Preface

This book is an evolution from my book *A First Course in Information Theory* published in 2002 when network coding was still at its infancy. The last few years have witnessed the rapid development of network coding into a research field of its own in information science. With its root in information theory, network coding has not only brought about a paradigm shift in network communications at large, but also had significant influence on such specific research fields as coding theory, networking, switching, wireless communications, distributed data storage, cryptography, and optimization theory. While new applications of network coding keep emerging, the fundamental results that lay the foundation of the subject are more or less mature. One of the main goals of this book therefore is to present these results in a unifying and coherent manner.

While the previous book focused only on information theory for discrete random variables, the current book contains two new chapters on information theory for continuous random variables, namely the chapter on differential entropy and the chapter on continuous-valued channels. With these topics included, the book becomes more comprehensive and is more suitable to be used as a textbook for a course in an electrical engineering department.

What Is in This book

Out of the 21 chapters in this book, the first 16 chapters belong to Part I, *Components of Information Theory*, and the last 5 chapters belong to Part II, *Fundamentals of Network Coding*. Part I covers the basic topics in information theory and prepares the reader for the discussions in Part II. A brief rundown of the chapters will give a better idea of what is in this book.

Chapter 1 contains a high-level introduction to the contents of this book. First, there is a discussion on the nature of information theory and the main results in Shannon's original paper in 1948 which founded the field. There are also pointers to Shannon's biographies and his works.

Chapter 2 introduces Shannon's information measures for discrete random variables and their basic properties. Useful identities and inequalities in

information theory are derived and explained. Extra care is taken in handling joint distributions with zero probability masses. There is a section devoted to the discussion of maximum entropy distributions. The chapter ends with a section on the entropy rate of a stationary information source.

Chapter 3 is an introduction to the theory of I-Measure which establishes a one-to-one correspondence between Shannon's information measures and set theory. A number of examples are given to show how the use of information diagrams can simplify the proofs of many results in information theory. Such diagrams are becoming standard tools for solving information theory problems.

Chapter 4 is a discussion of zero-error data compression by uniquely decodable codes, with prefix codes as a special case. A proof of the entropy bound for prefix codes which involves neither the Kraft inequality nor the fundamental inequality is given. This proof facilitates the discussion of the redundancy of prefix codes.

Chapter 5 is a thorough treatment of weak typicality. The weak asymptotic equipartition property and the source coding theorem are discussed. An explanation of the fact that a good data compression scheme produces almost i.i.d. bits is given. There is also an introductory discussion of the Shannon–McMillan–Breiman theorem. The concept of weak typicality will be further developed in Chapter 10 for continuous random variables.

Chapter 6 contains a detailed discussion of strong typicality which applies to random variables with finite alphabets. The results developed in this chapter will be used for proving the channel coding theorem and the rate-distortion theorem in the next two chapters.

The discussion in Chapter 7 of the discrete memoryless channel is an enhancement of the discussion in the previous book. In particular, the new definition of the discrete memoryless channel enables rigorous formulation and analysis of coding schemes for such channels with or without feedback. The proof of the channel coding theorem uses a graphical model approach that helps explain the conditional independence of the random variables.

Chapter 8 is an introduction to rate-distortion theory. The version of the rate-distortion theorem here, proved by using strong typicality, is a stronger version of the original theorem obtained by Shannon.

In Chapter 9, the Blahut–Arimoto algorithms for computing the channel capacity and the rate-distortion function are discussed, and a simplified proof for convergence is given. Great care is taken in handling distributions with zero probability masses.

Chapters 10 and 11 are devoted to the discussion of information theory for continuous random variables. Chapter 10 introduces differential entropy and related information measures, and their basic properties are discussed. The asymptotic equipartition property for continuous random variables is proved. The last section on maximum differential entropy distributions echoes the section in Chapter 2 on maximum entropy distributions.

Chapter 11 discusses a variety of continuous-valued channels, with the continuous memoryless channel being the basic building block. In proving the capacity of the memoryless Gaussian channel, a careful justification is given for the existence of the differential entropy of the output random variable. Based on this result, the capacity of a system of parallel/correlated Gaussian channels is obtained. Heuristic arguments leading to the formula for the capacity of the bandlimited white/colored Gaussian channel are given. The chapter ends with a proof of the fact that zero-mean Gaussian noise is the worst additive noise.

Chapter 12 explores the structure of the I-Measure for Markov structures. Set-theoretic characterizations of full conditional independence and Markov random field are discussed. The treatment of Markov random field here maybe too specialized for the average reader, but the structure of the I-Measure and the simplicity of the information diagram for a Markov chain are best explained as a special case of a Markov random field.

Information inequalities are sometimes called the laws of information theory because they govern the impossibilities in information theory. In Chapter 13, the geometrical meaning of information inequalities and the relation between information inequalities and conditional independence are explained in depth. The framework for information inequalities discussed here is the basis of the next two chapters.

Chapter 14 explains how the problem of proving information inequalities can be formulated as a linear programming problem. This leads to a complete characterization of all information inequalities provable by conventional techniques. These inequalities, called Shannon-type inequalities, can be proved by the World Wide Web available software package ITIP. It is also shown how Shannon-type inequalities can be used to tackle the implication problem of conditional independence in probability theory.

Shannon-type inequalities are all the information inequalities known during the first half century of information theory. In the late 1990s, a few new inequalities, called non-Shannon-type inequalities, were discovered. These inequalities imply the existence of laws in information theory beyond those laid down by Shannon. In Chapter 15, we discuss these inequalities and their applications.

Chapter 16 explains an intriguing relation between information theory and group theory. Specifically, for every information inequality satisfied by any joint probability distribution, there is a corresponding group inequality satisfied by any finite group and its subgroups and vice versa. Inequalities of the latter type govern the orders of any finite group and their subgroups. Group-theoretic proofs of Shannon-type information inequalities are given. At the end of the chapter, a group inequality is obtained from a non-Shannon-type inequality discussed in Chapter 15. The meaning and the implication of this inequality are yet to be understood.

Chapter 17 starts Part II of the book with a discussion of the butterfly network, the primary example in network coding. Variations of the butterfly

network are analyzed in detail. The advantage of network coding over store-and-forward in wireless and satellite communications is explained through a simple example. We also explain why network coding with multiple information sources is substantially different from network coding with a single information source.

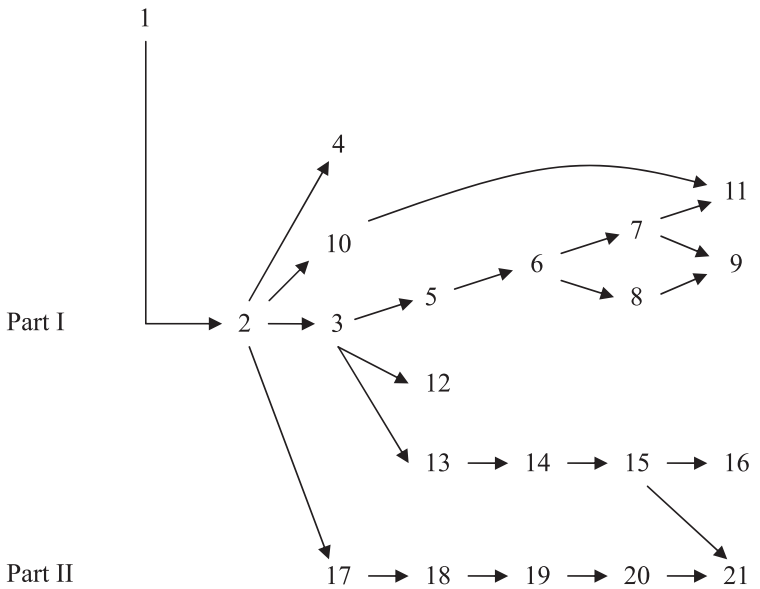
In Chapter 18, the fundamental bound for single-source network coding, called the max-flow bound, is explained in detail. The bound is established for a general class of network codes.

In Chapter 19, we discuss various classes of linear network codes on acyclic networks that achieve the max-flow bound to different extents. Static network codes, a special class of linear network codes that achieves the max-flow bound in the presence of channel failure, are also discussed. Polynomial-time algorithms for constructing these codes are presented.

In Chapter 20, we formulate and analyze convolutional network codes on cyclic networks. The existence of such codes that achieve the max-flow bound is proved.

Network coding theory is further developed in Chapter 21. The scenario when more than one information source are multicast in a point-to-point acyclic network is discussed. An implicit characterization of the achievable information rate region which involves the framework for information inequalities developed in Part I is proved.

How to Use This book



Part I of this book by itself may be regarded as a comprehensive textbook in information theory. The main reason why the book is in the present form is because in my opinion, the discussion of network coding in Part II is incomplete without Part I. Nevertheless, except for Chapter 21 on multi-source network coding, Part II by itself may be used satisfactorily as a textbook on single-source network coding.

An elementary course on probability theory and an elementary course on linear algebra are prerequisites to Part I and Part II, respectively. For Chapter 11, some background knowledge on digital communication systems would be helpful, and for Chapter 20, some prior exposure to discrete-time linear systems is necessary. The reader is recommended to read the chapters according to the above chart. However, one will not have too much difficulty jumping around in the book because there should be sufficient references to the previous relevant sections.

This book inherits the writing style from the previous book, namely that all the derivations are from the first principle. The book contains a large number of examples, where important points are very often made. To facilitate the use of the book, there is a summary at the end of each chapter.

This book can be used as a textbook or a reference book. As a textbook, it is ideal for a two-semester course, with the first and second semesters covering selected topics from Part I and Part II, respectively. A comprehensive instructor's manual is available upon request. Please contact the author at whyeung@ie.cuhk.edu.hk for information and access.

Just like any other lengthy document, this book for sure contains errors and omissions. To alleviate the problem, an errata will be maintained at the book homepage http://www.ie.cuhk.edu.hk/IT_book2/.

Hong Kong, China *Raymond W. Yeung*
December, 2007

2

Information Measures

Shannon's information measures refer to entropy, conditional entropy, mutual information, and conditional mutual information. They are the most important measures of information in information theory. In this chapter, we introduce these measures and establish some basic properties they possess. The physical meanings of these measures will be discussed in depth in subsequent chapters. We then introduce the informational divergence which measures the "distance" between two probability distributions and prove some useful inequalities in information theory. The chapter ends with a section on the entropy rate of a stationary information source.

2.1 Independence and Markov Chains

We begin our discussion in this chapter by reviewing two basic concepts in probability: independence of random variables and Markov chain. All the random variables in this book except for Chapters 10 and 11 are assumed to be discrete unless otherwise specified.

Let X be a random variable taking values in an alphabet $\mathcal{X}$. The probability distribution for X is denoted as $\{p_X(x), x \in \mathcal{X}\}$, with $p_X(x) = \Pr\{X = x\}$. When there is no ambiguity, $p_X(x)$ will be abbreviated as $p(x)$, and $\{p(x)\}$ will be abbreviated as $p(x)$. The *support* of X, denoted by $\mathcal{S}_X$, is the set of all $x \in \mathcal{X}$ such that $p(x) > 0$. If $\mathcal{S}_X = \mathcal{X}$, we say that p is *strictly positive*. Otherwise, we say that p is not strictly positive, or p contains zero probability masses. All the above notations naturally extend to two or more random variables. As we will see, probability distributions with zero probability masses are very delicate, and they need to be handled with great care.

Definition 2.1. *Two random variables X and Y are independent, denoted by $X \perp Y$, if*

$$p(x, y) = p(x)p(y) \tag{2.1}$$

for all x and y (i.e., for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$).

For more than two random variables, we distinguish between two types of independence.

Definition 2.2 (Mutual Independence). *For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are mutually independent if*

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{2.2}$$

for all $x_1, x_2, \cdots, x_n$.

Definition 2.3 (Pairwise Independence). *For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are pairwise independent if X_i and X_j are independent for all $1 \leq i < j \leq n$.*

Note that mutual independence implies pairwise independence. We leave it as an exercise for the reader to show that the converse is not true.

Definition 2.4 (Conditional Independence). *For random variables X, Y, and Z, X is independent of Z conditioning on Y, denoted by $X \perp Z|Y$, if*

$$p(x, y, z)p(y) = p(x, y)p(y, z) \tag{2.3}$$

for all x, y, and z, or equivalently,

$$p(x, y, z) = \begin{cases} \frac{p(x,y)p(y,z)}{p(y)} = p(x, y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.4}$$

The first definition of conditional independence above is sometimes more convenient to use because it is not necessary to distinguish between the cases $p(y) > 0$ and $p(y) = 0$. However, the physical meaning of conditional independence is more explicit in the second definition.

Proposition 2.5. *For random variables X, Y, and Z, $X \perp Z|Y$ if and only if*

$$p(x, y, z) = a(x, y)b(y, z) \tag{2.5}$$

for all x, y, and z such that $p(y) > 0$.

Proof. The 'only if' part follows immediately from the definition of conditional independence in (2.4), so we will only prove the 'if' part. Assume

$$p(x, y, z) = a(x, y)b(y, z) \tag{2.6}$$

for all x, y, and z such that $p(y) > 0$. Then for such x, y, and z, we have

$$p(x, y) = \sum_z p(x, y, z) = \sum_z a(x, y)b(y, z) = a(x, y) \sum_z b(y, z) \tag{2.7}$$

and

$$p(y,z) = \sum_x p(x,y,z) = \sum_x a(x,y)b(y,z) = b(y,z)\sum_x a(x,y). \tag{2.8}$$

Furthermore,

$$p(y) = \sum_z p(y,z) = \left(\sum_x a(x,y)\right)\left(\sum_z b(y,z)\right) > 0. \tag{2.9}$$

Therefore,

$$\frac{p(x,y)p(y,z)}{p(y)} = \frac{\left(a(x,y)\sum_z b(y,z)\right)\left(b(y,z)\sum_x a(x,y)\right)}{\left(\sum_x a(x,y)\right)\left(\sum_z b(y,z)\right)} \tag{2.10}$$

$$= a(x,y)b(y,z) \tag{2.11}$$

$$= p(x,y,z). \tag{2.12}$$

For x, y, and z such that $p(y) = 0$, since

$$0 \le p(x,y,z) \le p(y) = 0, \tag{2.13}$$

we have

$$p(x,y,z) = 0. \tag{2.14}$$

Hence, $X \perp Z|Y$ according to (2.4). The proof is accomplished. □

Definition 2.6 (Markov Chain). *For random variables $X_1, X_2, \cdots, X_n$, where $n \ge 3$, $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if*

$$\begin{aligned} &p(x_1,x_2,\cdots,x_n)p(x_2)p(x_3)\cdots p(x_{n-1}) \\ &= p(x_1,x_2)p(x_2,x_3)\cdots p(x_{n-1},x_n) \end{aligned} \tag{2.15}$$

for all $x_1, x_2, \cdots, x_n$, or equivalently,

$$p(x_1,x_2,\cdots,x_n) =$$

$$\begin{cases} p(x_1,x_2)p(x_3|x_2)\cdots p(x_n|x_{n-1}) & \text{if } p(x_2),p(x_3),\cdots,p(x_{n-1}) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.16}$$

We note that $X \perp Z|Y$ is equivalent to the Markov chain $X \to Y \to Z$.

Proposition 2.7. *$X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if $X_n \to X_{n-1} \to \cdots \to X_1$ forms a Markov chain.*

Proof. This follows directly from the symmetry in the definition of a Markov chain in (2.15). □

In the following, we state two basic properties of a Markov chain. The proofs are left as an exercise.

Proposition 2.8. *$X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if*

$$\begin{aligned} &X_1 \to X_2 \to X_3 \\ &(X_1, X_2) \to X_3 \to X_4 \\ &\qquad\vdots \\ &(X_1, X_2, \cdots, X_{n-2}) \to X_{n-1} \to X_n \end{aligned} \tag{2.17}$$

form Markov chains.

Proposition 2.9. *$X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if*

$$p(x_1, x_2, \cdots, x_n) = f_1(x_1, x_2) f_2(x_2, x_3) \cdots f_{n-1}(x_{n-1}, x_n) \tag{2.18}$$

for all $x_1, x_2, \cdots, x_n$ such that $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$.

Note that Proposition 2.9 is a generalization of Proposition 2.5. From Proposition 2.9, one can prove the following important property of a Markov chain. Again, the details are left as an exercise.

Proposition 2.10 (Markov Subchains). *Let $\mathcal{N}_n = \{1, 2, \cdots, n\}$ and let $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain. For any subset α of $\mathcal{N}_n$, denote $(X_i, i \in \alpha)$ by X_α. Then for any disjoint subsets $\alpha_1, \alpha_2, \cdots, \alpha_m$ of $\mathcal{N}_n$ such that*

$$k_1 < k_2 < \cdots < k_m \tag{2.19}$$

for all $k_j \in \alpha_j$, $j = 1, 2, \cdots, m$,

$$X_{\alpha_1} \to X_{\alpha_2} \to \cdots \to X_{\alpha_m} \tag{2.20}$$

forms a Markov chain. That is, a subchain of $X_1 \to X_2 \to \cdots \to X_n$ is also a Markov chain.

Example 2.11. Let $X_1 \to X_2 \to \cdots \to X_{10}$ form a Markov chain and $\alpha_1 = \{1, 2\}$, $\alpha_2 = \{4\}$, $\alpha_3 = \{6, 8\}$, and $\alpha_4 = \{10\}$ be subsets of $\mathcal{N}_{10}$. Then Proposition 2.10 says that

$$(X_1, X_2) \to X_4 \to (X_6, X_8) \to X_{10} \tag{2.21}$$

also forms a Markov chain.

We have been very careful in handling probability distributions with zero probability masses. In the rest of the section, we show that such distributions are very delicate in general. We first prove the following property of a strictly positive probability distribution involving four random variables.[1]

Proposition 2.12. *Let X_1, X_2, X_3, and X_4 be random variables such that $p(x_1, x_2, x_3, x_4)$ is strictly positive. Then*

$$\left.\begin{array}{l} X_1 \perp X_4 | (X_2, X_3) \\ X_1 \perp X_3 | (X_2, X_4) \end{array}\right\} \Rightarrow X_1 \perp (X_3, X_4) | X_2. \tag{2.22}$$

Proof. If $X_1 \perp X_4 | (X_2, X_3)$, then

$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_2, x_3) p(x_2, x_3, x_4)}{p(x_2, x_3)}. \tag{2.23}$$

On the other hand, if $X_1 \perp X_3 | (X_2, X_4)$, then

$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_2, x_4) p(x_2, x_3, x_4)}{p(x_2, x_4)}. \tag{2.24}$$

Equating (2.23) and (2.24), we have

$$p(x_1, x_2, x_3) = \frac{p(x_2, x_3) p(x_1, x_2, x_4)}{p(x_2, x_4)}. \tag{2.25}$$

Therefore,

$$p(x_1, x_2) = \sum_{x_3} p(x_1, x_2, x_3) \tag{2.26}$$

$$= \sum_{x_3} \frac{p(x_2, x_3) p(x_1, x_2, x_4)}{p(x_2, x_4)} \tag{2.27}$$

$$= \frac{p(x_2) p(x_1, x_2, x_4)}{p(x_2, x_4)} \tag{2.28}$$

or

$$\frac{p(x_1, x_2, x_4)}{p(x_2, x_4)} = \frac{p(x_1, x_2)}{p(x_2)}. \tag{2.29}$$

Hence from (2.24),

$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_2, x_4) p(x_2, x_3, x_4)}{p(x_2, x_4)} = \frac{p(x_1, x_2) p(x_2, x_3, x_4)}{p(x_2)} \tag{2.30}$$

for all x_1, x_2, x_3, and x_4, i.e., $X_1 \perp (X_3, X_4) | X_2$. □

[1] Proposition 2.12 is called the *intersection* axiom in Bayesian networks. See [287].

If $p(x_1, x_2, x_3, x_4) = 0$ for some x_1, x_2, x_3, and x_4, i.e., p is not strictly positive, the arguments in the above proof are not valid. In fact, the proposition may not hold in this case. For instance, let $X_1 = Y$, $X_2 = Z$, and $X_3 = X_4 = (Y, Z)$, where Y and Z are independent random variables. Then $X_1 \perp X_4|(X_2, X_3)$, $X_1 \perp X_3|(X_2, X_4)$, but $X_1 \not\perp (X_3, X_4)|X_2$. Note that for this construction, p is not strictly positive because $p(x_1, x_2, x_3, x_4) = 0$ if $x_3 \neq (x_1, x_2)$ or $x_4 \neq (x_1, x_2)$.

The above example is somewhat counter-intuitive because it appears that Proposition 2.12 should hold for all probability distributions via a continuity argument[2] which would go like this. For any distribution p, let $\{p_k\}$ be a sequence of strictly positive distributions such that $p_k \to p$ and p_k satisfies (2.23) and (2.24) for all k, i.e.,

$$p_k(x_1, x_2, x_3, x_4)p_k(x_2, x_3) = p_k(x_1, x_2, x_3)p_k(x_2, x_3, x_4) \tag{2.31}$$

and

$$p_k(x_1, x_2, x_3, x_4)p_k(x_2, x_4) = p_k(x_1, x_2, x_4)p_k(x_2, x_3, x_4). \tag{2.32}$$

Then by the proposition, p_k also satisfies (2.30), i.e.,

$$p_k(x_1, x_2, x_3, x_4)p_k(x_2) = p_k(x_1, x_2)p_k(x_2, x_3, x_4). \tag{2.33}$$

Letting $k \to \infty$, we have

$$p(x_1, x_2, x_3, x_4)p(x_2) = p(x_1, x_2)p(x_2, x_3, x_4) \tag{2.34}$$

for all x_1, x_2, x_3, and x_4, i.e., $X_1 \perp (X_3, X_4)|X_2$. Such an argument would be valid if there always exists a sequence $\{p_k\}$ as prescribed. However, the existence of the distribution $p(x_1, x_2, x_3, x_4)$ constructed immediately after Proposition 2.12 simply says that it is not always possible to find such a sequence $\{p_k\}$.

Therefore, probability distributions which are not strictly positive can be very delicate. For strictly positive distributions, we see from Proposition 2.5 that their conditional independence structures are closely related to the factorization problem of such distributions, which has been investigated by Chan [59].

2.2 Shannon's Information Measures

We begin this section by introducing the *entropy* of a random variable. As we will see shortly, all Shannon's information measures can be expressed as linear combinations of entropies.

[2] See Section 2.3 for a more detailed discussion on continuous functionals.

Definition 2.13. *The entropy $H(X)$ of a random variable X is defined as*

$$H(X) = -\sum_x p(x) \log p(x). \tag{2.35}$$

In the definitions of all information measures, we adopt the convention that summation is taken over the corresponding support. Such a convention is necessary because $p(x) \log p(x)$ in (2.13) is undefined if $p(x) = 0$.

The base of the logarithm in (2.13) can be chosen to be any convenient real number greater than 1. We write $H(X)$ as $H_\alpha(X)$ when the base of the logarithm is α. When the base of the logarithm is 2, the unit for entropy is the *bit*. When the base of the logarithm is e, the unit for entropy is the *nat*. When the base of the logarithm is an integer $D \geq 2$, the unit for entropy is the *D-it* (D-ary digit). In the context of source coding, the base is usually taken to be the size of the code alphabet. This will be discussed in Chapter 4.

In computer science, a bit means an entity which can take the value 0 or 1. In information theory, the entropy of a random variable is measured in bits. The reader should distinguish these two meanings of a bit from each other carefully.

Let $g(X)$ be any function of a random variable X. We will denote the expectation of $g(X)$ by $Eg(X)$, i.e.,

$$Eg(X) = \sum_x p(x)g(x), \tag{2.36}$$

where the summation is over $\mathcal{S}_X$. Then the definition of the entropy of a random variable X can be written as

$$H(X) = -E \log p(X). \tag{2.37}$$

Expressions of Shannon's information measures in terms of expectations will be useful in subsequent discussions.

The entropy $H(X)$ of a random variable X is a functional of the probability distribution $p(x)$ which measures the average amount of information contained in X or, equivalently, the average amount of *uncertainty* removed upon revealing the outcome of X. Note that $H(X)$ depends only on $p(x)$, not on the actual values in $\mathcal{X}$. Occasionally, we also denote $H(X)$ by $H(p)$.

For $0 \leq \gamma \leq 1$, define

$$h_b(\gamma) = -\gamma \log \gamma - (1-\gamma) \log(1-\gamma) \tag{2.38}$$

with the convention $0 \log 0 = 0$, so that $h_b(0) = h_b(1) = 0$. With this convention, $h_b(\gamma)$ is continuous at $\gamma = 0$ and $\gamma = 1$. h_b is called the *binary entropy function*. For a binary random variable X with distribution $\{\gamma, 1-\gamma\}$,

$$H(X) = h_b(\gamma). \tag{2.39}$$

Figure 2.1 is the plot of $h_b(\gamma)$ versus γ in the base 2. Note that $h_b(\gamma)$ achieves the maximum value 1 when $\gamma = \frac{1}{2}$.

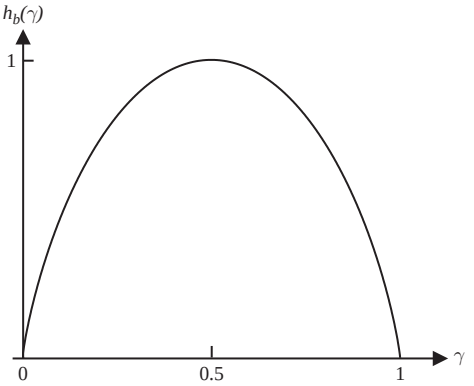


Fig. 2.1. $h_b(\gamma)$ versus γ in the base 2.

The definition of the *joint entropy* of two random variables is similar to the definition of the entropy of a single random variable. Extension of this definition to more than two random variables is straightforward.

Definition 2.14. *The joint entropy $H(X,Y)$ of a pair of random variables X and Y is defined as*

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) = -E \log p(X,Y). \tag{2.40}$$

For two random variables, we define in the following the *conditional entropy* of one random variable when the other random variable is given.

Definition 2.15. *For random variables X and Y, the conditional entropy of Y given X is defined as*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -E \log p(Y|X). \tag{2.41}$$

From (2.41), we can write

$$H(Y|X) = \sum_{x} p(x) \left[-\sum_{y} p(y|x) \log p(y|x) \right]. \tag{2.42}$$

The inner sum is the entropy of Y conditioning on a fixed $x \in \mathcal{S}_X$. Thus we are motivated to express $H(Y|X)$ as

$$H(Y|X) = \sum_{x} p(x) H(Y|X = x), \tag{2.43}$$

where

$$H(Y|X=x) = -\sum_y p(y|x)\log p(y|x). \tag{2.44}$$

Observe that the right-hand sides of (2.13) and (2.44) have exactly the same form. Similarly, for $H(Y|X,Z)$, we write

$$H(Y|X,Z) = \sum_z p(z)H(Y|X,Z=z), \tag{2.45}$$

where

$$H(Y|X,Z=z) = -\sum_{x,y} p(x,y|z)\log p(y|x,z). \tag{2.46}$$

Proposition 2.16.

$$H(X,Y) = H(X) + H(Y|X) \tag{2.47}$$

and

$$H(X,Y) = H(Y) + H(X|Y). \tag{2.48}$$

Proof. Consider

$$\begin{aligned} H(X,Y) &= -E\log p(X,Y) && (2.49)\\ &= -E\log[p(X)p(Y|X)] && (2.50)\\ &= -E\log p(X) - E\log p(Y|X) && (2.51)\\ &= H(X) + H(Y|X). && (2.52) \end{aligned}$$

Note that (2.50) is justified because the summation of the expectation is over $\mathcal{S}_{XY}$, and we have used the linearity of expectation[3] to obtain (2.51). This proves (2.47), and (2.48) follows by symmetry. □

This proposition has the following interpretation. Consider revealing the outcome of a pair of random variables X and Y in two steps: first the outcome of X and then the outcome of Y. Then the proposition says that the total amount of uncertainty removed upon revealing both X and Y is equal to the sum of the uncertainty removed upon revealing X (uncertainty removed in the first step) and the uncertainty removed upon revealing Y once X is known (uncertainty removed in the second step).

Definition 2.17. *For random variables X and Y, the mutual information between X and Y is defined as*

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = E\log\frac{p(X,Y)}{p(X)p(Y)}. \tag{2.53}$$

[3] See Problem 5 at the end of the chapter.

Remark $I(X;Y)$ is symmetrical in X and Y.

Proposition 2.18. *The mutual information between a random variable X and itself is equal to the entropy of X, i.e., $I(X;X) = H(X)$.*

Proof. This can be seen by considering

$$\begin{aligned} I(X;X) &= E \log \frac{p(X)}{p(X)^2} && (2.54)\\ &= -E \log p(X) && (2.55)\\ &= H(X). && (2.56) \end{aligned}$$

The proposition is proved. □

Remark The entropy of X is sometimes called the self-information of X.

Proposition 2.19.

$$I(X;Y) = H(X) - H(X|Y), \tag{2.57}$$

$$I(X;Y) = H(Y) - H(Y|X), \tag{2.58}$$

and

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{2.59}$$

provided that all the entropies and conditional entropies are finite (see Example 2.46 in Section 2.8).

The proof of this proposition is left as an exercise.

From (2.57), we can interpret $I(X;Y)$ as the reduction in uncertainty about X when Y is given or, equivalently, the amount of information about X provided by Y. Since $I(X;Y)$ is symmetrical in X and Y, from (2.58), we can as well interpret $I(X;Y)$ as the amount of information about Y provided by X.

The relations between the (joint) entropies, conditional entropies, and mutual information for two random variables X and Y are given in Propositions 2.16 and 2.19. These relations can be summarized by the diagram in Figure 2.2 which is a variation of the Venn diagram.[4] One can check that all the relations between Shannon's information measures for X and Y which are shown in Figure 2.2 are consistent with the relations given in Propositions 2.16 and 2.19. This one-to-one correspondence between Shannon's information measures and set theory is not just a coincidence for two random variables. We will discuss this in depth when we introduce the I-Measure in Chapter 3.

Analogous to entropy, there is a conditional version of mutual information called conditional mutual information.

[4] The rectangle representing the universal set in a usual Venn diagram is missing in Figure 2.2.

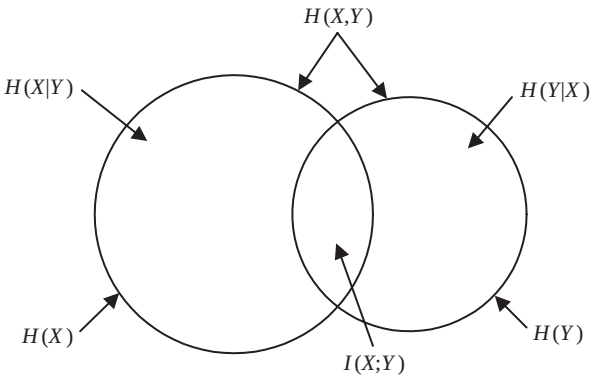


Fig. 2.2. Relationship between entropies and mutual information for two random variables.

Definition 2.20. *For random variables X, Y, and Z, the mutual information between X and Y conditioning on Z is defined as*

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = E \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}. \quad (2.60)$$

Remark $I(X;Y|Z)$ is symmetrical in X and Y.

Analogous to conditional entropy, we write

$$I(X;Y|Z) = \sum_{z} p(z) I(X;Y|Z=z), \quad (2.61)$$

where

$$I(X;Y|Z=z) = \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}. \quad (2.62)$$

Similarly, when conditioning on two random variables, we write

$$I(X;Y|Z,T) = \sum_{t} p(t) I(X;Y|Z,T=t), \quad (2.63)$$

where

$$I(X;Y|Z,T=t) = \sum_{x,y,z} p(x,y,z|t) \log \frac{p(x,y|z,t)}{p(x|z,t)p(y|z,t)}. \quad (2.64)$$

Conditional mutual information satisfies the same set of relations given in Propositions 2.18 and 2.19 for mutual information except that all the terms are now conditioned on a random variable Z. We state these relations in the next two propositions. The proofs are omitted.

Proposition 2.21. *The mutual information between a random variable X and itself conditioning on a random variable Z is equal to the conditional entropy of X given Z, i.e., $I(X;X|Z) = H(X|Z)$.*

Proposition 2.22.

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z), \tag{2.65}$$

$$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z), \tag{2.66}$$

and

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z), \tag{2.67}$$

provided that all the conditional entropies are finite.

Remark All Shannon's information measures are finite if the random variables involved have finite alphabets. Therefore, Propositions 2.19 and 2.22 apply provided that all the random variables therein have finite alphabets.

To conclude this section, we show that all Shannon's information measures are special cases of conditional mutual information. Let Φ be a degenerate random variable, i.e., Φ takes a constant value with probability 1. Consider the mutual information $I(X;Y|Z)$. When $X = Y$ and $Z = \Phi$, $I(X;Y|Z)$ becomes the entropy $H(X)$. When $X = Y$, $I(X;Y|Z)$ becomes the conditional entropy $H(X|Z)$. When $Z = \Phi$, $I(X;Y|Z)$ becomes the mutual information $I(X;Y)$. Thus all Shannon's information measures are special cases of conditional mutual information.

2.3 Continuity of Shannon's Information Measures for Fixed Finite Alphabets

In this section, we prove that for fixed finite alphabets, all Shannon's information measures are continuous functionals of the joint distribution of the random variables involved. To formulate the notion of continuity, we first introduce the *variational distance*[5] as a distance measure between two probability distributions on a common alphabet.

Definition 2.23. *Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. The variational distance between p and q is defined as*

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|. \tag{2.68}$$

[5] The variational distance is also referred to as the $\mathcal{L}^1$ distance in mathematics.

For a fixed finite alphabet $\mathcal{X}$, let $\mathcal{P}_{\mathcal{X}}$ be the set of all distributions on $\mathcal{X}$. Then according to (2.13), the entropy of a distribution p on an alphabet $\mathcal{X}$ is defined as

$$H(p) = -\sum_{x \in \mathcal{S}_p} p(x) \log p(x), \tag{2.69}$$

where $\mathcal{S}_p$ denotes the support of p and $\mathcal{S}_p \subset \mathcal{X}$. In order for $H(p)$ to be continuous with respect to convergence in variational distance[6] at a particular distribution $p \in \mathcal{P}_{\mathcal{X}}$, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|H(p) - H(q)| < \epsilon \tag{2.70}$$

for all $q \in \mathcal{P}_{\mathcal{X}}$ satisfying

$$V(p, q) < \delta, \tag{2.71}$$

or equivalently,

$$\lim_{p' \to p} H(p') = H\left(\lim_{p' \to p} p'\right) = H(p), \tag{2.72}$$

where the convergence $p' \to p$ is in variational distance.

Since $a \log a \to 0$ as $a \to 0$, we define a function $l : [0, \infty) \to \Re$ by

$$l(a) = \begin{cases} a \log a & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{cases}, \tag{2.73}$$

i.e., $l(a)$ is a continuous extension of $a \log a$. Then (2.69) can be rewritten as

$$H(p) = -\sum_{x \in \mathcal{X}} l(p(x)), \tag{2.74}$$

where the summation above is over all x in $\mathcal{X}$ instead of $\mathcal{S}_p$. Upon defining a function $l_x : \mathcal{P}_{\mathcal{X}} \to \Re$ for all $x \in \mathcal{X}$ by

$$l_x(p) = l(p(x)), \tag{2.75}$$

(2.74) becomes

$$H(p) = -\sum_{x \in \mathcal{X}} l_x(p). \tag{2.76}$$

Evidently, $l_x(p)$ is continuous in p (with respect to convergence in variational distance). Since the summation in (2.76) involves a finite number of terms, we conclude that $H(p)$ is a continuous functional of p.

We now proceed to prove the continuity of conditional mutual information which covers all cases of Shannon's information measures. Consider $I(X; Y|Z)$ and let p_{XYZ} be the joint distribution of X, Y, and Z, where the alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are assumed to be finite. From (2.47) and (2.67), we obtain

$$I(X; Y|Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \tag{2.77}$$

[6] Convergence in variational distance is the same as $\mathcal{L}^1$-convergence.

Note that each term on the right-hand side above is the unconditional entropy of the corresponding marginal distribution. Then (2.77) can be rewritten as

$$I_{X;Y|Z}(p_{XYZ}) = H(p_{XZ}) + H(p_{YZ}) - H(p_{XYZ}) - H(p_Z), \tag{2.78}$$

where we have used $I_{X;Y|Z}(p_{XYZ})$ to denote $I(X;Y|Z)$. It follows that

$$\begin{aligned}
&\lim_{p'_{XYZ} \to p_{XYZ}} I_{X;Y|Z}(p'_{XYZ}) \\
&= \lim_{p'_{XYZ} \to p_{XYZ}} [H(p'_{XZ}) + H(p'_{YZ}) - H(p'_{XYZ}) - H(p'_Z)] \qquad (2.79) \\
&= \lim_{p'_{XYZ} \to p_{XYZ}} H(p'_{XZ}) + \lim_{p'_{XYZ} \to p_{XYZ}} H(p'_{YZ}) \\
&\quad - \lim_{p'_{XYZ} \to p_{XYZ}} H(p'_{XYZ}) - \lim_{p'_{XYZ} \to p_{XYZ}} H(p'_Z). \qquad (2.80)
\end{aligned}$$

It can readily be proved, for example, that

$$\lim_{p'_{XYZ} \to p_{XYZ}} p'_{XZ} = p_{XZ}, \tag{2.81}$$

so that

$$\lim_{p'_{XYZ} \to p_{XYZ}} H(p'_{XZ}) = H\left(\lim_{p'_{XYZ} \to p_{XYZ}} p'_{XZ}\right) = H(p_{XZ}) \tag{2.82}$$

by the continuity of $H(\cdot)$ when the alphabets involved are fixed and finite. The details are left as an exercise. Hence, we conclude that

$$\begin{aligned}
&\lim_{p'_{XYZ} \to p_{XYZ}} I_{X;Y|Z}(p'_{XYZ}) \\
&= H(p_{XZ}) + H(p_{YZ}) - H(p_{XYZ}) - H(p_Z) \qquad (2.83) \\
&= I_{X;Y|Z}(p_{XYZ}), \qquad (2.84)
\end{aligned}$$

i.e., $I_{X;Y|Z}(p_{XYZ})$ is a continuous functional of p_{XYZ}.

Since conditional mutual information covers all cases of Shannon's information measures, we have proved that all Shannon's information measures are continuous with respect to convergence in variational distance under the assumption that the alphabets are fixed and finite. It is not difficult to show that under this assumption, convergence in variational distance is equivalent to $\mathcal{L}^2$-convergence, i.e., convergence in Euclidean distance (see Problem 8). It follows that Shannon's information measures are also continuous with respect to $\mathcal{L}^2$-convergence. The variational distance, however, is more often used as a distance measure between two probability distributions because it can be directly related with the informational divergence to be discussed in Section 2.5.

The continuity of Shannon's information measures we have proved in this section is rather restrictive and needs to be applied with caution. In fact, if the alphabets are not fixed, Shannon's information measures are everywhere discontinuous with respect to convergence in a number of commonly used distance measures. We refer the readers to Problems 28–31 for a discussion of these issues.

2.4 Chain Rules

In this section, we present a collection of information identities known as the chain rules which are often used in information theory.

Proposition 2.24 (Chain Rule for Entropy).

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}). \tag{2.85}$$

Proof. The chain rule for $n = 2$ has been proved in Proposition 2.16. We prove the chain rule by induction on n. Assume (2.85) is true for $n = m$, where $m \geq 2$. Then

$$\begin{aligned}
&H(X_1, \cdots, X_m, X_{m+1}) \\
&= H(X_1, \cdots, X_m) + H(X_{m+1}|X_1, \cdots, X_m) && (2.86)\\
&= \sum_{i=1}^{m} H(X_i|X_1, \cdots, X_{i-1}) + H(X_{m+1}|X_1, \cdots, X_m) && (2.87)\\
&= \sum_{i=1}^{m+1} H(X_i|X_1, \cdots, X_{i-1}), && (2.88)
\end{aligned}$$

where in (2.86) we have used (2.47) by letting $X = (X_1, \cdots, X_m)$ and $Y = X_{m+1}$, and in (2.87) we have used (2.85) for $n = m$. This proves the chain rule for entropy. □

The chain rule for entropy has the following conditional version.

Proposition 2.25 (Chain Rule for Conditional Entropy).

$$H(X_1, X_2, \cdots, X_n|Y) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y). \tag{2.89}$$

Proof. The proposition can be proved by considering

$$\begin{aligned}
&H(X_1, X_2, \cdots, X_n|Y) \\
&= H(X_1, X_2, \cdots, X_n, Y) - H(Y) && (2.90)\\
&= H((X_1, Y), X_2, \cdots, X_n) - H(Y) && (2.91)\\
&= H(X_1, Y) + \sum_{i=2}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y) - H(Y) && (2.92)\\
&= H(X_1|Y) + \sum_{i=2}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y) && (2.93)\\
&= \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y), && (2.94)
\end{aligned}$$

where (2.90) and (2.93) follow from Proposition 2.16, while (2.92) follows from Proposition 2.24.

Alternatively, the proposition can be proved by considering

$$\begin{aligned} & H(X_1, X_2, \cdots, X_n|Y) \\ &= \sum_y p(y) H(X_1, X_2, \cdots, X_n|Y = y) && (2.95) \\ &= \sum_y p(y) \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y = y) && (2.96) \\ &= \sum_{i=1}^{n} \sum_y p(y) H(X_i|X_1, \cdots, X_{i-1}, Y = y) && (2.97) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y), && (2.98) \end{aligned}$$

where (2.95) and (2.98) follow from (2.43) and (2.45), respectively, and (2.96) follows from an application of Proposition 2.24 to the joint distribution of $X_1, X_2, \cdots, X_n$ conditioning on $\{Y = y\}$. This proof offers an explanation to the observation that (2.89) can be obtained directly from (2.85) by conditioning on Y in every term. □

Proposition 2.26 (Chain Rule for Mutual Information).

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1}). \tag{2.99}$$

Proof. Consider

$$\begin{aligned} & I(X_1, X_2, \cdots, X_n; Y) \\ &= H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n|Y) && (2.100) \\ &= \sum_{i=1}^{n} [H(X_i|X_1, \cdots, X_{i-1}) - H(X_i|X_1, \cdots, X_{i-1}, Y)] && (2.101) \\ &= \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1}), && (2.102) \end{aligned}$$

where in (2.101), we have invoked both Propositions 2.24 and 2.25. The chain rule for mutual information is proved. □

Proposition 2.27 (Chain Rule for Conditional Mutual Information). *For random variables $X_1, X_2, \cdots, X_n$, Y, and Z,*

$$I(X_1, X_2, \cdots, X_n; Y|Z) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1}, Z). \tag{2.103}$$

Proof. This is the conditional version of the chain rule for mutual information. The proof is similar to that for Proposition 2.25. The details are omitted. □

2.5 Informational Divergence

Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. We very often want to measure how much p is different from q and vice versa. In order to be useful, this measure must satisfy the requirements that it is always nonnegative and it takes the zero value if and only if $p = q$. We denote the support of p and q by $\mathcal{S}_p$ and $\mathcal{S}_q$, respectively. The *informational divergence* defined below serves this purpose.

Definition 2.28. *The informational divergence between two probability distributions p and q on a common alphabet $\mathcal{X}$ is defined as*

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}, \tag{2.104}$$

where E_p denotes expectation with respect to p.

In the above definition, in addition to the convention that the summation is taken over $\mathcal{S}_p$, we further adopt the convention $c \log \frac{c}{0} = \infty$ for $c > 0$. With this convention, if $D(p\|q) < \infty$, then $p(x) = 0$ whenever $q(x) = 0$, i.e., $\mathcal{S}_p \subset \mathcal{S}_q$.

In the literature, the informational divergence is also referred to as *relative entropy* or the *Kullback-Leibler distance*. We note that $D(p\|q)$ is not symmetrical in p and q, so it is not a true *metric* or "distance." Moreover, $D(\cdot\|\cdot)$ does not satisfy the triangular inequality (see Problem 14).

In the rest of the book, the informational divergence will be referred to as *divergence* for brevity. Before we prove that divergence is always nonnegative, we first establish the following simple but important inequality called the *fundamental inequality* in information theory.

Lemma 2.29 (Fundamental Inequality). *For any $a > 0$,*

$$\ln a \le a - 1 \tag{2.105}$$

with equality if and only if $a = 1$.

Proof. Let $f(a) = \ln a - a + 1$. Then $f'(a) = 1/a - 1$ and $f''(a) = -1/a^2$. Since $f(1) = 0$, $f'(1) = 0$, and $f''(1) = -1 < 0$, we see that $f(a)$ attains its maximum value 0 when $a = 1$. This proves (2.105). It is also clear that equality holds in (2.105) if and only if $a = 1$. Figure 2.3 is an illustration of the fundamental inequality. □

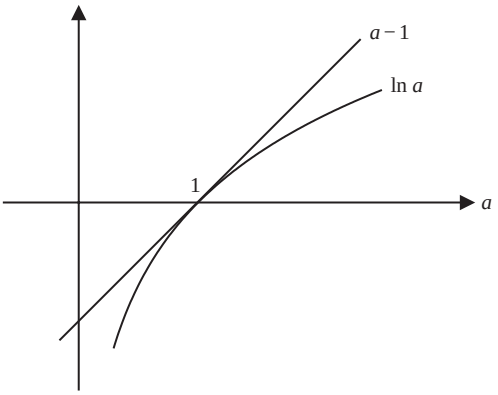


Fig. 2.3. The fundamental inequality $\ln a \le a - 1$.

Corollary 2.30. *For any $a > 0$,*

$$\ln a \ge 1 - \frac{1}{a} \tag{2.106}$$

with equality if and only if $a = 1$.

Proof. This can be proved by replacing a by $1/a$ in (2.105). □

We can see from Figure 2.3 that the fundamental inequality results from the convexity of the logarithmic function. In fact, many important results in information theory are also direct or indirect consequences of the convexity of the logarithmic function!

Theorem 2.31 (Divergence Inequality). *For any two probability distributions p and q on a common alphabet $\mathcal{X}$,*

$$D(p\|q) \ge 0 \tag{2.107}$$

with equality if and only if $p = q$.

Proof. If $q(x) = 0$ for some $x \in \mathcal{S}_p$, then $D(p\|q) = \infty$ and the theorem is trivially true. Therefore, we assume that $q(x) > 0$ for all $x \in \mathcal{S}_p$. Consider

$$D(p\|q) = (\log e) \sum_{x \in \mathcal{S}_p} p(x) \ln \frac{p(x)}{q(x)} \tag{2.108}$$

$$\ge (\log e) \sum_{x \in \mathcal{S}_p} p(x) \left(1 - \frac{q(x)}{p(x)}\right) \tag{2.109}$$

$$= (\log e) \left[\sum_{x \in \mathcal{S}_p} p(x) - \sum_{x \in \mathcal{S}_p} q(x)\right] \tag{2.110}$$

$$\ge 0, \tag{2.111}$$

where (2.109) results from an application of (2.106), and (2.111) follows from

$$\sum_{x\in\mathcal{S}_p} q(x) \le 1 = \sum_{x\in\mathcal{S}_p} p(x). \tag{2.112}$$

This proves (2.107).

For equality to hold in (2.107), equality must hold in (2.109) for all $x \in \mathcal{S}_p$ and also in (2.111). For the former, we see from Lemma 2.29 that this is the case if and only if

$$p(x) = q(x) \quad \text{for all } x \in \mathcal{S}_p, \tag{2.113}$$

which implies

$$\sum_{x\in\mathcal{S}_p} q(x) = \sum_{x\in\mathcal{S}_p} p(x) = 1, \tag{2.114}$$

i.e., (2.111) holds with equality. Thus (2.113) is a necessary and sufficient condition for equality to hold in (2.107).

It is immediate that $p = q$ implies (2.113), so it remains to prove the converse. Since $\sum_x q(x) = 1$ and $q(x) \ge 0$ for all x, $p(x) = q(x)$ for all $x \in \mathcal{S}_p$ implies $q(x) = 0$ for all $x \notin \mathcal{S}_p$, and therefore $p = q$. The theorem is proved. □

We now prove a very useful consequence of the divergence inequality called the *log-sum inequality*.

Theorem 2.32 (Log-Sum Inequality). *For positive numbers $a_1, a_2, \cdots$ and nonnegative numbers $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$,*

$$\sum_i a_i \log \frac{a_i}{b_i} \ge \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i} \tag{2.115}$$

with the convention that $\log \frac{a_i}{0} = \infty$. Moreover, equality holds if and only if $a_i/b_i = $ constant for all i.

The log-sum inequality can easily be understood by writing it out for the case when there are two terms in each of the summations:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \ge (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}. \tag{2.116}$$

Proof. Let $a'_i = a_i / \sum_j a_j$ and $b'_i = b_i / \sum_j b_j$. Then $\{a'_i\}$ and $\{b'_i\}$ are probability distributions. Using the divergence inequality, we have

$$0 \le \sum_i a'_i \log \frac{a'_i}{b'_i} \tag{2.117}$$

$$= \sum_i \frac{a_i}{\sum_j a_j} \log \frac{a_i / \sum_j a_j}{b_i / \sum_j b_j} \tag{2.118}$$

$$= \frac{1}{\sum_j a_j} \left[\sum_i a_i \log \frac{a_i}{b_i} - \left(\sum_i a_i\right) \log \frac{\sum_j a_j}{\sum_j b_j}\right], \tag{2.119}$$

which implies (2.115). Equality holds if and only if $a'_i = b'_i$ for all i or $a_i/b_i = constant$ for all i. The theorem is proved. □

One can also prove the divergence inequality by using the log-sum inequality (see Problem 20), so the two inequalities are in fact equivalent. The log-sum inequality also finds application in proving the next theorem which gives a lower bound on the divergence between two probability distributions on a common alphabet in terms of the variational distance between them. We will see further applications of the log-sum inequality when we discuss the convergence of some iterative algorithms in Chapter 9.

Theorem 2.33 (Pinsker's Inequality).

$$D(p\|q) \geq \frac{1}{2\ln 2} V^2(p,q). \tag{2.120}$$

Both divergence and the variational distance can be used as measures of the difference between two probability distributions defined on the same alphabet. Pinsker's inequality has the important implication that for two probability distributions p and q defined on the same alphabet, if $D(p\|q)$ or $D(q\|p)$ is small, then so is $V(p,q)$. Furthermore, for a sequence of probability distributions q_k, as $k \to \infty$, if $D(p\|q_k) \to 0$ or $D(q_k\|p) \to 0$, then $V(p,q_k) \to 0$. In other words, convergence in divergence is a stronger notion of convergence than convergence in variational distance.

The proof of Pinsker's inequality as well as its consequence discussed above is left as an exercise (see Problems 23 and 24).

2.6 The Basic Inequalities

In this section, we prove that all Shannon's information measures, namely entropy, conditional entropy, mutual information, and conditional mutual information, are always nonnegative. By this, we mean that these quantities are nonnegative for all joint distributions for the random variables involved.

Theorem 2.34. *For random variables X, Y, and Z,*

$$I(X;Y|Z) \geq 0, \tag{2.121}$$

with equality if and only if X and Y are independent when conditioning on Z.

Proof. Observe that

$$\begin{aligned} & I(X;Y|Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} && (2.122) \\ &= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} && (2.123) \\ &= \sum_{z} p(z) D(p_{XY|z} \| p_{X|z} p_{Y|z}), && (2.124) \end{aligned}$$

where we have used $p_{XY|z}$ to denote $\{p(x,y|z), (x,y) \in \mathcal{X} \times \mathcal{Y}\}$, etc. Since for a fixed z, both $p_{XY|z}$ and $p_{X|z}p_{Y|z}$ are joint probability distributions on $\mathcal{X} \times \mathcal{Y}$, we have

$$D(p_{XY|z} \| p_{X|z} p_{Y|z}) \geq 0. \tag{2.125}$$

Therefore, we conclude that $I(X;Y|Z) \geq 0$. Finally, we see from Theorem 2.31 that $I(X;Y|Z) = 0$ if and only if for all $z \in \mathcal{S}_z$,

$$p(x,y|z) = p(x|z)p(y|z) \tag{2.126}$$

or

$$p(x,y,z) = p(x,z)p(y|z) \tag{2.127}$$

for all x and y. Therefore, X and Y are independent conditioning on Z. The proof is accomplished. □

As we have seen in Section 2.2 that all Shannon's information measures are special cases of conditional mutual information, we already have proved that all Shannon's information measures are always nonnegative. The nonnegativity of all Shannon's information measures is called the *basic inequalities.*

For entropy and conditional entropy, we offer the following more direct proof for their nonnegativity. Consider the entropy $H(X)$ of a random variable X. For all $x \in \mathcal{S}_X$, since $0 < p(x) \leq 1$, $\log p(x) \leq 0$. It then follows from the definition in (2.13) that $H(X) \geq 0$. For the conditional entropy $H(Y|X)$ of random variable Y given random variable X, since $H(Y|X = x) \geq 0$ for each $x \in \mathcal{S}_X$, we see from (2.43) that $H(Y|X) \geq 0$.

Proposition 2.35. *$H(X) = 0$ if and only if X is deterministic.*

Proof. If X is deterministic, i.e., there exists $x^* \in \mathcal{X}$ such that $p(x^*) = 1$ and $p(x) = 0$ for all $x \neq x^*$, then $H(X) = -p(x^*) \log p(x^*) = 0$. On the other hand, if X is not deterministic, i.e., there exists $x^* \in \mathcal{X}$ such that $0 < p(x^*) < 1$, then $H(X) \geq -p(x^*) \log p(x^*) > 0$. Therefore, we conclude that $H(X) = 0$ if and only if X is deterministic. □

Proposition 2.36. $H(Y|X) = 0$ *if and only if* Y *is a function of* X.

Proof. From (2.43), we see that $H(Y|X) = 0$ if and only if $H(Y|X = x) = 0$ for each $x \in \mathcal{S}_X$. Then from the last proposition, this happens if and only if Y is deterministic for each given x. In other words, Y is a function of X. □

Proposition 2.37. $I(X;Y) = 0$ *if and only if* X *and* Y *are independent.*

Proof. This is a special case of Theorem 2.34 with Z being a degenerate random variable. □

One can regard (conditional) mutual information as a measure of (conditional) dependency between two random variables. When the (conditional) mutual information is exactly equal to 0, the two random variables are (conditionally) independent.

We refer to inequalities involving Shannon's information measures only (possibly with constant terms) as *information inequalities*. The basic inequalities are important examples of information inequalities. Likewise, we refer to identities involving Shannon's information measures only as *information identities*. From the information identities (2.47), (2.57), and (2.65), we see that all Shannon's information measures can be expressed as linear combinations of entropies provided that the latter are all finite. Specifically,

$$H(Y|X) = H(X,Y) - H(X), \tag{2.128}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{2.129}$$

and

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z). \tag{2.130}$$

Therefore, an information inequality is an inequality which involves only entropies.

As we will see later in the book, information inequalities form the most important set of tools for proving converse coding theorems in information theory. Except for a number of so-called non-Shannon-type inequalities, all known information inequalities are implied by the basic inequalities. Information inequalities will be studied systematically in Chapters 13, 14, and 15. In the next section, we will prove some consequences of the basic inequalities which are often used in information theory.

2.7 Some Useful Information Inequalities

In this section, we prove some useful consequences of the basic inequalities introduced in the last section. Note that the conditional versions of these inequalities can be proved by techniques similar to those used in the proof of Proposition 2.25.

Theorem 2.38 (Conditioning Does Not Increase Entropy).

$$H(Y|X) \leq H(Y) \tag{2.131}$$

with equality if and only if X and Y are independent.

Proof. This can be proved by considering

$$H(Y|X) = H(Y) - I(X;Y) \leq H(Y), \tag{2.132}$$

where the inequality follows because $I(X;Y)$ is always nonnegative. The inequality is tight if and only if $I(X;Y) = 0$, which is equivalent by Proposition 2.37 to X and Y being independent. □

Similarly, it can be shown that

$$H(Y|X,Z) \leq H(Y|Z), \tag{2.133}$$

which is the conditional version of the above proposition. These results have the following interpretation. Suppose Y is a random variable we are interested in, and X and Z are side-information about Y. Then our uncertainty about Y cannot be increased on the average upon receiving side-information X. Once we know X, our uncertainty about Y again cannot be increased on the average upon further receiving side-information Z.

Remark Unlike entropy, the mutual information between two random variables can be increased by conditioning on a third random variable. We refer the reader to Section 3.4 for a discussion.

Theorem 2.39 (Independence Bound for Entropy).

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{2.134}$$

with equality if and only if X_i, $i = 1, 2, \cdots, n$ are mutually independent.

Proof. By the chain rule for entropy,

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}) \tag{2.135}$$

$$\leq \sum_{i=1}^{n} H(X_i), \tag{2.136}$$

where the inequality follows because we have proved in the last theorem that conditioning does not increase entropy. The inequality is tight if and only if it is tight for each i, i.e.,

$$H(X_i|X_1,\cdots,X_{i-1}) = H(X_i) \tag{2.137}$$

for $1 \le i \le n$. From the last theorem, this is equivalent to X_i being independent of $X_1, X_2, \cdots, X_{i-1}$ for each i. Then

$$\begin{aligned} & p(x_1, x_2, \cdots, x_n) \\ &= p(x_1, x_2, \cdots, x_{n-1})p(x_n) && (2.138)\\ &= p(p(x_1, x_2, \cdots, x_{n-2})p(x_{n-1})p(x_n) && (2.139)\\ &\quad \vdots \\ &= p(x_1)p(x_2)\cdots p(x_n) && (2.140)\end{aligned}$$

for all $x_1, x_2, \cdots, x_n$, i.e., $X_1, X_2, \cdots, X_n$ are mutually independent.

Alternatively, we can prove the theorem by considering

$$\begin{aligned} & \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \cdots, X_n) \\ &= -\sum_{i=1}^{n} E \log p(X_i) + E \log p(X_1, X_2, \cdots, X_n) && (2.141)\\ &= -E \log[p(X_1)p(X_2)\cdots p(X_n)] + E \log p(X_1, X_2, \cdots, X_n) && (2.142)\\ &= E \log \frac{p(X_1, X_2, \cdots, X_n)}{p(X_1)p(X_2)\cdots p(X_n)} && (2.143)\\ &= D(p_{X_1X_2\cdots X_n} \| p_{X_1} p_{X_2} \cdots p_{X_n}) && (2.144)\\ &\ge 0, && (2.145)\end{aligned}$$

where equality holds if and only if

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{2.146}$$

for all $x_1, x_2, \cdots, x_n$, i.e., $X_1, X_2, \cdots, X_n$ are mutually independent. □

Theorem 2.40.

$$I(X; Y, Z) \ge I(X; Y), \tag{2.147}$$

with equality if and only if $X \to Y \to Z$ forms a Markov chain.

Proof. By the chain rule for mutual information, we have

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \ge I(X; Y). \tag{2.148}$$

The above inequality is tight if and only if $I(X; Z|Y) = 0$, or $X \to Y \to Z$ forms a Markov chain. The theorem is proved. □

Lemma 2.41. *If $X \to Y \to Z$ forms a Markov chain, then*

$$I(X; Z) \le I(X; Y) \tag{2.149}$$

and

$$I(X; Z) \le I(Y; Z). \tag{2.150}$$

Before proving this inequality, we first discuss its meaning. Suppose X is a random variable we are interested in, and Y is an observation of X. If we infer X via Y, our uncertainty about X on the average is $H(X|Y)$. Now suppose we process Y (either deterministically or probabilistically) to obtain a random variable Z. If we infer X via Z, our uncertainty about X on the average is $H(X|Z)$. Since $X \to Y \to Z$ forms a Markov chain, from (2.149), we have

$$\begin{aligned} H(X|Z) &= H(X) - I(X;Z) && (2.151)\\ &\geq H(X) - I(X;Y) && (2.152)\\ &= H(X|Y), && (2.153) \end{aligned}$$

i.e., further processing of Y can only increase our uncertainty about X on the average.

Proof. Assume $X \to Y \to Z$, i.e., $X \perp Z|Y$. By Theorem 2.34, we have

$$I(X;Z|Y) = 0. \tag{2.154}$$

Then

$$\begin{aligned} I(X;Z) &= I(X;Y,Z) - I(X;Y|Z) && (2.155)\\ &\leq I(X;Y,Z) && (2.156)\\ &= I(X;Y) + I(X;Z|Y) && (2.157)\\ &= I(X;Y). && (2.158) \end{aligned}$$

In (2.155) and (2.157), we have used the chain rule for mutual information. The inequality in (2.156) follows because $I(X;Y|Z)$ is always nonnegative, and (2.158) follows from (2.154). This proves (2.149).

Since $X \to Y \to Z$ is equivalent to $Z \to Y \to X$, we also have proved (2.150). This completes the proof of the lemma. □

From Lemma 2.41, we can prove the more general data processing theorem.

Theorem 2.42 (Data Processing Theorem). *If $U \to X \to Y \to V$ forms a Markov chain, then*

$$I(U;V) \leq I(X;Y). \tag{2.159}$$

Proof. Assume $U \to X \to Y \to V$. Then by Proposition 2.10, we have $U \to X \to Y$ and $U \to Y \to V$. From the first Markov chain and Lemma 2.41, we have

$$I(U;Y) \leq I(X;Y). \tag{2.160}$$

From the second Markov chain and Lemma 2.41, we have

$$I(U;V) \leq I(U;Y). \tag{2.161}$$

Combining (2.160) and (2.161), we obtain (2.159), proving the theorem. □

2.8 Fano's Inequality

In the last section, we have proved a few information inequalities involving only Shannon's information measures. In this section, we first prove an upper bound on the entropy of a random variable in terms of the size of the alphabet. This inequality is then used in the proof of Fano's inequality, which is extremely useful in proving converse coding theorems in information theory.

Theorem 2.43. *For any random variable X,*

$$H(X) \le \log |\mathcal{X}|, \tag{2.162}$$

where $|\mathcal{X}|$ denotes the size of the alphabet $\mathcal{X}$. This upper bound is tight if and only if X is distributed uniformly on $\mathcal{X}$.

Proof. Let u be the uniform distribution on $\mathcal{X}$, i.e., $u(x) = |\mathcal{X}|^{-1}$ for all $x \in \mathcal{X}$. Then

$$\begin{aligned}
&\log |\mathcal{X}| - H(X) \\
&= -\sum_{x \in \mathcal{S}_X} p(x) \log |\mathcal{X}|^{-1} + \sum_{x \in \mathcal{S}_X} p(x) \log p(x) && (2.163)\\
&= -\sum_{x \in \mathcal{S}_X} p(x) \log u(x) + \sum_{x \in \mathcal{S}_X} p(x) \log p(x) && (2.164)\\
&= \sum_{x \in \mathcal{S}_X} p(x) \log \frac{p(x)}{u(x)} && (2.165)\\
&= D(p\|u) && (2.166)\\
&\ge 0, && (2.167)
\end{aligned}$$

proving (2.162). This upper bound is tight if and if only $D(p\|u) = 0$, which from Theorem 2.31 is equivalent to $p(x) = u(x)$ for all $x \in \mathcal{X}$, completing the proof. □

Corollary 2.44. *The entropy of a random variable may take any nonnegative real value.*

Proof. Consider a random variable X defined on a fixed finite alphabet $\mathcal{X}$. We see from the last theorem that $H(X) = \log |\mathcal{X}|$ is achieved when X is distributed uniformly on $\mathcal{X}$. On the other hand, $H(X) = 0$ is achieved when X is deterministic. For $0 \le a \le |\mathcal{X}|^{-1}$, let

$$\begin{aligned}
g(a) &= H\left(\{1 - (|\mathcal{X}| - 1)a, a, \cdots, a\}\right) && (2.168)\\
&= -l(1 - (|\mathcal{X}| - 1)a) - (|\mathcal{X}| - 1)l(a), && (2.169)
\end{aligned}$$

where $l(\cdot)$ is defined in (2.73). Note that $g(a)$ is continuous in a, with $g(0) = 0$ and $g(|\mathcal{X}|^{-1}) = \log |\mathcal{X}|$. For any value $0 < b < \log |\mathcal{X}|$, by the intermediate value theorem of continuous functions, there exists a distribution for X such

that $H(X) = b$. Then we see that $H(X)$ can take any positive value by letting $|\mathcal{X}|$ be sufficiently large. This accomplishes the proof. □

Remark Let $|\mathcal{X}| = D$, or the random variable X is a D-ary symbol. When the base of the logarithm is D, (2.162) becomes

$$H_D(X) \le 1. \tag{2.170}$$

Recall that the unit of entropy is the D-it when the logarithm is in the base D. This inequality says that a D-ary symbol can carry at most 1 D-it of information. This maximum is achieved when X has a uniform distribution. We already have seen the binary case when we discuss the binary entropy function $h_b(p)$ in Section 2.2.

We see from Theorem 2.43 that the entropy of a random variable is finite as long as it has a finite alphabet. However, if a random variable has a countable alphabet,[7] its entropy may or may not be finite. This will be shown in the next two examples.

Example 2.45. Let X be a random variable such that

$$\Pr\{X = i\} = 2^{-i}, \tag{2.171}$$

$i = 1, 2, \cdots$. Then

$$H_2(X) = \sum_{i=1}^{\infty} i 2^{-i} = 2, \tag{2.172}$$

which is finite.

For a random variable X with a countable alphabet and finite entropy, we show in Appendix 2.A that the entropy of X can be approximated by the entropy of a truncation of the distribution of X.

Example 2.46. Let Y be a random variable which takes values in the subset of pairs of integers

$$\left\{(i,j) : 1 \le i < \infty \text{ and } 1 \le j \le \frac{2^{2^i}}{2^i}\right\} \tag{2.173}$$

such that

$$\Pr\{Y = (i,j)\} = 2^{-2^i} \tag{2.174}$$

for all i and j. First, we check that

$$\sum_{i=1}^{\infty} \sum_{j=1}^{2^{2^i}/2^i} \Pr\{Y = (i,j)\} = \sum_{i=1}^{\infty} 2^{-2^i} \left(\frac{2^{2^i}}{2^i}\right) = 1. \tag{2.175}$$

[7] An alphabet is countable means that it is either finite or countably infinite.

Then

$$H_2(Y) = -\sum_{i=1}^{\infty} \sum_{j=1}^{2^{2^i}/2^i} 2^{-2^i} \log_2 2^{-2^i} = \sum_{i=1}^{\infty} 1, \tag{2.176}$$

which does not converge.

Let X be a random variable and $\hat{X}$ be an estimate on X which takes value in the same alphabet $\mathcal{X}$. Let the probability of error P_e be

$$P_e = \Pr\{X \neq \hat{X}\}. \tag{2.177}$$

If $P_e = 0$, i.e., $X = \hat{X}$ with probability 1, then $H(X|\hat{X}) = 0$ by Proposition 2.36. Intuitively, if P_e is small, i.e., $X = \hat{X}$ with probability close to 1, then $H(X|\hat{X})$ should be close to 0. Fano's inequality makes this intuition precise.

Theorem 2.47 (Fano's Inequality). *Let X and $\hat{X}$ be random variables taking values in the same alphabet $\mathcal{X}$. Then*

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1), \tag{2.178}$$

where h_b is the binary entropy function.

Proof. Define a random variable

$$Y = \begin{cases} 0 \text{ if } X = \hat{X} \\ 1 \text{ if } X \neq \hat{X} \end{cases}. \tag{2.179}$$

The random variable Y is an indicator of the error event $\{X \neq \hat{X}\}$, with $\Pr\{Y = 1\} = P_e$ and $H(Y) = h_b(P_e)$. Since Y is a function X and $\hat{X}$,

$$H(Y|X, \hat{X}) = 0. \tag{2.180}$$

Then

$$\begin{aligned}
&H(X|\hat{X}) \\
&= H(X|\hat{X}) + H(Y|X, \hat{X}) && (2.181)\\
&= H(X, Y|\hat{X}) && (2.182)\\
&= H(Y|\hat{X}) + H(X|\hat{X}, Y) && (2.183)\\
&\leq H(Y) + H(X|\hat{X}, Y) && (2.184)\\
&= H(Y) + \sum_{\hat{x} \in \mathcal{X}} \Big[\Pr\{\hat{X} = \hat{x}, Y = 0\} H(X|\hat{X} = \hat{x}, Y = 0) \\
&\quad + \Pr\{\hat{X} = \hat{x}, Y = 1\} H(X|\hat{X} = \hat{x}, Y = 1) \Big]. && (2.185)
\end{aligned}$$

In the above, (2.181) follows from (2.180), (2.184) follows because conditioning does not increase entropy, and (2.185) follows from an application of

(2.43). Now X must take the value $\hat{x}$ if $\hat{X} = \hat{x}$ and $Y = 0$. In other words, X is conditionally deterministic given $\hat{X} = \hat{x}$ and $Y = 0$. Therefore, by Proposition 2.35,

$$H(X|\hat{X} = \hat{x}, Y = 0) = 0. \tag{2.186}$$

If $\hat{X} = \hat{x}$ and $Y = 1$, then X must take a value in the set $\{x \in \mathcal{X} : x \neq \hat{x}\}$ which contains $|\mathcal{X}| - 1$ elements. By Theorem 2.43, we have

$$H(X|\hat{X} = \hat{x}, Y = 1) \leq \log(|\mathcal{X}| - 1), \tag{2.187}$$

where this upper bound does not depend on $\hat{x}$. Hence,

$$\begin{aligned}
&H(X|\hat{X}) \\
&\leq h_b(P_e) + \left(\sum_{\hat{x} \in \mathcal{X}} \Pr\{\hat{X} = \hat{x}, Y = 1\}\right) \log(|\mathcal{X}| - 1) && (2.188)\\
&= h_b(P_e) + \Pr\{Y = 1\} \log(|\mathcal{X}| - 1) && (2.189)\\
&= h_b(P_e) + P_e \log(|\mathcal{X}| - 1), && (2.190)
\end{aligned}$$

which completes the proof. □

Very often, we only need the following simplified version when we apply Fano's inequality. The proof is omitted.

Corollary 2.48. $H(X|\hat{X}) < 1 + P_e \log |\mathcal{X}|$.

Fano's inequality has the following implication. If the alphabet $\mathcal{X}$ is finite, as $P_e \to 0$, the upper bound in (2.178) tends to 0, which implies $H(X|\hat{X})$ also tends to 0. However, this is not necessarily the case if $\mathcal{X}$ is countable, which is shown in the next example.

Example 2.49. Let $\hat{X}$ take the value 0 with probability 1. Let Z be an independent binary random variable taking values in $\{0, 1\}$. Define the random variable X by

$$X = \begin{cases} 0 & \text{if } Z = 0 \\ Y & \text{if } Z = 1 \end{cases}, \tag{2.191}$$

where Y is the random variable in Example 2.46 whose entropy is infinity. Let

$$P_e = \Pr\{X \neq \hat{X}\} = \Pr\{Z = 1\}. \tag{2.192}$$

Then

$$\begin{aligned}
&H(X|\hat{X}) && (2.193)\\
&= H(X) && (2.194)\\
&\geq H(X|Z) && (2.195)\\
&= \Pr\{Z = 0\} H(X|Z = 0) + \Pr\{Z = 1\} H(X|Z = 1) && (2.196)\\
&= (1 - P_e) \cdot 0 + P_e \cdot H(Y) && (2.197)\\
&= \infty && (2.198)
\end{aligned}$$

for any $P_e > 0$. Therefore, $H(X|\hat{X})$ does not tend to 0 as $P_e \to 0$.

2.9 Maximum Entropy Distributions

In Theorem 2.43, we have proved that for any random variable X,

$$H(X) \le \log |\mathcal{X}|, \tag{2.199}$$

with equality when X is distributed uniformly over $\mathcal{X}$. In this section, we revisit this result in the context that X is a real random variable.

To simplify our discussion, all the logarithms are in the base e. Consider the following problem:

> Maximize $H(p)$ over all probability distributions p defined on a countable subset $\mathcal{S}$ of the set of real numbers, subject to
>
> $$\sum_{x \in \mathcal{S}_p} p(x) r_i(x) = a_i \quad \text{for } 1 \le i \le m, \tag{2.200}$$
>
> where $\mathcal{S}_p \subset \mathcal{S}$ and $r_i(x)$ is defined for all $x \in \mathcal{S}$.

The following theorem renders a solution to this problem.

Theorem 2.50. *Let*

$$p^*(x) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(x)} \tag{2.201}$$

for all $x \in \mathcal{S}$, where $\lambda_0, \lambda_1, \cdots, \lambda_m$ are chosen such that the constraints in (2.200) are satisfied. Then p^* maximizes $H(p)$ over all probability distribution p on $\mathcal{S}$, subject to the constraints in (2.200).

Proof. For any p satisfying the constraints in (2.200), consider

$$\begin{aligned}
&H(p^*) - H(p) \\
&= -\sum_{x \in \mathcal{S}} p^*(x) \ln p^*(x) + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.202)\\
&= -\sum_{x \in \mathcal{S}} p^*(x) \left(-\lambda_0 - \sum_i \lambda_i r_i(x) \right) + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.203)\\
&= \lambda_0 \left(\sum_{x \in \mathcal{S}} p^*(x) \right) + \sum_i \lambda_i \left(\sum_{x \in \mathcal{S}} p^*(x) r_i(x) \right) + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.204)\\
&= \lambda_0 \cdot 1 + \sum_i \lambda_i a_i + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.205)\\
&= \lambda_0 \left(\sum_{x \in \mathcal{S}_p} p(x) \right) + \sum_i \lambda_i \left(\sum_{x \in \mathcal{S}_p} p(x) r_i(x) \right) + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.206)\\
&= -\sum_{x \in \mathcal{S}_p} p(x) \left(-\lambda_0 - \sum_i \lambda_i r_i(x) \right) + \sum_{x \in \mathcal{S}_p} p(x) \ln p(x) && (2.207)
\end{aligned}$$

$$= -\sum_{x\in\mathcal{S}_p} p(x)\ln p^*(x) + \sum_{x\in\mathcal{S}_p} p(x)\ln p(x) \tag{2.208}$$

$$= \sum_{x\in\mathcal{S}_p} p(x)\ln\frac{p(x)}{p^*(x)} \tag{2.209}$$

$$= D(p\|p^*) \tag{2.210}$$

$$\geq 0. \tag{2.211}$$

In the above, (2.207) is obtained from (2.203) by replacing $p^*(x)$ by $p(x)$ and $x \in \mathcal{S}$ by $x \in \mathcal{S}_p$ in the first summation, while the intermediate steps (2.204)–(2.206) are justified by noting that both p^* and p satisfy the constraints in (2.200). The last step is an application of the divergence inequality (Theorem 2.31). The proof is accomplished. □

Remark For all $x \in \mathcal{S}$, $p^*(x) > 0$, so that $\mathcal{S}_{p^*} = \mathcal{S}$.

The following corollary of Theorem 2.50 is rather subtle.

Corollary 2.51. *Let p^* be a probability distribution defined on $\mathcal{S}$ with*

$$p^*(x) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(x)} \tag{2.212}$$

for all $x \in \mathcal{S}$. Then p^* maximizes $H(p)$ over all probability distribution p defined on $\mathcal{S}$, subject to the constraints

$$\sum_{x\in\mathcal{S}_p} p(x)r_i(x) = \sum_{x\in\mathcal{S}} p^*(x)r_i(x) \quad \textit{for } 1 \leq i \leq m. \tag{2.213}$$

Example 2.52. Let $\mathcal{S}$ be finite and let the set of constraints in (2.200) be empty. Then

$$p^*(x) = e^{-\lambda_0}, \tag{2.214}$$

a constant that does not depend on x. Therefore, p^* is simply the uniform distribution over $\mathcal{S}$, i.e., $p^*(x) = |\mathcal{S}|^{-1}$ for all $x \in \mathcal{S}$. This is consistent with Theorem 2.43.

Example 2.53. Let $\mathcal{S} = \{0, 1, 2, \cdots\}$, and let the set of constraints in (2.200) be

$$\sum_x p(x)x = a, \tag{2.215}$$

where $a \geq 0$, i.e., the mean of the distribution p is fixed at some nonnegative value a. We now determine p^* using the prescription in Theorem 2.50. Let

$$q_i = e^{-\lambda_i} \tag{2.216}$$

for $i = 0, 1$. Then by (2.201),

$$p^*(x) = q_0 q_1^x. \tag{2.217}$$

Evidently, p^* is a geometric distribution, so that

$$q_0 = 1 - q_1. \tag{2.218}$$

Finally, we invoke the constraint (2.200) on p to obtain $q_1 = (a+1)^{-1}$. The details are omitted.

2.10 Entropy Rate of a Stationary Source

In the previous sections, we have discussed various properties of the entropy of a finite collection of random variables. In this section, we discuss the *entropy rate* of a discrete-time information source.

A discrete-time information source $\{X_k, k \geq 1\}$ is an infinite collection of random variables indexed by the set of positive integers. Since the index set is ordered, it is natural to regard the indices as time indices. We will refer to the random variables X_k as *letters*.

We assume that $H(X_k) < \infty$ for all k. Then for any finite subset A of the index set $\{k : k \geq 1\}$, we have

$$H(X_k, k \in A) \leq \sum_{k \in A} H(X_k) < \infty. \tag{2.219}$$

However, it is not meaningful to discuss $H(X_k, k \geq 1)$ because the joint entropy of an infinite collection of letters is infinite except for very special cases. On the other hand, since the indices are ordered, we can naturally define the *entropy rate* of an information source, which gives the average entropy per letter of the source.

Definition 2.54. *The entropy rate of an information source $\{X_k\}$ is defined as*

$$H_X = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) \tag{2.220}$$

when the limit exists.

We show in the next two examples that the entropy rate of a source may or may not exist.

Example 2.55. Let $\{X_k\}$ be an i.i.d. source with generic random variable X. Then

$$\lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) = \lim_{n \to \infty} \frac{nH(X)}{n} \tag{2.221}$$

$$= \lim_{n \to \infty} H(X) \tag{2.222}$$

$$= H(X), \tag{2.223}$$

i.e., the entropy rate of an i.i.d. source is the entropy of any of its single letters.

Example 2.56. Let $\{X_k\}$ be a source such that X_k are mutually independent and $H(X_k) = k$ for $k \geq 1$. Then

$$\frac{1}{n}H(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{k=1}^{n} k \tag{2.224}$$

$$= \frac{1}{n}\frac{n(n+1)}{2} \tag{2.225}$$

$$= \frac{1}{2}(n+1), \tag{2.226}$$

which does not converge as $n \to \infty$ although $H(X_k) < \infty$ for all k. Therefore, the entropy rate of $\{X_k\}$ does not exist.

Toward characterizing the asymptotic behavior of $\{X_k\}$, it is natural to consider the limit

$$H'_X = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}) \tag{2.227}$$

if it exists. The quantity $H(X_n|X_1, X_2, \cdots, X_{n-1})$ is interpreted as the conditional entropy of the next letter given that we know all the past history of the source, and H'_X is the limit of this quantity after the source has been run for an indefinite amount of time.

Definition 2.57. *An information source $\{X_k\}$ is stationary if*

$$X_1, X_2, \cdots, X_m \tag{2.228}$$

and

$$X_{1+l}, X_{2+l}, \cdots, X_{m+l} \tag{2.229}$$

have the same joint distribution for any $m, l \geq 1$.

In the rest of the section, we will show that stationarity is a sufficient condition for the existence of the entropy rate of an information source.

Lemma 2.58. *Let $\{X_k\}$ be a stationary source. Then H'_X exists.*

Proof. Since $H(X_n|X_1, X_2, \cdots, X_{n-1})$ is lower bounded by zero for all n, it suffices to prove that $H(X_n|X_1, X_2, \cdots, X_{n-1})$ is non-increasing in n to conclude that the limit H'_X exists. Toward this end, for $n \geq 2$, consider

$$H(X_n|X_1, X_2, \cdots, X_{n-1})$$

$$\leq H(X_n|X_2, X_3, \cdots, X_{n-1}) \tag{2.230}$$

$$= H(X_{n-1}|X_1, X_2, \cdots, X_{n-2}), \tag{2.231}$$

where the last step is justified by the stationarity of $\{X_k\}$. The lemma is proved. □

Lemma 2.59 (Cesáro Mean). *Let a_k and b_k be real numbers. If $a_n \to a$ as $n \to \infty$ and $b_n = \frac{1}{n}\sum_{k=1}^{n} a_k$, then $b_n \to a$ as $n \to \infty$.*

Proof. The idea of the lemma is the following. If $a_n \to a$ as $n \to \infty$, then the average of the first n terms in $\{a_k\}$, namely b_n, also tends to a as $n \to \infty$.

The lemma is formally proved as follows. Since $a_n \to a$ as $n \to \infty$, for every $\epsilon > 0$, there exists $N(\epsilon)$ such that $|a_n - a| < \epsilon$ for all $n > N(\epsilon)$. For $n > N(\epsilon)$, consider

$$|b_n - a| = \left| \frac{1}{n}\sum_{i=1}^{n} a_i - a \right| \tag{2.232}$$

$$= \left| \frac{1}{n}\sum_{i=1}^{n} (a_i - a) \right| \tag{2.233}$$

$$\le \frac{1}{n}\sum_{i=1}^{n} |a_i - a| \tag{2.234}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{N(\epsilon)} |a_i - a| + \sum_{i=N(\epsilon)+1}^{n} |a_i - a| \right) \tag{2.235}$$

$$< \frac{1}{n}\sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{(n - N(\epsilon))\epsilon}{n} \tag{2.236}$$

$$< \frac{1}{n}\sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon. \tag{2.237}$$

The first term tends to 0 as $n \to \infty$. Therefore, for any $\epsilon > 0$, by taking n to be sufficiently large, we can make $|b_n - a| < 2\epsilon$. Hence $b_n \to a$ as $n \to \infty$, proving the lemma. □

We now prove that H'_X is an alternative definition/interpretation of the entropy rate of $\{X_k\}$ when $\{X_k\}$ is stationary.

Theorem 2.60. *The entropy rate H_X of a stationary source $\{X_k\}$ exists and is equal to H'_X.*

Proof. Since we have proved in Lemma 2.58 that H'_X always exists for a stationary source $\{X_k\}$, in order to prove the theorem, we only have to prove that $H_X = H'_X$. By the chain rule for entropy,

$$\frac{1}{n} H(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{k=1}^{n} H(X_k | X_1, X_2, \cdots, X_{k-1}). \tag{2.238}$$

Since

$$\lim_{k \to \infty} H(X_k | X_1, X_2, \cdots, X_{k-1}) = H'_X \tag{2.239}$$

from (2.227), it follows from Lemma 2.59 that

$$H_X = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) = H'_X. \tag{2.240}$$

The theorem is proved. □

In this theorem, we have proved that the entropy rate of a random source $\{X_k\}$ exists under the fairly general assumption that $\{X_k\}$ is stationary. However, the entropy rate of a stationary source $\{X_k\}$ may not carry any physical meaning unless $\{X_k\}$ is also ergodic. This will be explained when we discuss the Shannon–McMillan–Breiman Theorem in Section 5.4.

Appendix 2.A: Approximation of Random Variables with Countably Infinite Alphabets by Truncation

Let X be a random variable with a countable alphabet $\mathcal{X}$ such that $H(X) < \infty$. Without loss of generality, $\mathcal{X}$ is taken to be the set of positive integers. Define a random variable $X(m)$ which takes values in

$$\mathcal{N}_m = \{1, 2, \cdots, m\} \tag{2.241}$$

such that

$$\Pr\{X(m) = k\} = \frac{\Pr\{X = k\}}{\Pr\{X \in \mathcal{N}_m\}} \tag{2.242}$$

for all $k \in \mathcal{N}_m$, i.e., the distribution of $X(m)$ is the truncation of the distribution of X up to m.

It is intuitively correct that $H(X(m)) \to H(X)$ as $m \to \infty$, which we formally prove in this appendix. For every $m \geq 1$, define the binary random variable

$$B(m) = \begin{cases} 1 \text{ if } X \leq m \\ 0 \text{ if } X > m \end{cases}. \tag{2.243}$$

Consider

$$\begin{aligned} H(X) = &- \sum_{k=1}^{m} \Pr\{X = k\} \log \Pr\{X = k\} \\ &- \sum_{k=m+1}^{\infty} \Pr\{X = k\} \log \Pr\{X = k\}. \end{aligned} \tag{2.244}$$

As $m \to \infty$,

$$- \sum_{k=1}^{m} \Pr\{X = k\} \log \Pr\{X = k\} \to H(X). \tag{2.245}$$

Since $H(X) < \infty$,

$$-\sum_{k=m+1}^{\infty} \Pr\{X=k\}\log\Pr\{X=k\} \to 0 \tag{2.246}$$

as $k \to \infty$. Now consider

$$\begin{aligned}
&H(X)\\
&= H(X|B(m)) + I(X;B(m)) &(2.247)\\
&= H(X|B(m)=1)\Pr\{B(m)=1\} + H(X|B(m)=0)\\
&\quad \times \Pr\{B(m)=0\} + I(X;B(m)) &(2.248)\\
&= H(X(m))\Pr\{B(m)=1\} + H(X|B(m)=0)\\
&\quad \times \Pr\{B(m)=0\} + I(X;B(m)). &(2.249)
\end{aligned}$$

As $m \to \infty$, $H(B(m)) \to 0$ since $\Pr\{B(m)=1\} \to 1$. This implies $I(X;B(m)) \to 0$ because

$$I(X;B(m)) \le H(B(m)). \tag{2.250}$$

In (2.249), we further consider

$$\begin{aligned}
&H(X|B(m)=0)\Pr\{B(m)=0\}\\
&= -\sum_{k=m+1}^{\infty} \Pr\{X=k\}\log\frac{\Pr\{X=k\}}{\Pr\{B(m)=0\}} &(2.251)\\
&= -\sum_{k=m+1}^{\infty} \Pr\{X=k\}(\log\Pr\{X=k\}\\
&\quad -\log\Pr\{B(m)=0\}) &(2.252)\\
&= -\sum_{k=m+1}^{\infty} (\Pr\{X=k\}\log\Pr\{X=k\})\\
&\quad +\left(\sum_{k=m+1}^{\infty} \Pr\{X=k\}\right)\log\Pr\{B(m)=0\} &(2.253)\\
&= -\sum_{k=m+1}^{\infty} \Pr\{X=k\}\log\Pr\{X=k\}\\
&\quad +\Pr\{B(m)=0\}\log\Pr\{B(m)=0\}. &(2.254)
\end{aligned}$$

As $m \to \infty$, the summation above tends to 0 by (2.246). Since $\Pr\{B(m)=0\} \to 0$, $\Pr\{B(m)=0\}\log\Pr\{B(m)=0\} \to 0$. Therefore,

$$H(X|B(m)=0)\Pr\{B(m)=0\} \to 0, \tag{2.255}$$

and we see from (2.249) that $H(X(m)) \to H(X)$ as $m \to \infty$.

Chapter Summary

Markov Chain: $X \to Y \to Z$ forms a Markov chain if and only if

$$p(x,y,z) = a(x,y)b(y,z)$$

for all x, y, and z such that $p(y) > 0$.

Shannon's Information Measures:

$$H(X) = -\sum_x p(x) \log p(x) = -E \log p(X),$$

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = E \log \frac{p(X,Y)}{p(X)p(Y)},$$

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -E \log p(Y|X),$$

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = E \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}.$$

Some Useful Identities:

$$\begin{aligned} H(X) &= I(X;X), \\ H(Y|X) &= H(X,Y) - H(X), \\ I(X;Y) &= H(X) - H(X|Y), \\ I(X;Y|Z) &= H(X|Z) - H(X|Y,Z). \end{aligned}$$

Chain Rule for Entropy:

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}).$$

Chain Rule for Mutual Information:

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1}).$$

Informational Divergence: For two probability distributions p and q on a common alphabet $\mathcal{X}$,

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}.$$

Fundamental Inequality: For any $a > 0$, $\ln a \le a - 1$, with equality if and only if $a = 1$.

Divergence Inequality: $D(p\|q) \geq 0$, with equality if and only if $p = q$.

Log-Sum Inequality: For positive numbers $a_1, a_2, \cdots$ and nonnegative numbers $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}.$$

Equality holds if and only if $\frac{a_i}{b_i} = \textit{constant}$ for all i.

The Basic Inequalities: All Shannon's information measures are nonnegative.

Some Useful Properties of Shannon's Information Measures:

1. $H(X) \leq \log |\mathcal{X}|$ with equality if and only if X is uniform.
2. $H(X) = 0$ if and only if X is deterministic.
3. $H(Y|X) = 0$ if and only if Y is a function of X.
4. $I(X;Y) = 0$ if and only X and Y are independent.

Fano's Inequality: Let X and $\hat{X}$ be random variables taking values in the same alphabet $\mathcal{X}$. Then

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1).$$

Conditioning Does Not Increase Entropy: $H(Y|X) \leq H(Y)$, with equality if and only if X and Y are independent.

Independence Bound for Entropy:

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if X_i, $i = 1, 2, \cdots, n$ are mutually independent.

Data Processing Theorem: If $U \to X \to Y \to V$ forms a Markov chain, then $I(U;V) \leq I(X;Y)$.

Maximum Entropy Distributions: Let

$$p^*(x) = e^{-\lambda_0 - \sum_{i=1}^{m} \lambda_i r_i(x)}$$

for all $x \in \mathcal{S}$, where $\lambda_0, \lambda_1, \cdots, \lambda_m$ are chosen such that the constraints

$$\sum_{x \in \mathcal{S}_p} p(x) r_i(x) = a_i \quad \text{for } 1 \leq i \leq m$$

are satisfied. Then p^* maximizes $H(p)$ over all probability distributions p on $\mathcal{S}$ subject to the above constraints.

Entropy Rate of a Stationary Source:

1. The entropy rate of an information source $\{X_k\}$ is defined as
$$H_X = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$
when the limit exists.
2. The entropy rate H_X of a stationary source $\{X_k\}$ exists and is equal to
$$H'_X = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}).$$

Problems

1. Let X and Y be random variables with alphabets $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4, 5\}$ and joint distribution $p(x, y)$ given by
$$\frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 1 & 1 \\ 0 & 3 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 3 \end{bmatrix}.$$
Calculate $H(X), H(Y), H(X|Y), H(Y|X)$, and $I(X;Y)$.
2. Prove Propositions 2.8, 2.9, 2.10, 2.19, 2.21, and 2.22.
3. Give an example which shows that pairwise independence does not imply mutual independence.
4. Verify that $p(x, y, z)$ as defined in Definition 2.4 is a probability distribution. You should exclude all the zero probability masses from the summation carefully.
5. *Linearity of expectation* It is well known that expectation is linear, i.e., $E[f(X) + g(Y)] = Ef(X) + Eg(Y)$, where the summation in an expectation is taken over the corresponding alphabet. However, we adopt in information theory the convention that the summation in an expectation is taken over the corresponding support. Justify carefully the linearity of expectation under this convention.
6. The identity $I(X;Y) = H(X) - H(X|Y)$ is invalid if $H(X|Y)$ (and hence $H(X)$) is equal to infinity. Give an example such that $I(X;Y)$ has a finite value but both $H(X)$ and $H(Y|X)$ are equal to infinity.
7. Let p'_{XY} and p_{XY} be probability distributions defined on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are fixed finite alphabets. Prove that
$$\lim_{p'_{XY}\to p_{XY}} p'_x = p_X,$$
where the limit is taken with respect to the variational distance.

8. Let p_k and p be probability distributions defined on a common finite alphabet. Show that as $k \to \infty$, if $p_k \to p$ in variational distance, then $p_k \to p$ in $\mathcal{L}^2$ and vice versa.
9. Consider any probability distribution $p(x, y, z)$ and let

$$q(x, y, z) = \begin{cases} p(x)p(y)p(z|x, y) & \text{if } p(x, y) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

   a) Show that $q(x, y, z)$ is in general not a probability distribution.
   b) By ignoring the fact that $q(x, y, z)$ may not be a probability distribution, application of the divergence inequality $D(p\|q) \geq 0$ would yield the inequality

$$H(X) + H(Y) + H(Z|X, Y) \geq H(X, Y, Z),$$

   which indeed holds for all jointly distributed random variables X, Y, and Z. Explain.
10. Let $C_\alpha = \sum_{n=2}^{\infty} \frac{1}{n(\log n)^\alpha}$.
   a) Prove that

$$C_\alpha \begin{cases} < \infty \text{ if } \alpha > 1 \\ = \infty \text{ if } 0 \leq \alpha \leq 1 \end{cases}.$$

   Then

$$p_\alpha(n) = [C_\alpha n(\log n)^\alpha]^{-1}, \quad n = 2, 3, \cdots$$

   is a probability distribution for $\alpha > 1$.
   b) Prove that

$$H(p_\alpha) \begin{cases} < \infty \text{ if } \alpha > 2 \\ = \infty \text{ if } 1 < \alpha \leq 2 \end{cases}.$$

11. Prove that $H(p)$ is concave in p, i.e., for $0 \leq \lambda \leq 1$ and $\bar{\lambda} = 1 - \lambda$,

$$\lambda H(p_1) + \bar{\lambda} H(p_2) \leq H(\lambda p_1 + \bar{\lambda} p_2).$$

12. Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$.
   a) Prove that for fixed $p(x)$, $I(X; Y)$ is a convex functional of $p(y|x)$.
   b) Prove that for fixed $p(y|x)$, $I(X; Y)$ is a concave functional of $p(x)$.
13. Do $I(X; Y) = 0$ and $I(X; Y|Z) = 0$ imply each other? If so, give a proof. If not, give a counterexample.
14. Give an example for which $D(\cdot\|\cdot)$ does not satisfy the triangular inequality.
15. Let X be a function of Y. Prove that $H(X) \leq H(Y)$. Interpret this result.
16. Prove that for any $n \geq 2$,

$$H(X_1, X_2, \cdots, X_n) \geq \sum_{i=1}^{n} H(X_i|X_j, j \neq i).$$

17. Prove that

$$H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3) \geq 2H(X_1, X_2, X_3).$$

Hint: Sum the identities

$$H(X_1, X_2, X_3) = H(X_j, j \neq i) + H(X_i | X_j, j \neq i)$$

for $i = 1, 2, 3$ and apply the result in Problem 16.

18. For a subset α of $\mathcal{N}_n = \{1, 2, \cdots, n\}$, denote $(X_i, i \in \alpha)$ by X_α. For $1 \leq k \leq n$, let

$$H_k = \frac{1}{\binom{n}{k}} \sum_{\alpha : |\alpha| = k} \frac{H(X_\alpha)}{k}.$$

Here H_k is interpreted as the average entropy per random variable when k random variables are taken from $X_1, X_2, \cdots, X_n$ at a time. Prove that

$$H_1 \geq H_2 \geq \cdots \geq H_n.$$

This sequence of inequalities, due to Han [147], is a generalization of the independence bound for entropy (Theorem 2.39). See Problem 6 in Chapter 21 for an application of these inequalities.

19. For a subset α of $\mathcal{N}_n = \{1, 2, \cdots, n\}$, let $\overline{\alpha} = \mathcal{N}_n \backslash \alpha$ and denote $(X_i, i \in \alpha)$ by X_α. For $1 \leq k \leq n$, let

$$H'_k = \frac{1}{\binom{n}{k}} \sum_{\alpha : |\alpha| = k} \frac{H(X_\alpha | X_{\overline{\alpha}})}{k}.$$

Prove that

$$H'_1 \leq H'_2 \leq \cdots \leq H'_n.$$

Note that H'_n is equal to H_n in the last problem. This sequence of inequalities is again due to Han [147]. See Yeung and Cai [406] for an application of these inequalities.

20. Prove the divergence inequality by using the log-sum inequality.
21. Prove that $D(p\|q)$ is convex in the pair (p, q), i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions on a common alphabet, then

$$D(\lambda p_1 + \overline{\lambda} p_2 \| \lambda q_1 + \overline{\lambda} q_2) \leq \lambda D(p_1 \| q_1) + \overline{\lambda} D(p_2 \| q_2)$$

for all $0 \leq \lambda \leq 1$, where $\overline{\lambda} = 1 - \lambda$.

22. Let p_{XY} and q_{XY} be two probability distributions on $\mathcal{X} \times \mathcal{Y}$. Prove that $D(p_{XY} \| q_{XY}) \geq D(p_X \| q_X)$.
23. *Pinsker's inequality.* Let $V(p, q)$ denote the variational distance between two probability distributions p and q on a common alphabet $\mathcal{X}$. We will determine the largest c which satisfies

$$D(p\|q) \geq cd^2(p, q).$$

a) Let $A = \{x : p(x) \geq q(x)\}$, $\hat{p} = \{p(A), 1 - p(A)\}$, and $\hat{q} = \{q(A), 1 - q(A)\}$. Show that $D(p\|q) \geq D(\hat{p}\|\hat{q})$ and $V(p, q) = V(\hat{p}, \hat{q})$.

b) Show that toward determining the largest value of c, we only have to consider the case when $\mathcal{X}$ is binary.

c) By virtue of (b), it suffices to determine the largest c such that

$$p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} - 4c(p - q)^2 \geq 0$$

for all $0 \leq p, q \leq 1$, with the convention that $0 \log \frac{0}{b} = 0$ for $b \geq 0$ and $a \log \frac{a}{0} = \infty$ for $a > 0$. By observing that equality in the above holds if $p = q$ and considering the derivative of the left-hand side with respect to q, show that the largest value of c is equal to $(2 \ln 2)^{-1}$.

24. Let p and $q_k, k \geq 1$ be probability distributions on a common alphabet. Show that if q_k converges to p in divergence, then it also converges to p in variational distance.

25. Find a necessary and sufficient condition for Fano's inequality to be tight.

26. Determine the probability distribution defined on $\{0, 1, \cdots, n\}$ that maximizes the entropy subject to the constraint that the mean is equal to m, where $0 \leq m \leq n$.

27. Show that for a stationary source $\{X_k\}$, $\frac{1}{n} H(X_1, X_2, \cdots, X_n)$ is non-increasing in n.

28. For real numbers $\alpha > 1$ and $\beta > 0$ and an integer $n \geq \alpha$, define the probability distribution

$$\mathcal{D}_n^{(\alpha,\beta)} = \left\{ 1 - \left(\frac{\log \alpha}{\log n} \right)^{\beta}, \underbrace{\frac{1}{n} \left(\frac{\log \alpha}{\log n} \right)^{\beta}, \cdots, \frac{1}{n} \left(\frac{\log \alpha}{\log n} \right)^{\beta}}_{n}, 0, 0, \cdots \right\}.$$

Let $\nu = \{1, 0, 0, \ldots\}$ be the deterministic distribution.

a) Show that $\lim_{n \to \infty} D\left(\nu || \mathcal{D}_n^{(\alpha,\beta)}\right) = 0$.

b) Determine $\lim_{n \to \infty} H\left(\mathcal{D}_n^{(\alpha,\beta)}\right)$.

29. *Discontinuity of entropy with respect to convergence in divergence.* Let $\mathcal{P}$ be the set of all probability distributions on a countable alphabet. A function $f : \mathcal{P} \to \Re$ is continuous with respect to convergence in divergence at $P \in \mathcal{P}$ if for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(P) - f(Q)| < \epsilon$ for all $Q \in \mathcal{P}$ satisfying $D(P\|Q) < \delta$; otherwise, f is discontinuous at P.

a) Let $H : \mathcal{P} \to \Re$ be the entropy function. Show that H is discontinuous at the deterministic distribution $\nu = \{1, 0, 0, \cdots, \}$. Hint: Use the results in Problem 28.

b) Show that H is discontinuous at $P = \{p_0, p_1, p_2, \cdots\}$ for all P such that $H(P) < \infty$. Hint: Consider the probability distribution

$$Q_n = \Big\{ p_0 - \frac{p_0}{\sqrt{\log n}}, p_1 + \frac{p_0}{n\sqrt{\log n}}, p_2 + \frac{p_0}{n\sqrt{\log n}}, \cdots,$$
$$p_n + \frac{p_0}{n\sqrt{\log n}}, p_{n+1}, p_{n+2}, \ldots \Big\}$$

for large n.

30. *Discontinuity of entropy with respect to convergence in variational distance.* Refer to Problem 29. The continuity of a function $f : \mathcal{P} \to \Re$ with respect to convergence in variational distance can be defined similarly.
    a) Show that if a function f is continuous with respect to convergence in variational distance, then it is also continuous with respect to convergence in divergence. Hint: Use Pinsker's inequality.
    b) Repeat (b) in Problem 29 with continuity defined with respect to convergence in variational distance.
31. *Continuity of the entropy function for a fixed finite alphabet.* Refer to Problems 29 and 30. Suppose the domain of H is confined to $\mathcal{P}'$, the set of all probability distributions on a fixed finite alphabet. Show that H is continuous with respect to convergence in divergence.
32. Let $\mathbf{p} = \{p_1, p_2, \cdots, p_n\}$ and $\mathbf{q} = \{q_1, q_2, \cdots, q_n\}$ be two sets of real numbers such that $p_i \geq p_{i'}$ and $q_i \geq q_{i'}$ for all $i < i'$. We say that $\mathbf{p}$ is *majorized* by $\mathbf{q}$ if $\sum_{i=1}^{m} p_i \leq \sum_{j=1}^{m} q_j$ for all $m = 1, 2, \ldots, n$, where equality holds when $m = n$. A function $f : \Re^n \to \Re$ is *Schur-concave* if $f(\mathbf{p}) \geq f(\mathbf{q})$ whenever $\mathbf{p}$ is majorized by $\mathbf{q}$. Now let $\mathbf{p}$ and $\mathbf{q}$ be probability distributions. We will show in the following steps that $H(\cdot)$ is Schur-concave.
    a) Show that for $\mathbf{p} \neq \mathbf{q}$, there exist $1 \leq j < k \leq n$ which satisfy the following:
        i) j is the largest index i such that $p_i < q_i$.
        ii) k is the smallest index i such that $i > j$ and $p_i > q_i$.
        iii) $p_i = q_i$ for all $j < i < k$.
    b) Consider the distribution $\mathbf{q}^* = \{q_1^*, q_2^*, \cdots, q_n^*\}$ defined by $q_i^* = q_i$ for $i \neq j, k$ and

$$(q_j^*, q_k^*) = \begin{cases} (p_j, q_k + (q_j - p_j)) & \text{if } p_k - q_k \geq q_j - p_j \\ (q_j - (p_k - q_k), p_k) & \text{if } p_k - q_k < q_j - p_j \end{cases}.$$

    Note that either $q_j^* = p_j$ or $q_k^* = p_k$. Show that
        i) $q_i^* \geq q_{i'}^*$ for all $i \leq i'$.
        ii) $\sum_{i=1}^{m} p_i \leq \sum_{i=1}^{m} q_i^*$ for all $m = 1, 2, \cdots, n$.
        iii) $H(\mathbf{q}^*) \geq H(\mathbf{q})$.
    c) Prove that $H(\mathbf{p}) \geq H(\mathbf{q})$ by induction on the Hamming distance between $\mathbf{p}$ and $\mathbf{q}$, i.e., the number of places where $\mathbf{p}$ and $\mathbf{q}$ differ.

In general, if a concave function f is symmetric, i.e., $f(\mathbf{p}) = f(\mathbf{p}')$ where $\mathbf{p}'$ is a permutation of $\mathbf{p}$, then f is Schur-concave. We refer the reader to [246] for the theory of majorization. (Hardy, Littlewood, and Pólya [154].)

Historical Notes

The concept of entropy has its root in thermodynamics. Shannon [322] was the first to use entropy as a measure of information. Informational divergence was introduced by Kullback and Leibler [214], and it has been studied extensively by Csiszár [80] and Amari [14].

Most of the materials in this chapter can be found in standard textbooks in information theory. The main concepts and results are due to Shannon [322]. Pinsker's inequality is due to Pinsker [292]. Fano's inequality has its origin in the converse proof of the channel coding theorem (to be discussed in Chapter 7) by Fano [107]. Generalizations of Fano's inequality which apply to random variables with countable alphabets have been obtained by Han and Verdú [153] and by Ho [165] (see also [168]). Maximum entropy, a concept in statistical mechanics, was expounded in Jaynes [186].