoted by $\mathsf{P}\{X \in \mathcal{A}\}$.

In accordance with the notation for constant vectors, we use $X_i^j = (X_i, \ldots, X_j)$ to denote a $(j - i + 1)$-sequence/column vector of random variables for $1 \le i \le j$. When $i = 1$, we always drop the subscript and use $X^j = (X_1, \ldots, X_j)$.

Let $(X_1, \ldots, X_k)$ be a tuple of k random variables and $\mathcal{J} \subseteq [1 : k]$. The subtuple of random variables with indices from $\mathcal{J}$ is denoted by $X(\mathcal{J}) = (X_j \colon j \in \mathcal{J})$. Similarly, given k random vectors $(X_1^n, \ldots, X_k^n)$,

$$X^n(\mathcal{J}) = (X_j^n \colon j \in \mathcal{J}) = (X_1(\mathcal{J}), \ldots, X_n(\mathcal{J})).$$

Sometimes we write $\mathbf{X}, \mathbf{Y}, \ldots$ for random (column) vectors with specified dimensions and X_j for the j-th component of $\mathbf{X}$. Let $\mathbf{X}(i)$ be a random vector indexed by time i and $X_j(i)$ be the j-th component of $\mathbf{X}(i)$. We denote the sequence of these vectors by $\mathbf{X}^n = (\mathbf{X}(1), \ldots, \mathbf{X}(n))$.

The following notation is used to specify random variables and random vectors.

- $X^n \sim p(x^n)$ means that $p(x^n)$ is the probability mass function (pmf) of the discrete random vector X^n. The function $p_{X^n}(\tilde{x}^n)$ denotes the pmf of X^n with argument $\tilde{x}^n$, i.e., $p_{X^n}(\tilde{x}^n) = \mathsf{P}\{X^n = \tilde{x}^n\}$ for all $\tilde{x}^n \in \mathcal{X}^n$. The function $p(x^n)$ without subscript is understood to be the pmf of the random vector X^n defined over $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.
- $X^n \sim f(x^n)$ means that $f(x^n)$ is the probability density function (pdf) of the continuous random vector X^n.
- $X^n \sim F(x^n)$ means that $F(x^n)$ is the cumulative distribution function (cdf) of X^n.
- $(X^n, Y^n) \sim p(x^n, y^n)$ means that $p(x^n, y^n)$ is the joint pmf of X^n and Y^n.
- $Y^n \,|\, \{X^n \in \mathcal{A}\} \sim p(y^n \,|\, X^n \in \mathcal{A})$ means that $p(y^n \,|\, X^n \in \mathcal{A})$ is the conditional pmf of Y^n given $\{X^n \in \mathcal{A}\}$.
- $Y^n \,|\, \{X^n = x^n\} \sim p(y^n|x^n)$ means that $p(y^n|x^n)$ is the conditional pmf of Y^n given $\{X^n = x^n\}$.
- $p(y^n|x^n)$ is a collection of (conditional) pmfs on $\mathcal{Y}^n$, one for every $x^n \in \mathcal{X}^n$. $f(y^n|x^n)$ and $F(y^n|x^n)$ are similarly defined.
- $Y^n \sim p_{X^n}(y^n)$ means that Y^n has the same pmf as X^n, i.e., $p(y^n) = p_{X^n}(y^n)$. Similar notation is used for conditional probability distributions.

Given a random variable X, the expected value of its function $g(X)$ is denoted by $\mathsf{E}_X(g(X))$, or $\mathsf{E}(g(X))$ in short. The conditional expectation of X given Y is denoted by $\mathsf{E}(X|Y)$. We use $\mathrm{Var}(X) = \mathsf{E}[(X - \mathsf{E}(X))^2]$ to denote the variance of X and $\mathrm{Var}(X|Y) = \mathsf{E}[(X - \mathsf{E}(X|Y))^2 \,|\, Y]$ to denote the conditional variance of X given Y.

For random vectors $\mathbf{X} = X^n$ and $\mathbf{Y} = Y^k$, $K_{\mathbf{X}} = \mathsf{E}[(\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{X} - \mathsf{E}(\mathbf{X}))^T]$ denotes the covariance matrix of $\mathbf{X}$, $K_{\mathbf{XY}} = \mathsf{E}[(\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{Y} - \mathsf{E}(\mathbf{Y}))^T]$ denotes the crosscovariance matrix of $(\mathbf{X}, \mathbf{Y})$, and $K_{\mathbf{X}|\mathbf{Y}} = \mathsf{E}[(\mathbf{X} - \mathsf{E}(\mathbf{X}|\mathbf{Y}))(\mathbf{X} - \mathsf{E}(\mathbf{X}|\mathbf{Y}))^T] = K_{\mathbf{X}-\mathsf{E}(\mathbf{X}|\mathbf{Y})}$ denotes the conditional covariance matrix of $\mathbf{X}$ given $\mathbf{Y}$, that is, the covariance matrix of the minimum mean squared error (MMSE) for estimating $\mathbf{X}$ given $\mathbf{Y}$.

We use the following notation for standard random variables and random vectors:

- $X \sim \text{Bern}(p)$: X is a Bernoulli random variable with parameter $p \in [0, 1]$, i.e.,
$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$
- $X \sim \text{Binom}(n, p)$: X is a binomial random variable with parameters $n \geq 1$ and $p \in [0, 1]$, i.e.,
$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in [0 : n].$$
- $X \sim \text{Unif}(\mathcal{A})$: X is a discrete uniform random variable over a finite set $\mathcal{A}$.
$X \sim \text{Unif}[i : j]$ for integers $j > i$: X is a discrete uniform random variable over $[i : j]$.
- $X \sim \text{Unif}[a, b]$ for $b > a$: X is a continuous uniform random variable over $[a, b]$.
- $X \sim \text{N}(\mu, \sigma^2)$: X is a Gaussian random variable with mean μ and variance σ^2.
$Q(x) = \mathsf{P}\{X > x\}$, $x \in \mathbb{R}$, where $X \sim \text{N}(0, 1)$.
- $\mathbf{X} = X^n \sim \text{N}(\boldsymbol{\mu}, K)$: $\mathbf{X}$ is a Gaussian random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix K, i.e.,
$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |K|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T K^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

We use the notation $\{X_i\} = (X_1, X_2, \ldots)$ to denote a discrete-time random process. The following notation is used for common random processes:

- $\{X_i\}$ is a $\text{Bern}(p)$ process means that $(X_1, X_2, \ldots)$ is a sequence of independent and identically distributed (i.i.d.) $\text{Bern}(p)$ random variables.
- $\{X_i\}$ is a $\text{WGN}(P)$ process means that $(X_1, X_2, \ldots)$ is a sequence of i.i.d. $\text{N}(0, P)$ random variables. More generally, $\{X_i, Y_i\}$ is a 2-$\text{WGN}(P, \rho)$ process means that (X_1, Y_1), $(X_2, Y_2), \ldots$ are i.i.d. jointly Gaussian random variable pairs with $\mathsf{E}(X_1) = \mathsf{E}(Y_1) = 0$, $\mathsf{E}(X_1^2) = \mathsf{E}(Y_1^2) = P$, and correlation coefficient $\rho = \mathsf{E}(X_1 Y_1)/P$.

We say that $X \to Y \to Z$ form a Markov chain if $p(x, y, z) = p(x)p(y|x)p(z|y)$. More generally, we say that $X_1 \to X_2 \to X_3 \to \cdots$ form a Markov chain if $p(x_i|x^{i-1}) = p(x_i|x_{i-1})$ for $i \geq 2$.

Common Functions

The following functions are used frequently. The logarithm function log is assumed to be base 2 unless specified otherwise.

- Binary entropy function: $H(p) = -p \log p - \bar{p} \log \bar{p}$ for $p \in [0, 1]$.
- Gaussian capacity function: $\mathsf{C}(x) = (1/2) \log(1 + x)$ for $x \geq 0$.
- Quadratic Gaussian rate function: $\mathsf{R}(x) = \max\{(1/2) \log x,\ 0\} = (1/2)[\log x]^+$.

ϵ–δ Notation

We use $\epsilon, \epsilon' > 0$ exclusively to denote "small" constants such that $\epsilon' < \epsilon$. We use $\delta(\epsilon) > 0$ to denote a function of ϵ that tends to zero as $\epsilon \to 0$. When there are multiple such functions $\delta_1(\epsilon), \delta_2(\epsilon), \ldots, \delta_k(\epsilon)$, we denote them all by a generic function $\delta(\epsilon)$ that tends to zero as $\epsilon \to 0$ with the understanding that $\delta(\epsilon) = \max\{\delta_1(\epsilon), \delta_2(\epsilon), \ldots, \delta_k(\epsilon)\}$. Similarly, we use $\epsilon_n \geq 0$ to denote a generic function of n that tends to zero as $n \to \infty$.

We say that $a_n \doteq 2^{nb}$ for some constant b if there exists some $\delta(\epsilon)$ (with ϵ defined in the context) such that for n sufficiently large,

$$2^{n(b-\delta(\epsilon))} \leq a_n \leq 2^{n(b+\delta(\epsilon))}.$$

Matrices

We use uppercase letters $A, B, \ldots$ to denote matrices. The entry in the i-th row and the j-th column of a matrix A is denoted by $A(i, j)$ or A_{ij}. A transpose of a matrix A is denoted by A^T, i.e., $A^T(i, j) = A(j, i)$. We use $\mathrm{diag}(a_1, a_2, \ldots, a_d)$ to denote a $d \times d$ diagonal matrix with diagonal elements $a_1, a_2, \ldots, a_d$. The $d \times d$ identity matrix is denoted by I_d. The subscript d is omitted when it is clear from the context. For a square matrix A, $|A| = \det(A)$ denotes the determinant of A and $\mathrm{tr}(A)$ denotes its trace.

A symmetric matrix A is said to be positive definite (denoted by $A \succ 0$) if $\mathbf{x}^T A \mathbf{x} > 0$ for all $\mathbf{x} \neq 0$. If instead $\mathbf{x}^T A \mathbf{x} \geq 0$ for all $\mathbf{x} \neq 0$, then the matrix A is said to be positive semidefinite (denoted by $A \succeq 0$). For symmetric matrices A and B of the same dimension, $A \succ B$ means that $A - B \succ 0$ and $A \succeq B$ means that $A - B \succeq 0$.

A singular value decomposition of an $r \times t$ matrix G of rank d is given by $G = \Phi \Gamma \Psi^T$, where Φ is an $r \times d$ matrix with $\Phi^T \Phi = I_d$, Ψ is a $t \times d$ matrix with $\Psi^T \Psi = I_d$, and $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_d)$ is a $d \times d$ positive diagonal matrix.

For a symmetric positive semidefinite matrix K with an eigenvalue decomposition $K = \Phi \Lambda \Phi^T$, we define its symmetric square root as $K^{1/2} = \Phi \Lambda^{1/2} \Phi^T$, where $\Lambda^{1/2}$ is a diagonal matrix with diagonal elements $\sqrt{\Lambda_{ii}}$. Note that $K^{1/2}$ is symmetric positive definite with $K^{1/2} K^{1/2} = K$. We define the symmetric square root inverse $K^{-1/2}$ of a symmetric positive definite matrix K as the symmetric square root of K^{-1}.

Order Notation

Let $g_1(N)$ and $g_2(N)$ be nonnegative functions on natural numbers.

- $g_1(N) = o(g_2(N))$ means that $g_1(N)/g_2(N)$ tends to zero as $N \to \infty$.
- $g_1(N) = O(g_2(N))$ means that there exist a constant a and an integer n_0 such that $g_1(N) \leq a g_2(N)$ for all $N > n_0$.
- $g_1(N) = \Omega(g_2(N))$ means that $g_2(N) = O(g_1(N))$.
- $g_1(N) = \Theta(g_2(N))$ means that $g_1(N) = O(g_2(N))$ and $g_2(N) = O(g_1(N))$.