(Hidden) Markov Processes: Theory and Applications to Biology

M. Vidyasagar

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Dedication goes here.

Contents

Preface

Every author aspiring to write a book should address two fundamental questions: (i) Who is the targeted audience, and (ii) what does he wish to say to them? In the world of literature, it is often said that every novel is autobiographical to some extent or another. Adapting this maxim to the current situation, I would say that every book I have ever written has been addressed to a reader who is in the situation in which I found myself before I embarked on the book-writing project. To put it another way, every book I have written has been an attempt to make it possible for my readers to circumvent some of the difficulties that I myself faced when learning a subject that was new to me.

In the present instance, for the past few years I have been interested in the broad area of computational biology. With the explosion in the sheer quantity of biological data, and an enhanced understanding of the fundamental mechanisms of genomics and proteomics, there is now greater interest than ever in this topic (computational biology). I got very interested in hidden Markov processes (HMPs) when I realized that several researchers in computational biology were applying HMPs to address various prediction and classification problems in genomics and proteomics. Thus, after virtually an entire research career spent in blissful ignorance of all matters stochastic, I got down to try and learn something about Markov processes and HMPs. At the same time, I was trying to learn enough about basic biology, and also to read the existing literature in the area of Markov and hidden Markov methods in computational biology.

I was faced with two sets of difficulties in this endeavour, one with the Markov process literature and another with the computational biology literature. In the paragraphs to follow, I will describe first the scope of the book; then I will describe the difficulties I faced, and how I hope to alleviate them in this book. My fond hope is that by reading this book others will have an easier time learning these topics than I did myself.

Hidden Markov processes (HMPs) were introduced into the statistics literature as far back as 1966 [13]. Starting in the mid 1970's [9, 10], HMPs have been used in speech recognition, which is perhaps the earliest *application* of HMPs in a non-mathematical context. The paper [43] contains a wonderful survey of most of the relevant theory of HMPs. In recent years, HMPs have also been used in problems of computational biology, such as finding genes from the genome (DNA sequence) of an organism [74, 30], or classifying proteins into one of several families [73]. Markov models underlie some of the significant advances in sequence alignment such as the BLAST algorithm and its variants [65, 4, 66, 5], and popular algorithms for finding genes from a genome, as exemplified by GENSCAN [23] and GLIMMER and its extensions [30]. Accordingly, the current book contains two distinct themes: (i) the theory of Markov processes and hidden Markov processes, and (ii) the application of this theory to some problems in computational biology.

Now let me describe the difficulties I found with the existing books on Markov processes. These books invariably focus on processes with *infinite* state spaces. Books such as [69, 103] restrict themselves to Markov processes with *countable* state spaces, since in this case many of the technicalities associated with uncountable state spaces disappear. From a mathematician's standpoint, the case of a finite state space is not worth expounding separately, since the extension from a finite state space to a countably infinite state space usually "comes for free." However, even the "simplified" theory as in [103] is inaccessible to many if not most engineers, and certainly to most biologists. Such readers can handle Markov processes with *finite* state spaces but nothing more, because they can understand matrices, eigenvalues, eigenvectors and the like, but in general cannot follow more advanced topics. At the same time, books on Markov processes with *finite* state spaces seldom go beyond computing stationary distributions, and almost completely ignore advanced topics such as ergodicity, mixing, parameter estimation, and the like, that are vital in any application of the theory. A notable exception is [99], which at least talks about ergodicity, but does not discuss mixing properties of Markov chains, nor parameter estimation. For purposes such as analyzing the statistical significance of an inferencing or a modelling algorithm, mere ergodicity is too weak a property, and some form of mixing is required.

Thus the current situation with respect to books on Markov processes can be summarized as follows: There is no treatment of "advanced" notions using only "elementary" techniques, and in an "elementary" setting. In contrast, in the present book the focus is almost exclusively on stochastic processes assuming values in a finite set, so that technicalities are kept to an absolute minimum. By restricting attention to Markov chains with finite state spaces, I am able to capture most of the interesting phenomena such as ergodicity and mixing, while giving elementary proofs that are accessible to anyone who knows undergraduate level linear algebra.

In the area of HMPs, much of the existing material is dedicated to the computation of various likelihoods (such as the most likely state trajectory corresponding to a given observation sequence), or to the determination of the most likely parameter set for a hidden Markov model *of a given, prespecified order*. In contrast, very little attention seems to have been paid to realization theory, that is, *determining the order of a hidden Markov model* on the basis of a set of observations. To me it appears practically a tautology to declare that the complexity of the model should be based on the observations, and not fixed *a priori* just to make a few problems tractable.

And yet realization theory is not given the same importance in the HMP literature as likelihood computation. In the present book, I have attempted to remedy the situation by including a thorough discussion of realization theory. Both "partial" realization and "complete" realization problems are studied. I not only give solutions to each, but also address issues such as the accuracy and confidence one has in the estimated parameters of the hidden Markov model.

The difficulty I faced with the existing literature in computational biology can now be described. At present there are several engineers and mathematicians who would like to study problems in computational biology and suggest suitable algorithms. There are of course some obvious difficulties, such as the need to learn (I am tempted to say "memorize") a great deal of unfamiliar terminology. Mathematicians are accustomed to a "reductionist" approach to their subject whereby everything follows from a few simply stated axioms. Such persons are handicapped by the huge differences in the styles of exposition between the engineering/mathematics community on the one hand and the biology community on the other hand. Even in the most "theoretical" biology journals, usually an algorithm is *not* described to the same level of depth or detail as it would be in the engineering or mathematics literature. Thus a person who wishes to understand statistical algorithms for gene prediction for example discovers very quickly that there is no place where these algorithms are discussed with sufficient detail for him/her to suggest improvements. The main purpose of the book is to address such a requirement.

Computational biology is a vast subject, and is constantly evolving. In choosing topics from computational biology for inclusion in the book, I restricted myself to genomics and proteomics, as these are perhaps the two aspects of biology that are the most "reductionist" in the sense described above. Even within genomics and proteomics, I have restricted myself to those algorithms that have a close connection with the Markov and HMP theory described here. Thus I have omitted any discussion of, to cite just one example, neural network-based methods. Readers wishing to find an encyclopaedic treatment of many aspects of computational biology are referred to [12, 46].

The chapter on biological applications begins with a basic introduction to biology, that is of necessity overly simplified. A "true" biologist will, in all probability, find much to quarrel with in this section. But it serves the immediate purpose, namely, to formulate some problems in genomics and proteomics in statistical terms. A reader who is eager to understand the relationship between the theory presented in the early chapters and its application to biological situations can jump ahead straightaway to Sections 11.1 and 11.2.

There is a fundamental difference between HMPs as they are used in speech recognition and HMPs as they are used in biology, namely: the length of the sample paths. In speech recognition, the sample paths are perhaps a dozen symbols long. As a result, methods such as Viterbi decoding can be effectively applied. In biological problems, the sample paths are several hundred symbols long. As a consequence, some of the methods used in speech recognition will not work at all in biological problems. Instead, one uses multi-step Markov processes (which are not truly "hidden" Markov processes). Hence issues such as parameter estimation become simpler in biological applications than in speech recognition, but issues such as estimating the likelihood of misclassification become more difficult. Nevertheless, in the book I will discuss all of the techniques used by either type of HMPs, so as to broaden the appeal of the book.

I hope that the book would not only assist biologists and other users of the theory to gain a better understanding of the methods they use, but would also spur the engineering and statistics research community to study some new and interesting research problems.

Chapter One

Introduction to Computational Biology

It would be more accurate to name this chapter 'A view of biology from the perspective of a computationalist.' The chapter represents my attempt to put forward a *simplified view* of the basic issues in genomics and proteomics, which are perhaps the two aspects of biology that lend themselves most readily to a 'reductionist' approach that is very familiar to mathematicians. Mathematically trained individuals prefer their problems to be stated simply and precisely. It does not matter how *difficult* the problem is, but it should be abundantly clear *what the problem is*, so that the community can be sure that they are all talking about the same problem. To persons with a mathematical training, biology appears to be a bewildering array of terminology (often unpronouncable by outsiders), and conventions. This chapter therefore represents an attempt to simplify the subject for the benefit of those who wish to understand what the basic issues are, at least from a computational standpoint, and then move on to tackle some of the outstanding problems.

Because of the simplification involved, some of what is written here may possibly raise the hackles of biologists. For instance, later on I will say that the only difference between DNA and RNA is that the letter T gets replaced by the letter U. Biologically speaking of course, this statement is potentially misleading. T (Thymine) has a different chemical structure and biological function from U (Uracil). But for the intended readers of this book, the key point is that if the DNA is viewed as a long string over the four symbol alphabet $\{A, C, G, T\}$ (as explained below), then the corresponding RNA is obtained simply by substituting the symbol U for the symbol T wherever the latter occurs. Thus, *from this standpoint*, there is no difference between T and U. Thus readers are advised to keep this point in mind, and not treat everything said here as being 'biologically true.' In two of my earlier books, I studied a specific form of machine learning known as 'probably approximately correct' (or PAC) learning. I could perhaps borrow that phrase and say that all the biological statements in this chapter are 'probably approximately correct.' However, all of the specific problem statements given here are 'computationally true' – there is no lack of precision either in the problem formulations or in the solutions put forward.

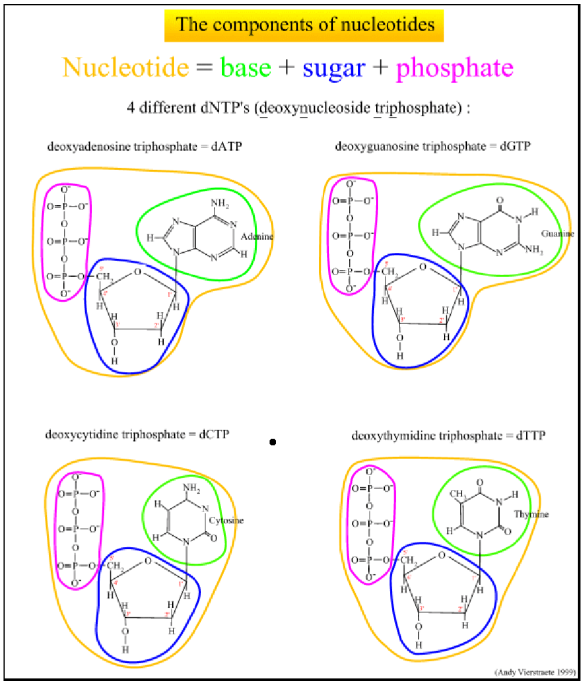

The components of nucleotides
Nucleotide = base + sugar + phosphate
4 different dNTP's (deoxynucleoside triphosphate) :
deoxyadenosine triphosphate = dATP
deoxyguanosine triphosphate = dGTP
deoxycytidine triphosphate = dCTP
deoxythymidine triphosphate = dTTP
Adenine
Guanine
Cytosine
Thymine
(Andy Vierstraete 1999)

Figure 1.1 The Four Nucleotides

1.1 THE GENOME

The genetic material in all living things is DNA, or Deoxyribonucleic acid. DNA is an enormously complex molecule built up out of just four building blocks, known as nucleotides. The nucleotides were discovered by Phoebus Levene in 1929. The four nucleotides share a common backbone, but contain different nucleic acids, or bases. For this reason, the four nucleotides are distinguished by the bases they contain, namely: Adenine, Cytosine, Guanine and Thymine. It is customary to denote the four nucleotides by the initial letters of the bases they contain, as A, C, G, T. There is no particular preferred order for listing the nucleotides, so they are listed here in the English alphabetical order. The diagram below[1] shows the four nucleotides including both the backbone and the base.

Because all four nucleotides share a common backbone, any nucleotide can ‘plug into’ any other nucleotide, like a lego toy. The so-called 3’ end of one nucleotide forms a very strong covalent chemical bond with the 5’ end of the next nucleotide. Thus it is possible to assemble the nucleotides in any order we wish. Such strings of nucleotides, assembled as per our specifications, are referred to as a strand of ‘oligonucleotides.’ While it has been known for

[1] Original sources for all diagrams are given in the References.

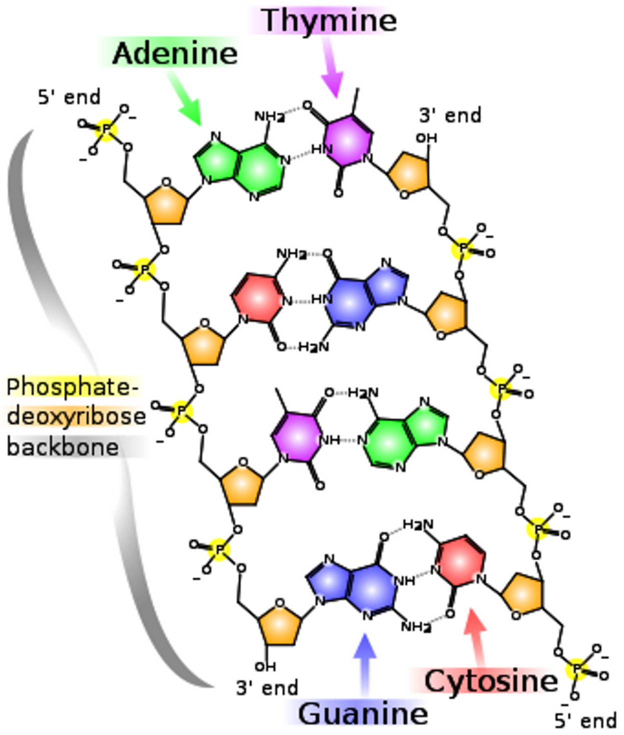
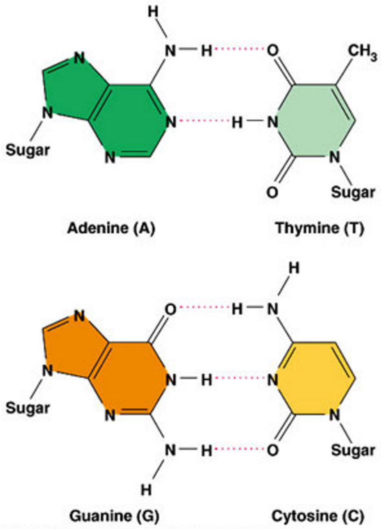


Figure 1.2 A DNA Fragment Showing Reverse Complimentarity

decades that it is possible to produce 'arbitrary' sequences of nucleotides, it is only within the past few years that it has become commercially feasible to produce oligonucleotide sequences on demand, thus leading to an entirely new field known as 'synthetic biology.'

The 5' end is deemed to be the start of the strand, and the 3' end is deemed to be the end of the strand. The DNA molecule consists of two strands of oligonucleotides that run in opposite directions. In addition to the very strong covalent bond between successive nucleotides on the same strand, the DNA molecule also has much weaker hydrogen bonds between nucleotides on opposite strands, which is a property known as 'reverse complementarity.' Thus if one strand contains A, then its counterpart on the opposite side must be T. Similarly, if one strand contains C, then its counterpart on the opposite must be G. Figure 1.1 below depicts the reverse complemetarity property. In it one can see both the covalent bond between adjacent nucleotides on the same strand, as well as the hydrogen bonds across strands. Figure 1.1 depicts the hydrogen bonds between A and T, and between C and G. It can be seen from these figures that the $A \leftrightarrow T$ bond is somewhat weaker than the $C \leftrightarrow G$ bond, because the former consists of two hydrogen atoms



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 1.3 Hydrogen Bonds Between A and T, and between C and G

bonding, while the latter consists of three hydrogen atoms bonding.

The DNA molecule is shaped like a double helix as shown in Figure 1.1. The discovery of the shape of the DNA molecule and the chemical bonds that hold it together (the horizontal beads in Figure 1.3) was made by James Watson and Francis Crick in 1953. Determining the structure of DNA was considered to be a very important problem at that time, and several persons were working on it, including the famous chemist Linus Pauling. Experimental evidence for the helical structure of DNA was obtained by Rosalind Franklin based on x-ray diffraction technique, though she (according to some reports) did not readily believe in the helical structure of DNA. Maurice Wilkins was a colleague of hers at King's College in London who was at the time trying to construct models based on available experimental evidence. Watson, Crick and Wilkes shared the Nobel Prize in Medicine in 1962 for their discoveries. Unfortunately Rosalind Franklin had passed away by that time, so one can only speculate as to whether she too would have received recognition in some form for the discovery. For Watson's own account of the discovery of the double helix, see [115]. The other two winners have also written their own version of events [27, 118], in the case of Wilkins 50 years after the event. For a counter-viewpoint that claims that Rosalind Franklin deserves far more credit than she has received, see [98]. In some 'lesser' organisms such as viruses and bacterial phages, the DNA molecules folds back on itself, but still retains the double helix structure.

Each cell within a living organism contains a copy of the DNA. So for example there are about 10^{11} cells in the human body, which implies that

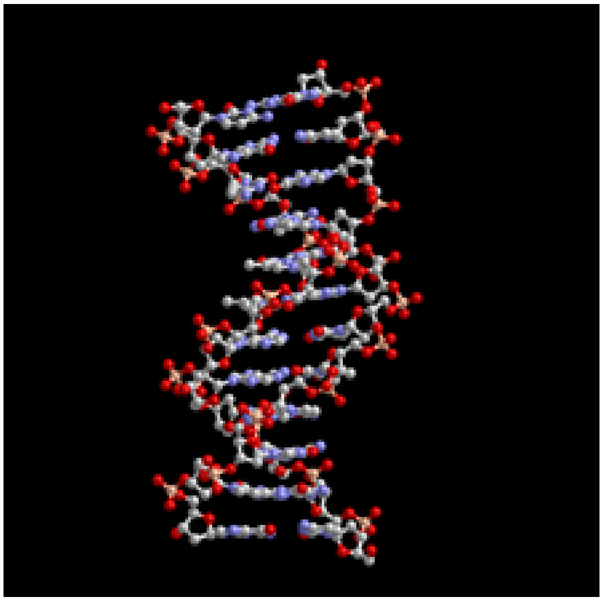


Figure 1.4 Double Helix Structure of DNA

there are 10^{11} copies of the DNA in the body. The DNA is an enormously complex molecule. As discussed in greater detail below, the human DNA consists of roughly *3.3 billion* nucleotides in each strand of the double helix. If the two strands of the double helix were to be separated and one strand were to be stretched out, the total length would be about *three meters*! And yet the chemical interactions are so strong that the DNA is very tightly wound up within itself, and each of the 10^{11} or so cells within the human body contains a copy of this three meter-long molecule. It is clear that, while one dimension of the DNA is very high, the other two dimensions are very tiny, which is why this compactification is possible.[2]

It is important to understand that each strand of the double-helix DNA has a definite spatial direction. The starting point for each strand is called the 5' end, while the ending point is called the 3' end. If we think of a strand of DNA as a 'tape,' then there is only one way to read the tape – it cannot be 'read backwards.' Because of the definite spatial direction, expressions such as 'the previous nucleotide,' 'the next nucleotide,' 'upstream' and 'downstream' are completely unambiguous. This feature permits us to model the *spatial orientation* by a *temporal orientation* and use modelling methods based on time series. If the spatial orientation were to be arbitrary, we could not do this.

[2] The following purely hyperbolic statement brings out this point more powerfully: If all the DNA molecules within the human body were to be stretched out and placed end-to-end, they would reach from the earth to the sun!

As stated above, the two strands of the helix run in opposite directions. Thus the 5' end of one strand is opposite the 3' end of the other strand. Moreover, again as stated above, in order for the two strands to have a series of hydrogen bonds, the two sides must satisfy the 'reverse complementarity' property. Thus, if one strand contains A (Adenine) in one location, the other side must contain T (Thyamine). Similarly, C (Cytosine) and G (Guanine) occur opposite each other.

The 'genome' of an organism is just a listing out, symbol by symbol, of the sequence of nucleotides that makes up one strand of the DNA. Because of reverse complementarity, if we know the listing of one strand, we know unambiguously the listing of the other strand. Since DNA occurs in two strands and the bases in each strand must 'pair up' according to reverse complementarity, the length of a genome is specified in 'base pairs.' The typical length of the genome varies depending upon the nature and complexity of the organism. Viruses, which cannot survive on their own but need a host in order to replicate, typically have genomes that several thousand base pairs long. Bacteria, which are the simplest self-sustaining life forms (meaning that they can reproduce on their own without a host, in contrast to viruses) have genomes that are a few million base pairs long. The mosquito has a genome that is about 300 million base pairs, the mouse genome is about 2.4 billion base pairs, the human genome is about 3.3 billion base pairs, and finally, the rice genome has about 10 billion base pairs! That last statistic, namely that the rice genome is three times as long as the human genome, ought to dispel the idea that the length of the genome is somehow monotonically related to the 'intelligence' of the organism.

The determination of the genome of organisms is one of the great triumphs of experimental biology, because the genome is one of the most 'unambiguous' representations of a life-form. See the first chapter of [88] for an excellent summary of the experimental methods and computational algorithms involved in 'sequencing' and 'assembling' a genome, that is, determing the string of symbols that comprise the genome. Moreover, it is noteworthy that the genome is not an 'analog' representation of life, but is a 'digital' representation, in the sense that the symbols at each location in the genome can have only a finite number of possible values (four). In general, the genomes of two exemplars of a species will have the same length. However, in the case of organisms such as the HIV virus which reproduces itself very sloppily, this statement is not always true. And of course, this statement is not valid at all when organisms have been experimentally modified in a laboratory. However, the genomes of two exemplars of an organism need not be *identical.* If the genome of an organism is 100% reproduced to create another organism, the second one is called a 'clone' of the first.

A draft of the human genome was determined and published simultaneously in February 2001 by two groups: The International Human Genome Research Consortium (IHGSC) [60] and Celera Genomics [106], a private company that subsequently went out of business. The human genome is about 3.3 billion base pairs long. The exact length is not known precisely, and experiments are still under way to refine the draft genome further.

In spite of its enormous length, it appears that there is a great deal of redundancy in the human genome. The overlap between the human genome and the mouse genome is about 80%, whereas between a human and a chimpanzee (our nearest neighbour in the animal kingdom) is about 98%. Between two humans the overlap is still more striking. It is estimated that the genome sequences of two humans will agree in about 99.9% of the locations, and differ only in about 0.1%, or about 3 million base pairs. All that distinguishes one human from another, be it height, weight, colour of eyes, colour of hair, etc. can presumably be attributed to this tiny variation in the genome (aside from environmental factors of course). These variations from the 'consensus' human genome (see below) are called Single Nucleotide Polymorphisms (SNPs), often pronounced as 'snips'. Even 'identical' twins will not have identifical genomes; rather, the overlap in such a case will be about 99.99%, as opposed to 99.9% in the case of two unrelated humans. A 100% replication of a genome is known as a 'clone' as stated earlier. The cloning of life forms is both a fascinating as well as a controversial subject.

It is widely accepted that there is indeed a 'consensus' human genome. In other words, it is believed that at any given location, an overwhelming majority of humans will have just one of the four nucleotides. Moreover, in case there is a deviation from the consensus genome, even though there are three variations possible in theory, in reality only one variation seems to occur in an overwhelming majority of cases. It is *not* the case, for example, that at a particular location about half of the population will have a T, another half will have a G, while A or C occur in a tiny fraction of the population. It is also not the case that virtually all humans have, say, a T at a particular location, whereas the symbols A, C, G occur among the remaining small minority with roughly equal frequency. If at all there is a variation, *only one* of the remaining three symbols will occur in almost all the rest of humanity. To contrast the situation, there is no 'consensus blood group' for example. While the O type blood is the most common, the percentage of the population that has other blood types is still significant. The draft human genome published in February 2001 by Celera is actually the (approximate) sequence of the DNA of no fewer than six different individuals, not that of just one person. Since the estimated error in the published draft (2% or so) is considerably more than the variations amongst individuals (0.1% or so, as mentioned above), this mixing up of DNA from different individuals did not matter. The above remarks about the existence of a consensus genome apply also to other organisms.

As technology improves, we can aspire to a situation whereby it will be both quick and inexpensive to determine the genome of every human on the planet, or at least a large number of them. As stated above, it would be wasteful to capture the DNA of a specific individual and sequence that. It would be more efficient to (i) determine the consensus genome very accurately, and (ii) determine the SNP's of an individual, that is, variations from the consensus genome. It appears reasonable that Step (ii) above should be, at least in principle, less expensive than an *ab initio* sequencing of the entire genome of an individual, since the variations from the consensus genome are expected in only 0.1% of the locations. The SNP profile of an individual is also known as the 'genotype.'

One of the most exciting challenges in computational biology is correlating an individual's genotype with his/her 'phenotype,' for example, the person's propensity to disease, responsiveness to a drug or treatment regime, or even potential adverse reactions to a drug. The current status is that in *some very specific situations*, we know that a particular SNP causes a specific disorder. Among the very first disorders to be tied unambiguously to a specific SNP is Sickle cell anemia, which causes red blood cells to be shaped like a crescent (or a sickle), as opposed to the round shape of a normal red blood cell. It was discovered that sickle cell anemia is caused by just one substitution: The codon GAG that codes for glutamine gets replaced by GTG, which codes for valine.[3] Thus *exactly one SNP* at just the right place in the genome can cause sickle cell anemia. Other examples of genetic disorders are Huntington Chorea and cystic fibrosis. Cystic fibrosis is in many ways well-suited for study using computational techniques, because while the *location* of the mutations that causes the disorder is well known, there are literally hundreds of mutations that have been discovered thus far among patients afflicted by this disorder. It would therefore be very interesting to correlate, using computational techniques, the particular mutation with the particular manifestation of the disorder. Unfortunately, most disorders are far more complex, and cannot be related to the malfunction of one specific gene. The paradigm 'one gene, one protein, one function' is not valid in many cases, especially where disorders are involved.

1.2 GENES AND PROTEINS

1.2.1 The Genetic Code

As we have already seen, the genome of an organism is just an enormously long string over the four-symbol nucleotide alphabet $\{A, C, G, T\}$. The genome of an organism is the most 'low level' description of an organism. Here the expression 'low level' is used to mean that it is the most basic description, and when we know that, we don't actually know very much. One can think of the genome as a kind of 'raw data' that needs to be turned into 'information.' This is one of the classic challenges of computational biology.

The next level of complexity in the genome arises from genes and proteins. Proteins are the sustenance of life, and DNA must continually replicate itself so that the production of proteins can go on uninterrupted. The DNA of an organism consists in effect of two parts: (i) the genes whose function is to produce proteins, and (ii) the intergenic regions, often referred to 'junk' DNA. I myself dislike the expression 'junk' DNA. It seems to me that there

[3] Codons are introduced later in this section.

is a vast difference between saying 'This stretch of DNA has no function' and 'This stretch of DNA has no *known* function.' The first is a statement of fact, while the second is a statement about our ignorance. Actually the second sentence accurately describes what we call 'junk DNA.' The situation could change, and we may one day discover *why* there is 'junk' DNA. It is also interesting why it is called 'junk' and not 'trash,' 'garbage,' or even 'spam.'

Proteins were discovered as early as the beginning of the nineteenth century, in fact much earlier than genes. Practically all the proteins discovered at that time were essential dietary ingredients; for example, vitamins are proteins or combinations of proteins. By the 1840's it was already known that every protein consists of a sequence of amino acids, which are twenty in number. These twenty amino acids are denoted either by a single letter, or by two or three letters. Just as the four nucleotides are the building blocks of DNA, the twenty amino acids are the building blocks of proteins. Thus, just we can think of the genome as a string over the four-symbol alphabet of nucleotides, we can think of a protein as a string over the twenty-symbol alphabet of amino acids. The listing out of a protein in terms of its sequence of amino acids is called the **primary structure**. Thus we can think of describing a protein in terms of its primary structure as being analogous to describing an organism in terms of its genome. Both are the 'lowest level' descriptions, and additional work is needed to extract useful information in either case. Typically a protein consists of several hundred, or perhaps a few thousand, amino acids. At the other end, proteins consisting of as few as fifty amino acids are also known.

Once the double helix structure of DNA was discovered in 1953, the scientific community attempted to understand how DNA gets converted to proteins. The working hypothesis, which is by now quite universally accepted, states that first the double-stranded DNA molecule is separated into its two individual strands. Then particular stretches of DNA get cut out and this forms the template for conversion to RNA. In the process, Thymine (T) gets replaced by Uracil (U). RNA is a single-stranded molecule and is thus somewhat unstable chemically. (However, double-stranded RNA has been discovered recently). This process is known as 'transcription.' Then triplets of RNA nucleotides A, C, G, U get converted into amino acids through a process known as 'translation.' The entire hypothesis is labelled as 'The Central Dogma' of biology. This much was understood by 1960, but what happened next was still not clear. Figure 1.2.1 below depicts the central dogma.

Recall that the four nucleotides that make up DNA were discovered in 1929. The basis of the conversion of DNA to proteins, called the 'genetic code,' was discovered in full only in the 1960's. In 1961, Marshall Nirenberg succeeded in showing that the triplet UUU produced the amino acid phenylaline. In quick succession he and his colleagues succeeded in showing that several amino acids were produced by various triplets of nucleotides. It was left to Hargobind Khorana to complete the picture by showing which triplet of nucleotides produced which amino acid. Since there are $4^3 = 64$ triplets

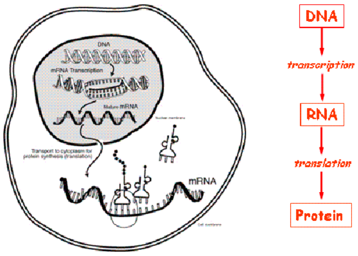


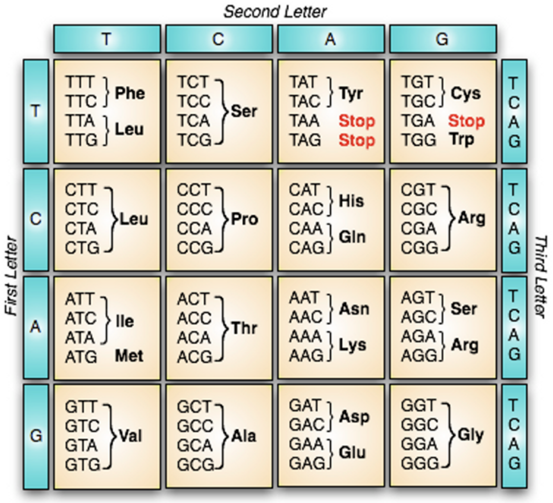
Figure 1.5 Central Dogma of Biology

(called ‘codons’) and only 20 amino acids, there had to be some redundancy. Khorana further showed that each protein-coding RNA ended with one of three sequences, called the **stop codons**, namely UAA, UAG, UGA. Finally, he also showed that every protein-coding RNA began with the ‘start’ codons AUG or GUG. However, while a stop codon cannot occur in the middle of a protein sequence, a ‘start’ codon can also occur in the middle of a protein, where it codes for the amino acid methionine. A key aspect of Khorana’s work was that he was the first one to synthesize oligonucleotides, that is, to create artificially ‘arbitrary’ strings of nucleotides. This synthesis technique allowed him to study the amino acids produced by all possible triplets of nucleotides. In recognition of this seminal work, Nirenberg and Khorana shared the Nobel Prize in Medicine in 1968, along with Robert Holley, who discovered tRNA (translation RNA).

Figure 1.2.1 below depicts the genetic code in compact form. Strictly speaking, RNA codons consist of triplets from the RNA alphabet $\{A, C, G, U\}$. However, since the RNA sequence is obtained simply by replacing the symbol T by the symbol U (ignoring the chemical significance of such a substitution), we can think of ‘codons’ as either triplets over the alphabet $\{A, C, G, T\}$ or over the alphabet $\{A, C, G, U\}$. We use both conventions interchangeably, depending on convenience. Thus we can think of TAA, TAG, TGA as stop codons, and of ATG, GTG as the start codons.

From the above table, it is clear that there is a great deal of subtlety in the genetic code. If we think of the genetic code as a map from the 64-symbol set $\{A, C, G, U\}^3$ into the twenty-one symbol set consisting of the twenty amino acids plus the stop codon, then the structure of the map is not at all clear. The size of the preimage of the various ‘output symbols’ (amino acids or stop codon) ranges from a high of 6 for Leucine to a low of one for many amino acids. Several persons have proposed various speculative explanations for the structure of the genetic code, but until there is no universally accepted explanation.

An interesting aspect of the above chain of discoveries is that *physicists* played a central role in motivating much of this work. After the discovery of individual nucleotides in 1929, the famous physicist Erwin Schrödinger



	Second Letter				
First Letter	T	C	A	G	Third Letter
T	TTT TTC } Phe TTA TTG } Leu	TCT TCC TCA TCG } Ser	TAT TAC } Tyr TAA Stop TAG Stop	TGT TGC } Cys TGA Stop TGG Trp	T C A G
C	CTT CTC CTA CTG } Leu	CCT CCC CCA CCG } Pro	CAT CAC } His CAA CAG } Gln	CGT CGC CGA CGG } Arg	T C A G
A	ATT ATC ATA } Ile ATG Met	ACT ACC ACA ACG } Thr	AAT AAC } Asn AAA AAG } Lys	AGT AGC } Ser AGA AGG } Arg	T C A G
G	GTT GTC GTA GTG } Val	GCT GCC GCA GCG } Ala	GAT GAC } Asp GAA GAG } Glu	GGT GGC GGA GGG } Gly	T C A G

Figure 1.6 The Genetic Code in Tabular Form

suggested very strongly that *there must be a genetic code*, that is, a way of associating strings of nucleotides with amino acids. Schrödinger's best known contribution is of course the 'wave function' formulation of quantum mechanics, which eventually supplanted the earlier 'matrix mechanics' formulation put forth by Werner Heisenberg. Another physicist, George Gamow, suggested on the basis of numerical arguments ($4^3 > 20$) that the genetic code consisted of a map from triplets of nucleotides, that is codons, into amino acids. A very readable description of the entire discovery process can be found in the web site of the Nobel Prize under either Khorana or Nirenberg.

1.2.2 Genes and the Gene-Finding Problem

Roughly speaking, a 'gene' is a stretch of DNA that gets converted into a protein. Within the continuous stretch of a single gene, some regions are called the **coding regions** while others are called **noncoding regions**. The conversion of a gene to a protein can happen in one of two ways. In so-called 'prokaryotes' or 'lower-level' organisms, each gene consists of one continuous stretch of DNA. In so-called 'eukaryotes' or higher organisms, the gene can actually consist of several 'exons' separated by 'introns'. When the gene produces the corresponding protein, the noncoding regions all get cut out, and all the coding regions come together. When this happens, the concatenation of all the coding regions gets converted to a protein according to the genetic code above. Figure 1.2.2 below depicts this process.

It goes without saying that the total length of all the coding regions put

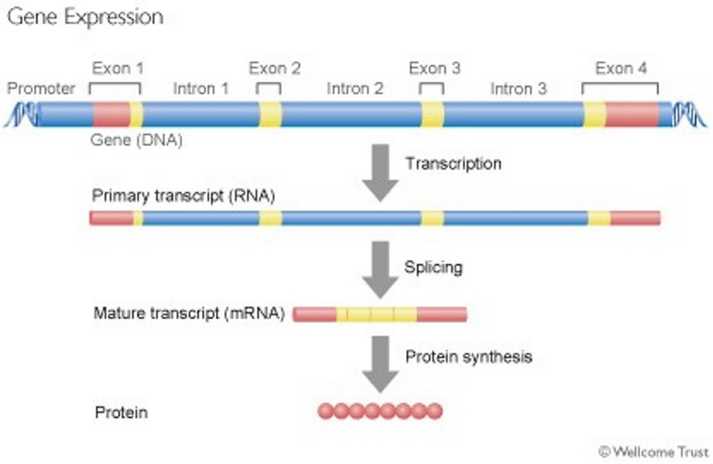


Figure 1.7 Coding by Genes for Proteins

together is an exact multiple of three (so that there is an integer number of codons). Moreover, the last codon is one of the three stop codons TAG, TGA, TAA, while the first codon is one of the start codons ATG, GTG. However, as mentioned above, a start codon can also occur as an intermediate codon. Finally, it has been observed that about 10 to 15 places 'upstream' from the start codon, the tetramer[4] $TATA$ must occur. This tetramer is called the $TATA$-box. Taking all of these features into account, a stretch of DNA that possesses the following features is a possible gene, and is referred to as an ORF (Open Reading Frame):

1. The sequence begins with a start codon ATG or GTG.
2. There is a $TATA$-box 10 to 15 positions upstream of the sequence.
3. The sequence ends with one of the three stop codons TAA, TAG, TGA.
4. The total length of the sequence is 'reasonable,' not less than 300 nucleotides, and not more than 6,000 nucleotides.

The last convention merely indicates that the shortest known protein is 57 amino acids long, corresponding to 161 nucleotides, while the longest known proteins contain about 3,500 amino acids, corresponding to 10,500 nucleotides.

In the case of eukaryotes (higher organisms) an additional complication arises: A gene need not consist of one continuous stretch of DNA; instead, in general it can consist of several 'exons' interspersed by 'introns.' When a gene produces the corresponding protein, first the introns get 'cut out'

[4] A tetramer is a quadruplet of nucleotides. Other commonly used expressions such as trimer, hexamer, etc. are self-explanatory.

and all the exons come together. Then the concatenation of the exons gets converted into a sequence of amino acids via the genetic code discussed earlier. The boundary between an exon and intron is called a 'donor site' and is the dimer AG, while the boundary between and intron and an exon is called an 'acceptor site' and consists of the dimer GT. Of course, the main difficulty is that while every donor site is the dimer AG, the converse is not at all true: Just because we see the dimer AG at some point in the genome, we cannot automatically conclude that it is a donor site. In fact it is fairly easy to see that *only a very small fraction* of the occurences of AG correspond to donor sites. Similar remarks apply to acceptor sites. Together the donor sites and acceptor sites are referred to as 'splice sites.' In general, introns tend to be considerably longer than exons. In eukaryotes, the genes are separated by intergenic regions, the so-called 'junk DNA.' Thus, in order to determine where a gene begins and ends, we need first to weed out the intergenic regions, and then, within each gene, weed out the introns, leaving only the exons.

It is easy to see from the above that determining the ORF's from the genome is completely straight-forward. However, the difficulty is that not all ORF's are genes. (If they were, life would be very simple indeed.) Thus one of the fundamental challenges in genomics is to determine which ORF's are actually genes. One can think of ORFs as a 'candidate genes', or 'putative genes' as some persons prefer to call them.

There are two distinct kinds of algorithms used in the literature to solve the gene-finding problem. These can be described as *ab initio* methods and 'bootstrapping' methods.[5] In *ab initio* methods, one begins with the 'raw' genome and does not assume anything at all about which sections of the genome actually correspond to genes. In prokaryotes, algorithms such as the various versions of GLIMMER [96, 30, 97] begin with the assumption that *all* ORFs that are longer than 500 base pairs are genes. In reality, the vast majority (more than 80%) of ORFs are *not* genes; and yet algorithms based on this assumption seem to work remarkably well. Moreover, such an assumption allows one to analyze a given genome without *any* prior knowledge, which is the meaning of the expression *ab initio*. In bootstrapping algorithms, one needs at least a few ORFs that are 'known' to be genes. By 'known' genes, we mean either that the ORF has been experimentally verified to be a gene, or else that the ORF sequence is sufficiently close to an experimentally verified gene in some other organism, that we can be very confident that the ORF really is a gene without bothering with experimental verification of this particular ORF. In either case, this kind of 'known' gene is commonly referred to as an 'annotated' gene in the literature. When there are a few 'known' (or annotated) genes, these are used as the starting point to analyze other ORFs and to make predictions as to whether those ORFs are genes or not. If the algorithm predicts some ORFs (whose status is previously unknown) to be genes, then the experimenters would go to work to

[5] This terminology is my own and is not at all standard.

validate these predictions. If the predictions are accurate and the predicted gene is indeed a gene, then the predicted (and now validated) gene is added to the database of annotated genes. If the prediction is not borne out by experiment, presumably the originators of the prediction algorithm would introspect on how to improve the accuracy of their algorithms.

For this purpose, two distinct kinds of algorithms are used, namely: (i) deterministic methods based on sequence alignment, and (ii) stochastic algorithms based on statistical analysis.[6] In 'deterministic' algorithms, the premise is that an ORF is actually a gene provided it is sufficiently similar *at a symbol for symbol level* with a known gene. For example, it can be reasonably assumed that the gene that regulates the supply of insulin in humans is similar *at a symbol for symbol level* with the gene of the same function in mice. Thus, if one predicts that an ORF is actually a gene based on symbol for symbol matching, then one would also have a pretty good idea of the *function* of the gene, which is very useful to have. In contrast, in the case of stochastic algorithms, the premise is that different genes within the same organism (or the same family of organisms) will have roughly similar *statistical properties*, even if they don't match at a symbol for symbol level.

To illustrate the difference between deterministic and stochastic algorithms, let us consider the following hypothetical problem. Suppose one is given two different sequences of heads (H) and tails (T) and is asked whether the same coin could have produced both sequences. Clearly, even if both sequences were produced by the same coin, it is extremely unlikely that heads and tails will appear in exactly the same sequence. Thus a deterministic algorithm based on aligning the two sequences would not yield good results. On the other hand, if we compute *the fraction of heads (or tails)* in the two sequences, and the fractions are quite close, then we can state with some confidence that the same coin produced the two sequences.

One of the most popular amongst deterministic algorithms is 'optimal gapped alignment,' which is discussed in Section 9.1. As it turns out, this algorithm is impractical when one wishes simultaneously to align a very large number of sequences. To address this difficulty, a statistical (as opposed to stochastic) algorithm known as BLAST has been developed. This algorithm is discussed in detail in Chapter 9; however it must be mentioned that only the original version of BLAST is discussed in this chapter, and not its subsequent variants. BLAST theory is based on estimating the probability (or likelihood) of rare events; this kind of problem is known as 'tail probability estimation.' BLAST theory uses something called 'the method of types,' which is an essential tool in so-called large deviation theory. Large deviation theory gives us very precise formulas for the 'rate' at which empirically estimated quantities (such as frequencies of occurence of various events) converge to their true values. Large deviation theory is discussed in Chapter 8. Stochastic algorithms for various problems in computational biology are often based on interpreting a sequence of symbols (for example nucleotides or

[6] Here too the terminology is my own and not standard.

amino acids) as the realization of a stochastic process. While it is possible to construct quite complex models for these stochastic processes, Markov models and hidden Markov models are very popular in biology. Such models are introduced in Chapters 5 through 8, while stochastic algorithms for various problems in gene finding and protein classification are discussed in Chapter 9.

1.2.3 Proteins and the Protein Classification Problem

Proteins are at the next level of complexity after genes. As stated above, the central dogma of biology describes how genes get converted into proteins. The original and rather simplistic recipe of 'one gene, one protein, one function' has long since been revised in favor of far more subtle models. For instance, it is now known that the same gene can, in different kinds of cells, code for different protein. These subtleties of biochemistry are beyond the scope of this book. For present purposes, we will stick with the simple model whereby the relevant parts of a gene come together during translation, codons get converted into amino acids as per the genetic code, and the resulting sequence of amino acids forms the protein. Thus, once we know that a particular stretch of DNA represents a gene, and we know the functional parts of the gene (coding regions and/or exons), we can unambiguously determine the sequence of amino acids produced by that gene.

The sequence of amino acids is known as the 'primary structure' of a protein. It is the lowest level description of a protein, just as the genome is the lowest level description of the DNA of an organism. As befits a 'low level' description, knowing the primary structure of a protein does not get us very far in terms of knowing how a protein 'works.' In order to understand how a protein performs its assigned function, it is highly desirable to know how the protein 'folds,' that is, the three-dimensional structure of the protein, which is known as the tertiary structure of the protein, and also its so-called 'active sites'. The tertiary structure of a protein corresponds to the 3-D conformation that minimizes the potential energy of the conformation. While this simple-sounding statement is consistent with the laws of physics, in reality the potential energy function of a protein is a highly complex function of its conformation. The 'conformation' itself consists of $2(n-1)$ angles where n is the number of amino acids, representing the two degrees of freedom at each joint between two amino acid molecules. Each amino acid molecule can be thought as being essentially 'rigid'; however, the orientation at each joint represents two degrees of freedom. The potential energy term must also include the interaction of each amino acid with the surrounding medium, usually water.

Figure 1.2.3 below shows the tertiary structure of a few molecules. A somewhat simplified description of the tertiary structure is called the 'secondary structure,' and consists of just three elements: α-helices, β-sheets, and strands. The same figure also shows the secondary structures corresponding to each tertiary structure.

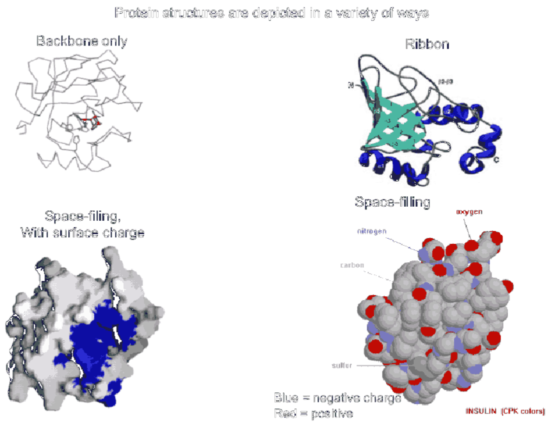


Figure 1.8 Secondary and Tertiary Structures of Various Proteins

Finally, when two or more proteins bond, the result is a molecule that is still more complex. The 3-D structure of the protein-protein complex is referred to as the quaternary structure. Figure 1.2.3 below shows the quaternary structure of hemoglobin, one of the molecules that is vital to life. It is a combination of four different proteins.

If a protein can be crystallized, then its tertiary structure can be determined on the basis of experimental methods such as NMR or x-ray diffraction. However, many proteins of interest cannot be crystallized. In such a case one is forced to resort to other methods to 'predict' the 3-D structure of the protein. Even if a protein can be crystallized, the procedure for determining the structure is both time-consuming and expensive. Thus there is a definite need for computational methods for predicting the tertiary structure of a protein.

As mentioned earlier, in principle it is possible to determine the structure of a protein by minimizing its potential energy. In practice however, the minimization problem is intractable for all but extremely short proteins. This has led to a number of methods for predicting protein structure, which are discussed at length in [12]. These methods include some *ab initio* methods, as well as methods based on neural networks. Amongst the most popular are 'homology-based' methods, whereby several proteins whose structures are known are grouped into a small number of families, typically three or four families. Then the protein of interest is 'classified' as being most similar to one of the protein families. If the similarity is sufficiently high, then one makes a guess that the 3-D structure of the new protein is similar to those of the 'known' proteins. This approach has the advantage (from the standpoint of the present book) of being based on hidden Markov models.

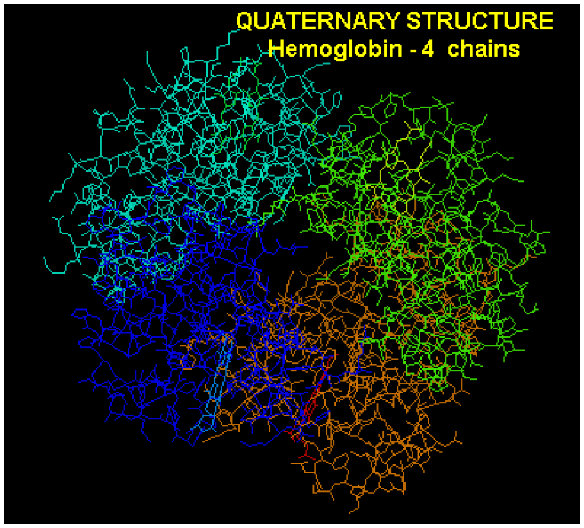


Figure 1.9 Quaternary Structure of the Hemoglobin Molecule

Chapter Two

Introduction to Probability and Random Variables

2.1 INTRODUCTION TO RANDOM VARIABLES

2.1.1 Motivation

Probability theory is an attempt to formalize the notion of uncertainty in the outcome of an experiment. For instance, suppose an urn contains four balls, colored red, blue, white and green respectively. Suppose we dip our hand in the urn and pull out one of the balls 'at random.' What is the likelihood that the ball we pull out will be red? What is the likelihood that we have to draw a ball at least ten times (replacing the drawn ball each time and shaking the urn thoroughly) before we draw a red ball for the first time? Probability theory provides a mathematical abstraction and a framework where we can address such issues.

When there are only finitely many possible outcomes, probability theory becomes relatively simple. For instance, in the above example, when we draw a ball there are only four possible outcomes, namely: $\{R, B, W, G\}$ with the obvious notation. If we draw two balls, after replacing the first ball drawn, then there $4^2 = 16$ possible outcomes, represented as $\{RR, \ldots, GG\}$. In such situations, one can get by with simple 'counting' arguments. The counting approach can also be made to work when the set of possible outcomes is *countably* infinite.[1] However, in probability theory infinity is never very far away, and counting arguments can lead to serious logical inconsistencies if applied to situations where the set of possible outcomes is *uncountably* infinite. The great Russian mathematician A. N. Kolmogorov invented axiomatic probability theory in the 1930's precisely to address the issues thrown up by having uncountably many possible outcomes. Subsequent development of probability theory has been based on this axiomatic foundation.

Example 2.1 Let us return to the example above. Suppose that all the four balls are identical in size and shape, and differ only in their color. Then it is reasonable to suppose that drawing any one color is as likely, neither more nor less, than any other color. This leads to the observation that the likelihood of drawing a red ball (or any other ball) is $1/4 = 0.25$.

Example 2.2 Now suppose that the four balls are all spherical, and that

[1] Recall that a set S is said to be **countable** if it can be place in one-to-one correspondence with the set of natural numbers $\mathbb{N} = \{1, 2, \ldots\}$.

their diameters are in the ratio $4 : 3 : 2 : 1$ in the order red, blue, white and green. We can suppose that the likelihood of our fingers touching and drawing a particular ball is proportional to its surface area. In this case, it follows that the likelihoods of drawing the four balls are in the proportion $4^2 : 3^2 : 2^2 : 1^2$ or $16 : 9 : 4 : 1$ in the order red, blue, white and green. This leads to the conclusion that

$$P(R) = 16/30, P(B) = 9/30, P(W) = 4/30, P(G) = 1/30.$$

Example 2.3 There can be instances where such analytical reasoning can fail. Suppose the red ball is coated with an adhesive resin that makes it more likely to stick to our fingers when we touch it. The complicated interaction between the surface adhesion of our fingers and the surface of the ball may be too difficult to analyze, so we have no recourse other than to draw balls repeatedly and see how many times the red ball comes out. Suppose we make 1,000 draws, and the outcomes are: 451 red, 187 blue, 174 white and 188 green. Then we can write

$$\hat{P}(R) = 0.451, \hat{P}(B) = 0.187, \hat{P}(W) = 0.174, \hat{P}(G) = 0.188.$$

The symbol $\hat{P}$ is used instead of P to highlight the fact that these are simply *observed frequencies*, and not the true but unknown probabilities. It is tempting to treat the observed frequencies as true probabilities, but that would not be correct. The reason is that if the experiment is repeated, the outcomes may be quite different. The reader can convince himself/herself of the difference between frequencies and probabilities by tossing a coin ten times, and another ten times. It is extremely unlikely that the same set of results will turn up both times. One of the major questions addressed in this book is: Just how close are the observed *frequencies* to the true but unknown *probabilities*, and just how quickly do these observed frequencies converge to their true values (namely, the true probabilities)? Such questions are addressed in Chapter 6.

2.1.2 Definition of a Random Variable and Probability

Suppose we wish to study the behaviour of a 'random' variable $\mathcal{X}$ that can assume one of only a finite set of values belonging to a set $\mathbb{A} = \{a_1, \ldots, a_n\}$. The set $\mathbb{A}$ of possible values is often referred to as the 'alphabet' of the random variable. For example, in the ball-drawing experiment discussed in the preceding subsection, $\mathcal{X}$ can be thought of as the color of the ball drawn, and assumes values in the set $\{R, B, W, G\}$. This example, incidentally, serves to highlight the fact that the set of outcomes can consist of abstract *symbols*, and need not consist of *numbers*. This usage, adopted in this book, is at variance from the usual convention in mathematics texts, where it is assumed that $\mathbb{A}$ is a subset of the real numbers $\mathbb{R}$. However, since biological applications are a prime motivator for this book, it makes no sense to restrict $\mathbb{A}$ in this way. In genomics, for example, $\mathbb{A}$ consists of the four symbol set of nucleic acids, or nucleotides, usually denoted by $\{A, C, G, T\}$. Moreover, by allowing $\mathbb{A}$ to consist of arbitrary symbols, we also allow explicitly the possibility that *there is no natural ordering of these symbols.* For instance, in this book the nucleotides are written in the order A, C, G, T purely to follow the English alphabetical ordering. But there is no consensus on the ordering in biology texts. Thus any method of analysis that is developed here must be *permutation independent.* In other words, if we choose to order the symbols in the set $\mathbb{A}$ in some other fashion, the methods of analysis must give the same answers as before.

Now we give a general definition of the notion of probability, and introduce the notation that is used throughout the book.

Definition 2.1 *Given an integer n, the* **n-dimensional simplex** $\mathbb{S}_n$ *is defined as*

$$\mathbb{S}_n = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq 0 \; \forall i, \sum_{i=1}^{n} x_i = 1\}. \tag{2.1}$$

Thus $\mathbb{S}_n$ consists of all nonnegative vectors whose components add up to one.

Definition 2.2 *Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set. Then a* **probability distribution** *on the set $\mathbb{A}$ is any vector $\boldsymbol{\mu} \in \mathbb{S}_n$.*

The interpretation of a probability distribution $\boldsymbol{\mu}$ on the set $\mathbb{A}$ is that we say

$$\Pr\{\mathcal{X} = a_i\} = \mu_i$$

to be read as 'the probability that the random variable $\mathcal{X}$ equals x_i is μ_i.' Thus, if $\mathbb{A} = \{R, B, W, G\}$ and $\boldsymbol{\mu} = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25]$, then all the four outcomes of drawing the various colored balls are equally likely. This is the case in Example 2.1. If the situation is as in Example 2.2, where the balls have different diameters in the proportion 4 : 3 : 2 : 1, the probability distribution is

$$\boldsymbol{\mu} = [16/30 \;\; 9/30 \;\; 4/30 \;\; 1/30].$$

If we now choose to reorder the elements of the set $\mathbb{A}$ in the form $\{R, W, G, B\}$, then the probability distribution gets reordered correspondingly, as

$$\boldsymbol{\mu} = [16/30 \;\; 4/30 \;\; 1/30 \;\; 9/30].$$

Thus, when we speak of the probability distribution $\boldsymbol{\mu}$ on the set $\mathbb{A}$, we need to specify the ordering of the elements of the set.

The way we have defined it above, a probability distribution associates a *weight* with each element of the set $\mathbb{A}$ of possible outcomes. Thus $\boldsymbol{\mu}$ can be thought of as a map from $\mathbb{A}$ into the interval $[0, 1]$. This notion of a weight of *individual elements* can be readily extended to define the weight of *each* *subset* of $\mathbb{A}$. This is called the probability **measure** P_μ associated with the distribution $\boldsymbol{\mu}$. Suppose $A \subseteq \mathbb{A}$. Then we define

$$P_\mu(A) := \Pr\{\mathcal{X} \in A\} = \sum_{i=1}^{n} \mu_i I_A(x_i), \tag{2.2}$$

where $I_A(\cdot)$ is the so-called **indicator function** of the set A, defined by

$$I_A(x) = \begin{cases} 1 & \text{if} \quad x \in A, \\ 0 & \text{if} \quad x \notin A. \end{cases} \tag{2.3}$$

So (2.2) states that the **probability measure** of the set A, denoted by $P_\mu(A)$, is the sum of the probability weights of the individual elements of the set A. Thus, whereas $\boldsymbol{\mu}$ maps the set $\mathbb{A}$ into $[0, 1]$, the corresponding probability measure P_μ maps the 'power set' $2^{\mathbb{A}}$ (that is, the collection of all subsets of $\mathbb{A}$) into the interval $[0, 1]$.

In this text, we need to deal with three kinds of objects:

- A probability distribution $\boldsymbol{\mu}$ on a finite set $\mathbb{A}$.
- A random variable $\mathcal{X}$ assuming values in $\mathbb{A}$, with the probability distribution $\boldsymbol{\mu}$.
- A probability measure P_μ on the power set $2^{\mathbb{A}}$, associated with the probability disbribution $\boldsymbol{\mu}$.

We will use whichever interpretation is most convenient and natural in the given context. As for notation, throughout the text, bold face greek letters such as $\boldsymbol{\mu}$ denote probability distributions. The probability measure corresponding to $\boldsymbol{\mu}$ is denoted by P_μ. Strictly speaking, we should write $P_{\boldsymbol{\mu}}$, but for reasons of aesthetics and appearence we prefer to use P_μ. Similar notation applies to all other bold face greek letters.

From (2.2), it follows readily that the empty set $\emptyset$ has probability measure zero, while the complete set $\mathbb{A}$ has probability measure one. This is true irrespective of what the underlying probability distribution $\boldsymbol{\mu}$ is. Moreover, the following additional observations are easy consequences of (2.2):

Theorem 2.3 *Suppose $\mathbb{A}$ is a finite set and $\boldsymbol{\mu}$ is a probability distribution on $\mathbb{A}$, and let P_μ denote the corresponding probability measure. Then*

1. $0 \leq P_\mu(A) \leq 1 \; \forall A \subseteq \mathbb{A}$.
2. *If A, B are disjoint subsets of $\mathbb{A}$, then*
$$P_\mu\,(A \cup B) = P_\mu(A) + P_\mu(B). \tag{2.4}$$

In the next paragraph we give a brief glimpse of axiomatic probability theory in a general setting, where the set $\mathbb{A}$ of possible outcomes is not necessarily finite. This paragraph is *not needed* to understand the remainder of the book, and therefore the reader can skip it with no after-effects. In axiomatic probability theory, one actually *begins* with generalizations of the two properties above. One starts with a collection of subsets $\mathcal{S}$ of $\mathbb{A}$ that has three properties:

1. Both the empty set $\emptyset$ and $\mathbb{A}$ itself belong to $\mathcal{S}$.

2. If A belongs to $\mathcal{S}$, so does its complement A^c.

3. If $\{A_1, A_2, \ldots\}$ is a *countable* collection of sets belonging to $\mathcal{S}$, then their union $B := \cup_{i=1}^{\infty} A_i$ also belongs to $\mathcal{S}$.

Then the probability measure P is defined to be a function that assigns a number $P(A) \in [0, 1]$ to each set A belonging to $\mathcal{S}$ such that two properties hold. First, $P(\emptyset) = 0$ and $P(\mathbb{A}) = 1$. Second, if $\{A_1, A_2, \ldots\}$ is a *countable* collection of *pairwise disjoint* sets belonging to $\mathcal{S}$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

In the case where the set $\mathbb{A}$ is finite, we can as well take $\mathcal{S}$ to be the collection of *all* subsets of $\mathbb{A}$. If $\mathbb{A}$ is finite, and if $\{A_1, A_2, \ldots\}$ is a countable collection of pairwise disjoint sets, then the only possibility is that all but finitely many sets are empty. So Property 3 above can be simplified to:

$$P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset,$$

which is precisely Property 2 from Theorem 2.3. In this case, assigning a probability measure to *each subset* of $\mathbb{A}$ while satisfying the above restriction is the same as assigning a nonnegative weight to *each element* of $\mathbb{A}$ while ensuring that the weights add up to one. If $\mathbb{A}$ is infinite but countable, we can still use the same approach. That is, we can assign nonnegative weights to each element of $\mathbb{A}$, such that the weights add up to one, we can take $\mathcal{S}$ to consist of *all* subsets of $\mathbb{A}$, and for every subset A of $\mathbb{A}$, we can define $P(A)$ by (2.2). This is why most ‘elementary’ books on Markov chains, for example, assume that the underlying set $\mathbb{A}$ is either finite or countably infinite. But if $\mathbb{A}$ is an uncountable infinite set (such as the real numbers, for example), this approach based on assigning weights does not work.

At this point the reader can well ask: But what does it all *mean*? As with much of mathematics, probability theory exists at many distinct levels. It can be viewed as an exercise in pure reasoning, an intellectual pastime, a challenge to one’s wits. While that may satisfy some persons, the theory would have very little by way of *application* to ‘real’ situations unless the notion of probability is given a little more concrete interpretation. So we can think of the probability distribution $\boldsymbol{\mu}$ as arising in one of two ways. First, the distribution can be *postulated*, as in the previous subsection. Thus if we are drawing from an urn containing four balls that are identical in all respects save their color, it makes sense to *postulate* that each of the four outcomes is equally likely. Similarly, if the balls are identical except for their diameter, and if we believe that the likelihood of drawing a ball is proportional to the surface area, then once again we can *postulate* that the four components of $\boldsymbol{\mu}$ are in proportion to the areas (or equivalently, to the diameter squared) of the four balls. Then the requirement that the components of $\boldsymbol{\mu}$ must add up to one gives the normalizing constant. Second, the distribution can be *estimated*, as with the adhesive-coated balls in Example 2.3. Thus we can declare that there is a true but unknown probability vector $\boldsymbol{\mu}$, and that our estimate of it, based on 1,000 draws of balls, is $\hat{\boldsymbol{\mu}} = [0.451 \;\; 0.187 \;\; 0.174 \;\; 0.188]$. Then we can try to develop theories that allow us to say *how close $\hat{\boldsymbol{\mu}}$ is to $\boldsymbol{\mu}$, and with what confidence we can make this statement.* The reader is referred to Chapter 6 for a discussion of such topics.

2.1.3 Function of a Random Variable, Expected Value

Suppose $\mathcal{X}$ is a random variable assuming values in a finite set $\mathbb{A} = \{a_1, \ldots, a_n\}$, with the probability measure P_μ and the probability distribution $\boldsymbol{\mu}$. Suppose f is a function mapping the set $\mathbb{A}$ into another set $\mathbb{B}$. Since $\mathbb{A}$ is finite, it is clear that the set $\{f(a_1), \ldots, f(a_n)\}$ is finite. So there is no loss of generality in assuming that the set $\mathbb{B}$ (the range of the function f) is also a finite set. Moreover, it is *not* assumed that the values $f(a_1), \ldots, f(a_n)$ are distinct. Thus the image of the set $\mathbb{A}$ under the function f can have fewer than n elements. Now $f(\mathcal{X})$ is itself a random variable. Moreover, the distribution of $f(\mathcal{X})$ can be computed readily from the distribution of $\mathcal{X}$. Suppose $\boldsymbol{\mu} \in \mathcal{S}_n$ is the distribution of $\mathcal{X}$. Thus $\mu_i = \Pr\{\mathcal{X} = x_i\}$. To compute the distribution of $f(\mathcal{X})$, we need to address the possibility that $f(a_1), \ldots, f(a_n)$ may not be distinct elements. Let $\mathbb{B} = \{b_1, \ldots, b_m\}$ denote the set of all possible outcomes of $f(\mathcal{X})$, and note that $m \leq n$. Then

$$\Pr\{f(\mathcal{X}) = b_j\} = \sum_{a_i \in f^{-1}(b_j)} \mu_i.$$

In other words, the probability that $f(\mathcal{X}) = b_j$ is the sum of the probabilities of all the preimages of b_j under the function f.

Example 2.4 Let the set $\mathcal{R}$ equal $\{A, C, G, U\}$, the set of RNA nucleotides. While chemically DNA and RNA are quite different, as explained in Chapter 1, during the transcription phase of DNA reproduction, Thymine (T) gets replaced by Uracil (U). Now let the set $\mathbb{A}$ equal $\mathcal{R}^3$, the set of all triplets over the alphabet $\{A, C, G, U\}$. As discussed in Chapter 1, each triplet is called a ‘codon.’ Clearly X contains $4^3 = 64$ elements. Now the 64 codons get mapped into the 20 amino acids and the STOP symbol, as shown in Table 1.1. So we can define the function $f : \mathcal{R}^3 \to \mathbb{B}$, where $\mathbb{B}$ is a set of cardinality 21 containing the symbols for the 20 amino acids and the STOP symbol. From Table 1.1 it is clear that the size of the preimages $f^{-1}(b)$ varies quite considerably for each of the 21 elements of $\mathbb{B}$, ranging from 6 down to 1. Thus, if we know the frequency of distribution of codons in a particular stretch of genome (a frequency distribution on X), we can convert this into a corresponding frequency distribution of amino acids and stop codons.

Example 2.5 Suppose that in Example 2.1, we receive a payoff of \$2 if we draw a green ball, we pay a penalty of \$1 if we draw a red ball, and we neither pay a penalty nor receive a payment if we draw a white ball or a blue ball. In this case $f(R) = -1$, $f(G) = 2$, and $f(W) = f(B) = 0$.

Definition 2.4 *Suppose $\mathcal{X}$ is a* real-valued *random variable assuming values $\mathbb{A} = \{a_1, \ldots, a_n\}$, with the probability distribution $\boldsymbol{\mu}$, and associated probability measure P_μ. Then the* **expected value** *of $\mathcal{X}$ is defined as*

$$E[\mathcal{X}] := \sum_{i=1}^{n} a_i \mu_i. \tag{2.5}$$

It is important to note that, while the notion of probability can be defined for *any* random variable (for example, the set of nucleotides or the set of amino acids), the notion of an expected value can be defined only for *real-valued* random variables.

Suppose now that $\mathcal{X}$ is a random variable assuming values in some finite set $\mathbb{A}$ (not necessarily a subset of $\mathbb{R}$), and f is a function mapping the set $\mathbb{A}$ into the real numbers $\mathbb{R}$. Thus to each element $a_i \in \mathbb{A}$, the function f assigns a real number $f(a_i)$. Let $\boldsymbol{\mu}$ denote the distribution of $\mathcal{X}$ and let P_μ denote the associated probability measure.

Definition 2.5 *The* **expected value** *of the function f is denoted by $E[f, P_\mu]$ and is defined by*

$$E[f, P_\mu] := \sum_{i=1}^{n} f(a_i)\mu_i = \sum_{i=1}^{n} f(a_i) \Pr\{\mathcal{X} = a_i\}. \tag{2.6}$$

It is left to the reader to verify that the above equation is the same as the expected value of the random variable $f(\mathcal{X})$. The point to note is that the formula (2.6) is valid even if the real numbers $f(a_1), \ldots, f(a_n)$ are not all distinct.

There is a small bit of pedantry that needs an explanation. Note that we write $E[\mathcal{X}]$ for the expected value of a random variable, without explicitly displaying the underlying probability measure P_μ. The reasoning is that if $\mathcal{X}, \mathcal{Y}$ both assume values in the same set $\mathbb{A}$ but have different different probability distributions, then they are in reality distinct random variables. On the other hand, in the above definition, f is a *fixed* function from $\mathbb{A}$ into $\mathbb{B}$. So if $\mathcal{X}, \mathcal{Y}$ are distinct random variables assuming values in $\mathbb{A}$, then in principle the expected values of $f(\mathcal{X}), f(\mathcal{Y})$ could be different. So we write $E[f, P_\mu]$; we could also write $E[f(\mathcal{X})]$ in which case the measure P_μ is implicit through $\mathcal{X}$.

Observe that the expected value is *linear* in the function f. Thus, if f, g are two functions of a random variable $\mathcal{X}$ with probability measure P_μ and the probability distribution $\boldsymbol{\mu}$, and α, β are two real numbers, then

$$E[\alpha f + \beta g, P_\mu] = \alpha E[f, P_\mu] + \beta E[g, P_\mu].$$

Suppose $\mathcal{X}$ is a *real-valued* random variable assuming values in $\mathbb{A} = \{a_1, \ldots, a_n\}$, with probability measure P_μ and the probability distribution

$\boldsymbol{\mu}$. Then the quantity defined earlier as the expected value of $\mathcal{X}$, namely

$$E[\mathcal{X}] = \sum_{i=1}^{n} a_i \mu_i,$$

is also called the **mean value** or just simply the **mean** of $\mathcal{X}$. The quantity $E[(\mathcal{X} - E(\mathcal{X}, P_\mu))^2, P_\mu]$ is called the **variance** of $\mathcal{X}$. The square root of the variance is called the **standard deviation** of $\mathcal{X}$. In many books, the symbols $\mu(\mathcal{X}), \sigma(\mathcal{X})$ are commonly used to denote the mean and standard deviation respectively. However we do not use that notation.

Note that we can also define the variance of $\mathcal{X}$ as $E[\mathcal{X}^2, P_\mu] - (E[\mathcal{X}])^2$. This is because, by the linearity of the expected value operation, we have

$$\begin{aligned} E[(\mathcal{X} - E(\mathcal{X}))^2, P_\mu] &= E[\mathcal{X}^2, P_\mu] - 2(E[\mathcal{X}])^2 + (E[\mathcal{X}])^2 \\ &= E[\mathcal{X}^2, P_\mu] - (E[\mathcal{X}])^2. \end{aligned} \tag{2.7}$$

The above argument also shows that, for every random variable $\mathcal{X}$, we have

$$E[\mathcal{X}^2, P_\mu] \geq (E[\mathcal{X}])^2.$$

This is a special case of a very general result known as Schwarz' inequality.

2.1.4 Total Variation Distance Between Two Probability Measures

Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set, and $\boldsymbol{\mu}, \boldsymbol{\nu}$ are two probability distributions on X. Let P_μ, P_ν denote the corresponding probability measures. In this section, we show how to quantify the 'difference' between the two measures.

Definition 2.6 *Let P_μ, P_ν be two probability measures on a finite set $\mathbb{A} = \{a_1, \ldots, a_n\}$, corresponding to the distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$ respectively. Then the* **total variation distance** *between P_μ and P_ν (or between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$), denoted by $\rho(P_\mu, P_\nu)$ (or $\rho(\boldsymbol{\mu}, \boldsymbol{\nu})$), is defined as*

$$\rho(P_\mu, P_\nu), \rho(\boldsymbol{\mu}, \boldsymbol{\nu}) := \max_{A \subseteq \mathbb{A}} |P_\mu(A) - P_\nu(A)|. \tag{2.8}$$

Now it is shown that $\rho(\cdot, \cdot)$ is indeed a proper metric or 'distance' on $\mathbb{S}_n$, which can be identified with the set of all probability distributions on the set $\mathbb{A}$ of cardinality n.

Lemma 2.7 *The function $\rho(\cdot, \cdot)$ defined in (2.8) satisfies the following properties:*

1. *$\rho(P_\mu, P_\nu) \geq 0$ for all $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$.*
2. *$\rho(P_\mu, P_\nu) = 0$ if and only if $\boldsymbol{\mu} = \boldsymbol{\nu}$.*
3. *$\rho(P_\mu, P_\nu) = \rho(P_\nu, P_\mu)$ for all $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$.*
4. *The so-called 'triangle inequality' is satisfied, namely:*

$$\rho(P_\mu, P_\nu) \leq \rho(P_\mu, P_\phi) + \rho(P_\phi, P_\nu) \; \forall \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\phi} \in \mathbb{S}_n. \tag{2.9}$$

Proof. Property No. 1 is obvious. To prove Property No. 2, note that $\rho(P_\mu, P_\nu) = 0$ if $\boldsymbol{\mu} = \boldsymbol{\nu}$. Thus the key observation is the converse, or equivalently, $\rho(P_\mu, P_\nu) > 0$ if $\boldsymbol{\mu} \neq \boldsymbol{\nu}$. Note that if $\boldsymbol{\mu} \neq \boldsymbol{\nu}$, then $\mu_i \neq \nu_i$ for at least one index i (actually for at least two indices). Let $A = \{i\}$, where i is an index such that $\mu_i \neq \nu_i$. Then

$$P_\mu(A) = \mu_i \neq \nu_i = P_\nu(A).$$

Hence $\rho(P_\mu, P_\nu) > 0$. Property No. 3 is again obvious. Finally, Property No. 4 follows from the triangle inequality for real numbers, namely:

$$|x - y| \leq |x - z| + |z - y|, \; \forall x, y, z \in \mathbb{R}.$$

Now suppose $A \subseteq \mathbb{A}$ is arbitrary. Then the triangle inequality for real numbers implies that

$$|P_\mu(A) - P_\nu(A)| \leq |P_\mu(A) - P_\phi(A)| + |P_\phi(A) - P_\nu(A)|, \; \forall A \subseteq \mathbb{A}.$$

Taking the maximum over all $A \subseteq \mathbb{A}$ proves Property No. 4. □

As defined in (2.8), $\rho(P_\mu, P_\nu)$ is the maximum difference between $P_\mu(A)$ and $P_\nu(A)$ as A varies over the 2^n subsets of $\mathbb{A}$. Clearly, (2.8) is an impractical formula for actually *computing* the number $\rho(P_\mu, P_\nu)$. The next theorem gives a number of equivalent formulas for computing $\rho(P_\mu, P_\nu)$. Note that, given a real number $x \in \mathbb{R}$, the symbol x_+ denotes the positive part of x, that is, $\max\{x, 0\}$. Similarly x_- denotes $\min\{x, 0\}$.

Theorem 2.8 *Suppose* $\mathbb{A} = \{a_1, \ldots, a_n\}$ *is a finite set, and that* P_μ, P_ν *are two probability measures on* $\mathbb{A}$ *with associated distributions* $\boldsymbol{\mu}$ *and* $\boldsymbol{\nu}$ *respectively. Then*

$$\rho(P_\mu, P_\nu) = \sum_{i=1}^n (\mu_i - \nu_i)_+ \tag{2.10}$$

$$= -\sum_{i=1}^n (\mu_i - \nu_i)_- \tag{2.11}$$

$$= \frac{1}{2} \sum_{i=1}^n |\mu_i - \nu_i|. \tag{2.12}$$

Proof. Define $\delta_i := \mu_i - \nu_i$, for $i = 1, \ldots, n$. Then, since $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$, it follows that $\sum_{i=1}^n \delta_i = 0$. Moreover, for any set $A \subseteq \mathbb{A}$, we have that

$$P_\mu(A) - P_\nu(A) = \sum_{a_i \in A} \delta_i = \sum_{i=1}^n I_A(a_i) \delta_i,$$

where $I_A(\cdot)$ is the indicator function of the set A. Now, let us look at the 2^n numbers $P(A) - Q(A)$ generated by varying A over all subsets of $\mathbb{A}$. (These numbers may not all be distinct.) Let $S \subseteq \mathbb{R}$ denote the set of all these numbers. The first point to note is that

$$P_\mu(A^c) - P_\nu(A^c) = [1 - P_\mu(A)] - [1 - P_\nu(A)] = -[P_\mu(A) - P_\nu(A)].$$

Hence the set S is symmetric: If $x \in S$ (corresponding to a set A), then $-x \in S$ (corresponding to the set A^c). Observe that $\rho(P_\mu, P_\nu)$ is the largest value of $|x|$ for all $x \in S$. However, because of the symmetry of the set S, $\rho(P_\mu, P_\nu)$ also equals the largest value of $x \in S$, and also

$$\rho(P_\mu, P_\nu) = -\min\{x \in S\}.$$

So if we can find the largest or the smallest element in S, then we would have found $\rho(P_\mu, P_\nu)$.

Next, let $\mathcal{N}_+ \subseteq \{1, \ldots, n\}$ denote the set of indices i for which $\delta_i \geq 0$, and let $\mathcal{N}_-$ denote the set of indices i for which $\delta_i < 0$. Then

$$P_\mu(A) - P_\nu(A) = \sum_{i \in \mathcal{N}_+} I_A(a_i)\delta_i + \sum_{i \in \mathcal{N}_-} I_A(a_i)\delta_i.$$

Now the first summation consists of only nonnegative numbers, while the second summation consists only of nonpositive numbers. Therefore the largest possible value of $P_\mu(A) - P_\nu(A)$ is

$$\sum_{i \in \mathcal{N}_+} I_A(a_i)\delta_i = \sum_{i=1}^{n} (\delta_i)_+,$$

and corresponds to the choice $A = \{a_i : i \in \mathcal{N}_+\}$. By the discussion in the preceding paragraph, it follows that

$$\rho(P_\mu, P_\nu) = \sum_{i=1}^{n} (\delta_i)_+,$$

which is precisely (2.10). Similarly, the smallest value of $P_\mu(A) - P_\nu(A)$ is

$$\sum_{i \in \mathcal{N}_-} I_A(a_i)\delta_i = \sum_{i=1}^{n} (\delta_i)_-,$$

corresponding to the choice $A = \{a_i : i \in \mathcal{N}_-\}$. Again from the discussion in the previous paragraph, it follows that

$$\rho(P_\mu, P_\nu) = -\sum_{i=1}^{n} (\delta_i)_-,$$

which is precisely (2.11). Finally, observe that

$$\sum_{i=1}^{n} |\delta_i| = \sum_{i=1}^{n} (\delta_i)_+ - \sum_{i=1}^{n} (\delta_i)_- = 2\rho(P_\mu, P_\nu),$$

which establishes (2.12). □

From the definition (2.8), it is immediate that $\rho(P_\mu, P_\nu) \in [0, 1]$. This is because both $P_\mu(A)$ and $P_\nu(A)$ lie in the range $[0, 1]$, and so $P_\mu(A) - P_\nu(A) \in [-1, 1]$. Now the proof of Theorem 2.8 shows that, for every pair of probability measures P_μ and P_ν, there actually exists a set A such that $P_\mu(A) - P_\nu(A) = \rho(P_\mu, P_\nu)$; one such choice is $A = \{a_i : i \in \mathcal{N}_+\}$. Now, if

$\rho(P_\mu, P_\nu)$ actually equals one, this implies that $P_\mu(A) = 1$ and $P_\nu(A) = 0$ (and also that $P_\mu(A^c) = 0$ and $P_\nu(A^c) = 1$). In such a case the two measures P_μ and P_ν are said to be **mutually singular**, because their weights are supported on disjoint sets: The weights μ_i are concentrated on the set A whereas the weights ν_i are concentrated on the set A^c.

Lemma 2.9 *Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set, and P_μ, P_ν are two probability measures on $\mathbb{A}$. Suppose $f : \mathbb{A} \to [-1, 1]$. Then*

$$|E[f, P_\mu] - E[f, P_\nu]| \leq 2\rho(P_\mu, P_\nu). \tag{2.13}$$

Proof. The proof follows by direct substitution. We have that

$$\begin{aligned}
|E[f, P_\mu] - E[f, P_\nu]| &= \left| \sum_{i=1}^{n} f(a_i)(\mu_i - \nu_i) \right| \\
&\leq \sum_{i=1}^{n} |f(a_i)| \cdot |\mu_i - \nu_i| \\
&\leq \sum_{i=1}^{n} |\mu_i - \nu_i| \text{ since } |f(a_i)| \leq 1 \\
&= 2\rho(P_\mu, P_\nu).
\end{aligned}$$

□

Lemma 2.10 *Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set, and P_μ, P_ν are two probability measures on $\mathbb{A}$. Suppose $f : \mathbb{A} \to [0, 1]$. Then*

$$|E[f, P_\mu] - E[f, P_\nu]| \leq \rho(P_\mu, P_\nu). \tag{2.14}$$

Proof. This lemma can be derived as a corollary of Lemma 2.9 by observing that $f(\mathcal{X}) - 0.5$ assumes values in $[-0.5, 0.5]$; but we will give an alternate proof. We have that

$$\begin{aligned}
E[f, P_\mu] - E[f, P_\nu] &= \sum_{i=1}^{n} f(a_i)(\mu_i - \nu_i) \\
&\leq \sum_{i=1}^{n} f(a_i) \cdot (\mu_i - \nu_i)_+ \text{ since } 0 \leq f(a_i) \ \forall i \\
&\leq \sum_{i=1}^{n} (\mu_i - \nu_i)_+ \text{ since } f(a_i) \leq 1 \ \forall i \\
&= \rho(P_\mu, P_\nu).
\end{aligned}$$

By entirely analogous reasoning, it follows that

$$E[f, P_\mu] - E[f, P_\nu] \geq \sum_{i=1}^{n} (\mu_i - \nu_i)_- = -\rho(P_\mu, P_\nu).$$

The desired bound now follows by combining these two inequalities. □

Problem 2.1 Suppose $\mathcal{X}$ is a 'binary' random variable, assuming just two real values, namely 0 and 1, with $\Pr\{\mathcal{X} = 1\} = \alpha \in (0, 1)$. We denote this by $\mathcal{X} = \mathcal{B}(1, \alpha)$. Compute the mean and standard deviation of $\mathcal{X}$.

Problem 2.2 Suppose an urn contains both white and black balls in the proportion α to $1 - \alpha$.[2] Let $\mathcal{X}$ be the associated binary random variable as defined in Problem 2.1 above. Now suppose we draw n balls from the urn, one after the other, replacing the ball drawn after each trial. Let $\mathcal{Y}_n$ denote the number of white balls drawn after n trials. Then $\mathcal{Y}_n$ is a 'binomially distributed' random variable, whereby

$$\Pr\{Y_n = i\} = \left(\begin{array}{c} n \\ i \end{array} \right) \alpha^i (1-\alpha)^{n-i},$$

where

$$\left(\begin{array}{c} n \\ i \end{array} \right) = \frac{n!}{i!(n-i)!}$$

is called the combinatorial parameter. $\mathcal{Y}_n$ is the number of different sequences of n draws containing precisely i white balls. We denote this by $\mathcal{Y}_n = \mathcal{B}(n, \alpha)$. Compute the mean and standard deviation of $\mathcal{Y}_n$.

In case α is an irrational number, the urn and balls interpretation is not meaningful. Instead we should think in terms of generating n independent outcomes $\mathcal{X}_1, \ldots, \mathcal{X}_n$ where each $\mathcal{X}_i$ is binary with $\Pr\{X_i = 1\} = \alpha$. Such a sequence of random variables is known as a 'Bernoulli process' or a set of 'Bernoulli trials'. If we equate an outcome of 1 with 'success' and 0 with 'failure', then $\mathcal{Y}_n$ is the number of successes in n Bernoulli trials.

Problem 2.3 A counterpart of the binomial distribution is the hypergeometric distribution. Suppose an urn contains N balls, out of which M are white. Now suppose we draw n balls one after the other, but this time *without replacing* the ball drawn. Let $\mathcal{Z}$ denote the resulting number of white balls. Show that

$$\Pr\{\mathcal{Z} = i\} = \frac{\left(\begin{array}{c} M \\ i \end{array} \right) \left(\begin{array}{c} N-M \\ n-i \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}.$$

We denote this by $\mathcal{Z} = \mathcal{H}(n, \alpha.N)$ where $\alpha = M/N$ is the fraction of white balls in the urn.

Problem 2.4 Suppose $\mathcal{Z} = \mathcal{H}(n, \alpha, N)$ have the hypergeometric distribution. Show that, as $N \to \infty$, the hypergeometric distribution approaches the binomial distribution. Can you explain why?

[2] This suggests that α is a rational number, but the problem makes sense even without this assumption.

Problem 2.5 Suppose $\boldsymbol{\mu} = [0.4 \;\; 0.6], \boldsymbol{\nu} = [0.6 \;\; 0.4]$. Compute the total variation distance $\rho(\boldsymbol{\mu}, \boldsymbol{\nu})$.

Problem 2.6 Let $n = 10$, and let $\mathcal{Y}, \mathcal{Z}$ be binomially distributed random variables, with $Y_n = \mathcal{B}(10, 0.6), \mathcal{Z} = \mathcal{B}(10, 0.4)$. Compute the total variation distance between the probability distributions of $\mathcal{Y}$ and $\mathcal{Z}$.

Problem 2.7 Given a finite set $\mathbb{A}$, let $\mathcal{F}(\mathbb{A})$ denote all functions mapping $\mathbb{A}$ into the interval $[0, 1]$. Suppose $\boldsymbol{\mu}, \boldsymbol{\nu}$ are two probability distributions on $\mathbb{A}$. Show that

$$\max_{f \in \mathcal{F}(\mathbb{A})} |E[f, P_\mu] - E[f, P_\nu]| = \rho(P_\mu, P_\nu).$$

Problem 2.8 Prove the following generalization of Lemma 2.9: Suppose $f : \mathbb{A} \to [a, b]$ where a, b are real numbers, and that $\boldsymbol{\mu}, \boldsymbol{\nu}$ are probability distributions on $\mathbb{A}$. Then

$$|E[f, P_\mu] - E[f, P_\nu]| \leq (b - a)\rho(P_\mu, P_\nu).$$

Is this the best possible bound?

2.2 MULTIPLE RANDOM VARIABLES

2.2.1 Joint and Marginal Distributions

Up to now we have discussed only one random variable. It is also possible to have more than one random variable, each assuming values in its own set. Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ are finite sets. Then the cartesian product $\mathbb{A} \times \mathbb{B}$ has cardinality nm and consists of all pairs of the form (a, b). Thus

$$\mathbb{A} \times \mathbb{B} = \{(a, b) : a \in \mathbb{A}, b \in \mathbb{B}\}.$$

Suppose $\boldsymbol{\phi} \in \mathbb{S}_{nm}$. We can think of $\boldsymbol{\phi}$ as the probability distribution of some random variable $\mathcal{Z}$ that assumes values in the set $\mathbb{A} \times \mathbb{B}$. We could of course think of $\mathcal{Z}$ as a random variable that can assume one of nm values. But the fact that the range of values of $\mathcal{Z}$ is a cartesian product allows us to carry out a refined analysis. Because of the product nature of the underlying set $\mathbb{A} \times \mathbb{B}$, we refer to $\mathcal{Z}$ as a **joint random variable** $(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ is a random variable assuming values in $\mathbb{A}$ and $\mathcal{Y}$ is a random variable assuming values in $\mathbb{B}$. The probability distribution $\boldsymbol{\phi}$ on the set $\mathbb{A} \times \mathbb{B}$ is called the **joint distribution** of the variables $\mathcal{X}$ and $\mathcal{Y}$. Thus

$$\Pr\{\mathcal{Z} = (a_i, b_j)\} = \Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j\} = \phi_{ij}, \; \forall a_i \in X, b_j \in Y.$$

So we can arrange the nm elements of ϕ in an array, as shown below.

$$\begin{bmatrix} \phi_{11} & \cdots & \phi_{1m} \\ \vdots & \vdots & \vdots \\ \phi_{n1} & \cdots & \phi_{nm} \end{bmatrix}.$$

Up to now we have gained nothing by arranging the nm values of the probability distribution in an array. But now can take the analysis to another level.

Let us define the vectors $\boldsymbol{\phi}_{\mathcal{X}}$ and $\boldsymbol{\phi}_{\mathcal{Y}}$ as follows.

$$(\boldsymbol{\phi}_{\mathcal{X}})_i := \sum_{j=1}^{m} \phi_{ij}, i = 1, \ldots, n, \tag{2.15}$$

$$(\boldsymbol{\phi}_{\mathcal{Y}})_j := \sum_{i=1}^{n} \phi_{ij}, j = 1, \ldots, m. \tag{2.16}$$

Then it follows that $\boldsymbol{\phi}_{\mathcal{X}} \in \mathbb{S}_n$ and $\boldsymbol{\phi}_{\mathcal{Y}} \in \mathbb{S}_m$. This is because $\boldsymbol{\phi}$ is a probability distribution and as a result we have

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij} = 1 \;\Rightarrow\; \sum_{i=1}^{n}\left[\sum_{j=1}^{m} \phi_{ij}\right] = 1 \;\Rightarrow\; \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i = 1.$$

Similarly

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij} = 1 \;\Rightarrow\; \sum_{j=1}^{m} (\boldsymbol{\phi}_{\mathcal{Y}})_j = 1.$$

So $\boldsymbol{\phi}_{\mathcal{X}} \in \mathbb{S}_n$ and $\boldsymbol{\phi}_{\mathcal{Y}} \in \mathbb{S}_m$. The distribution $\boldsymbol{\phi}_{\mathcal{X}}$ is called the **marginal distribution** of the random variable $\mathcal{X}$. Similarly $\boldsymbol{\phi}_{\mathcal{Y}}$ is called the marginal distribution of the random variable $\mathcal{Y}$. Depending on the context, it may be more natural to write $\boldsymbol{\phi}_{\mathbb{A}}$ in the place of $\boldsymbol{\phi}_{\mathcal{X}}$ and $\boldsymbol{\phi}_{\mathbb{B}}$ in the place of $\boldsymbol{\phi}_{\mathbf{y}}$. Mimicking earlier notation, we refer to the measure corresponding to the distribution $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ as the **joint measure** P_{ϕ} of the joint random variable $(\mathcal{X}, \mathcal{Y})$. We refer to the measure corresponding to the distribution $\boldsymbol{\phi}_{\mathcal{X}} \in \mathbb{S}_n$ as the **marginal measure** of $\mathcal{X}$ (or the marginal measure on the set $\mathbb{A}$), and denote it by $P_{\phi,\mathcal{X}}$ or $P_{\phi,\mathbb{A}}$. The symbols $\mathcal{P}_{\phi,\mathcal{Y}}$ and $P_{\phi,\mathbb{B}}$ are defined analogously.

Now we proceed to show that indeed it is the case that

$$\Pr\{\mathcal{X} = a_i\} = (\boldsymbol{\phi}_{\mathcal{X}})_i, \; \forall i = 1, \ldots, n.$$

To see this, fix an index i and observe that the sets $\{(a_i, b_1)\}$ through $\{(a_i, b_m)\}$ are all pairwise disjoint subsets of $\mathbb{A} \times \mathbb{B}$. Moreover, it is clear that

$$\{(\mathcal{X}, \mathcal{Y}) \in \mathbb{A} \times \mathbb{B} : \mathcal{X} = a_i\} = \bigcup_{j=1}^{m} \{(a_i, b_j)\}.$$

Hence from Property 2 of Theorem 2.3, we can conclude that

$$\begin{aligned}\Pr\{\mathcal{X} = a_i\} &= \Pr\{(\mathcal{X}, \mathcal{Y}) \in \mathbb{A} \times \mathbb{B} : \mathcal{X} = a_i\} \\ &= P_\phi \left(\bigcup_{j=1}^{m} \{(a_i, b_j)\} \right) \\ &= \sum_{j=1}^{m} P_\phi\{(a_i, b_j)\} \\ &= \sum_{j=1}^{m} \phi_{ij} = (\boldsymbol{\phi}_{\mathcal{X}})_i.\end{aligned}$$

By entirely analogous reasoning, it follows that

$$\Pr\{\mathcal{Y} = b_j\} = \sum_{i=1}^{n} \phi_{ij} = (\boldsymbol{\phi}_{\mathcal{Y}})_j, \ \forall j = 1, \ldots, m.$$

2.2.2 Independence, Conditional Distributions

Up to now we have introduced the notion of a joint distribution of the two random variables $\mathcal{X}$ and $\mathcal{Y}$, as well as their individual distributions, which can be obtained as the marginal distributions of the joint distribution. The next notion is perhaps *the* fundamental notion of probability theory.

Definition 2.11 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively. Let P_ϕ denote their joint probability measure, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Then the two random variables are said to be* **independent** *under the measure P_ϕ (or the distribution $\boldsymbol{\phi}$) if*

$$\phi_{ij} = (\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}_{\mathcal{Y}})_j, \ \forall i = 1, \ldots, n, j = 1, \ldots, m. \tag{2.17}$$

An equivalent way of stating (2.17) is:

$$\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j\} = \Pr\{\mathcal{X} = a_i\} \cdot \Pr\{\mathcal{Y} = b_j\}, \ \forall a_i \in \mathbb{A}, b_j \in \mathbb{B}. \tag{2.18}$$

The above definition can be made a little more intuitive by introducing the notion of a 'product' distribution. Suppose $\boldsymbol{\mu} \in \mathbb{S}_n, \boldsymbol{\nu} \in \mathbb{S}_m$ are distributions on the sets $\mathbb{A}, \mathbb{B}$ respectively. Then their **product distribution $\boldsymbol{\mu} \times \boldsymbol{\nu}$** on the set $\mathbb{A} \times \mathbb{B}$ is defined by

$$(\boldsymbol{\mu} \times \boldsymbol{\nu})_{ij} = \mu_i \cdot \nu_j, \ \forall i, j. \tag{2.19}$$

With this definition, it follows that the two random variables $\mathcal{X}, \mathcal{Y}$ are independent under the distribution $\boldsymbol{\phi}$ if and only if $\boldsymbol{\phi} = \boldsymbol{\phi}_{\mathcal{X}} \times \boldsymbol{\phi}_{\mathcal{Y}}$.

In the sequel we will often have occasion to speak about 'independent and identically distributed' random variables. This notion can be made to fit into the above frame work by using product distributions where each of the marginals is the same. Thus if $\boldsymbol{\mu} \in \mathbb{S}_n$, then the product distribution $\boldsymbol{\mu}^2 \in \mathbb{S}_{n^2}$ is defined by

$$(\boldsymbol{\mu}^2)_{ij} = \mu_i \cdot \mu_j, \ \forall i, j.$$

The associated probability measure is often denoted by P_μ^2. Higher 'powers' of $\boldsymbol{\mu}$ and P_μ are defined in an entirely analogous fashion.

Theorem 2.12 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively. Suppose $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ is their joint distribution, that P_ϕ is their joint measure, and that $\mathcal{X}, \mathcal{Y}$ are independent under the measure P_ϕ. Suppose further that $f : \mathbb{A} \to \mathbb{R}, g : \mathbb{B} \to \mathbb{R}$ are functions on $\mathbb{A}$ and $\mathbb{B}$ respectively. Then*

$$E[f(\mathcal{X})g(\mathcal{Y}), P_\phi] = E[f(\mathcal{X}), P_{\phi,\mathcal{X}}] \cdot E[g(\mathcal{Y}), P_{\phi,\mathcal{Y}}]. \tag{2.20}$$

The point of the theorem is this: If f is a function of $\mathcal{X}$ alone and g is a function of $\mathcal{Y}$ alone, then fg is a function of $\mathcal{X}$ and $\mathcal{Y}$ and the pair $(\mathcal{X}, \mathcal{Y})$ has the joint measure P_ϕ. In general, we cannot say anything about the expected value of the function $f(\mathcal{X})g(\mathcal{Y})$ under the measure P_ϕ. However, if the two random variables are *independent* under the measure P_ϕ, then the expected value factors neatly into the product of two different expected values, namely the expected value of f under the marginal measure $P_{\phi,\mathcal{X}}$, and the expected value of g under the marginal measure $P_{\phi,\mathcal{Y}}$.

Proof. The proof is a ready consequence of (2.17). We have

$$\begin{aligned}
E[f(\mathcal{X})g(\mathcal{Y}), P_\phi] &= \sum_{i=1}^{n}\sum_{j=1}^{m} f(a_i)g(b_j)\phi_{ij} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} f(a_i)g(b_j)(\boldsymbol{\phi}_{\mathcal{X}})_i(\boldsymbol{\phi}_{\mathcal{Y}})_j \\
&= \left[\sum_{i=1}^{n} f(a_i)(\boldsymbol{\phi}_{\mathcal{X}})_i\right] \cdot \left[\sum_{j=1}^{m} g(b_j)(\boldsymbol{\phi}_{\mathcal{Y}})_j\right] \\
&= E[f(\mathcal{X}), P_{\phi,\mathcal{X}}] \cdot E[g(\mathcal{Y}), P_{\phi,\mathcal{Y}}].
\end{aligned}$$

This is precisely the desired conclusion. □

The above observation motivates the notion of the correlation coefficient between two real-valued random variables.

Definition 2.13 *Suppose $\mathcal{X}, \mathcal{Y}$ are real-valued random variables assuming values in finite sets $\mathbb{A}, \mathbb{B} \subseteq \mathbb{R}$ respectively. Let $\boldsymbol{\phi}$ denote their joint distribution, and $\boldsymbol{\phi}_{\mathcal{X}}, \boldsymbol{\phi}_{\mathcal{Y}}$ the two marginal distributions. Let $E[\mathcal{X}\mathcal{Y}, \boldsymbol{\phi}]$, $E[\mathcal{X}, \boldsymbol{\phi}_{\mathcal{X}}]$, $E[\mathcal{Y}, \boldsymbol{\phi}_{\mathcal{Y}}]$ denote expectations, and let $\sigma(\mathcal{X}), \sigma(\mathcal{Y})$ denote the standard deviations of $\mathcal{X}, \mathcal{Y}$ under their respective marginal distributions. Then the quantity*

$$C(\mathcal{X}, \mathcal{Y}) := \frac{E[\mathcal{X}\mathcal{Y}, \boldsymbol{\phi}] - E[\mathcal{X}, \boldsymbol{\phi}_{\mathcal{X}}]E[\mathcal{Y}, \boldsymbol{\phi}_{\mathcal{Y}}]}{\sigma(\mathcal{X})\sigma(\mathcal{Y})} \tag{2.21}$$

is called the **correlation coefficient** *between $\mathcal{X}$ and $\mathcal{Y}$.*

Note that some authors refer to $C(\mathcal{X}, \mathcal{Y})$ as the 'Pearson' correlation coefficient after its inventor. It can be shown that the correlation coefficient

$C(\mathcal{X}, \mathcal{Y})$ always lies in the interval $[-1, 1]$. It is often said that $\mathcal{X}, \mathcal{Y}$ are **uncorrelated** if $C(\mathcal{X}, \mathcal{Y}) = 0$, **positively correlated** if $C(\mathcal{X}, \mathcal{Y}) > 0$, and **negatively correlated** if $C(\mathcal{X}, \mathcal{Y}) < 0$. One of the advantages of the correlation coefficient is that it is invariant under both scaling and centering. In other words, if $\alpha, \beta, \gamma, \delta$ are any real numbers, then

$$C(\alpha\mathcal{X} + \beta, \gamma\mathcal{Y} + \delta) = C(\mathcal{X}, \mathcal{Y}). \tag{2.22}$$

If two random variables are independent, then they are uncorrelated. However, the converse statement is most definitely not true; see Problem 2.11.

The next definition is almost as important as the notion of independence.

Definition 2.14 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. The* **conditional probability** *of $\mathcal{X}$ given an observation $\mathcal{Y} = b_j$ is defined as*

$$\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j\} := \frac{\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j\}}{\Pr\{Y = b_j\}} = \frac{\phi_{ij}}{\sum_{i'=1}^{n} \phi_{i'j}}. \tag{2.23}$$

In case $\Pr\{Y = b_j\} = 0$, *we define* $\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j\} = \Pr\{\mathcal{X} = a_i\} = (\boldsymbol{\phi}_{\mathcal{X}})_i$.

Let us use the notation $\phi_{\{a_i|b_j\}}$ as a shorthand for $\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j\}$. Then the vector

$$\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}} := [\phi_{\{a_1|b_j\}} \ldots \phi_{\{a_n|b_j\}}] \in \mathbb{S}_n. \tag{2.24}$$

This is obvious from (2.23). So $\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}$ is a probability distribution on the set $\mathbb{A}$; it is referred to as the **conditional distribution** of $\mathcal{X}$, given that $\mathcal{Y} = b_j$. The corresponding probability measure is denoted by $P_{\phi,\{\mathcal{X}|\mathcal{Y}=b_j\}}$ and is referred to as the **conditional measure** of $\mathcal{X}$, given that $\mathcal{Y} = b_j$. We also use the simplified notation $\boldsymbol{\phi}_{\mathcal{X}|b_j}$ and $P_{\phi,\mathcal{X}|b_j}$ if the variable $\mathcal{Y}$ is clear from the context.

Now we briefly introduce the notion of convex combinations of vectors; we will discuss this idea in greater detail in Section 3.1. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are n-dimensional vectors and $\lambda \in [0, 1]$, then the vector $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$ is called a **convex combination** of $\mathbf{x}$ and $\mathbf{y}$. More generally, if $\mathbf{x}_1, \ldots, \mathbf{x}_l \in \mathbb{R}^n$ and $\boldsymbol{\lambda} \in \mathbb{S}_l$, then the vector $\sum_{i=1}^{l} \lambda_i \mathbf{x}_i$ is called a **convex combination** of the vectors $\mathbf{x}_1$ through $\mathbf{x}_l$. In the present context, it is easy to see that

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{m} (\boldsymbol{\phi}_{\mathcal{Y}})_j \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}. \tag{2.25}$$

Thus, the marginal distribution $\boldsymbol{\phi}_{\mathcal{X}}$ is a convex combination of the m conditional distributions $\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}, j = 1, \ldots, m$. The proof of (2.25) is a straightforward consequence of the definitions and is left as an exercise.

Thus far we have introduced a lot of terminology and notation, so let us recapitulate. Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables, assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Then they have a *joint probability measure* P_ϕ, defined on the product set $\mathbb{A} \times \mathbb{B}$. Associated with P_ϕ is a *marginal probability* $P_{\phi,\mathcal{X}}$, which is a measure on $\mathbb{A}$, and a *marginal probability* $P_{\phi,\mathcal{Y}}$, which is a measure on $\mathbb{B}$. Finally, for each of the m possible values of $\mathcal{Y}$, there is an associated *conditional probability* $P_{\phi,\{\mathcal{X}|\mathcal{Y}=b_j\}}$, which is a measure on $\mathbb{A}$. Similarly, for each of the n possible values of $\mathcal{X}$, there is an associated *conditional probability* $P_{\phi,\{\mathcal{Y}|\mathcal{X}=a_i\}}$, which is a measure on $\mathbb{B}$.

Example 2.6 Let us return to the problem studied earlier of an urn containing four uniform balls with colors red, blue, green and yellow. Suppose we draw two balls from the urn, one after the other, but *without* replacing the first ball before drawing the second ball. Let $\mathcal{X}$ denote the color of the first ball, and $\mathcal{Y}$ the color of the second ball. We can ask: What is the probability of drawing a red ball the second time? The answer is somewhat counter-intuitive because, as shown below, the answer is 0.25. We know that, when we make the second draw, there are only three balls in the urn, and which three colors they represent depends on $\mathcal{X}$, the outcome of the first draw. Nevertheless, the probability of drawing a red ball (or any other colored ball) turns out to be 0.25, as is shown next.

Let us first compute the marginal or 'unconditional' distribution of $\mathcal{X}$, the outcome of the first draw. Since the balls are assumed to be uniform and there are four balls when we draw for the first time, we can define $\mathbb{A} = \{R, B, G, Y\}$ and with this definition the distribution $\boldsymbol{\phi}_\mathcal{X}$ of $\mathcal{X}$ is given by

$$\boldsymbol{\phi}_\mathcal{X} = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25].$$

Now let us compute the conditional probability of $\mathcal{Y}$ given $\mathcal{X}$. If $\mathcal{X} = R$, then at the second draw there are only B, G, Y in the urn. So we can say that

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=R\}} = [0 \;\; 1/3 \;\; 1/3 \;\; 1/3].$$

Similarly,

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=B\}} = [1/3 \;\; 0 \;\; 1/3 \;\; 1/3],$$

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=G\}} = [1/3 \;\; 1/3 \;\; 0 \;\; 1/3],$$

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=Y\}} = [1/3 \;\; 1/3 \;\; 1/3 \;\; 0].$$

Therefore

$$\begin{aligned}\Pr\{\mathcal{Y} = R\} = {} & \Pr\{\mathcal{Y} = R|\mathcal{X} = R\} \cdot \Pr\{\mathcal{X} = R\} + \cdots \\ & + \Pr\{\mathcal{Y} = R|\mathcal{X} = Y\} \cdot \Pr\{\mathcal{X} = Y\},\end{aligned}$$

and so on for the other three colors. Doing this routine calculation shows that

$$\boldsymbol{\phi}_\mathcal{Y} = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25].$$

This somewhat counter-intuitive result can be explained as follows: When we make the second draw to determine $\mathcal{Y}$, there are indeed only three balls in the urn. However, *which three* they are depends on $\mathcal{X}$, the outcome of the first draw. There are four possible sets of three colors, and each of them is equally likely. Hence the probability of getting a red ball the first time is exactly the same as the probability of getting a red ball the second time, even though we are not replacing the first ball drawn.

Example 2.7 The purpose of this example is to show that it is necessary to verify the condition (2.26) for *every* possible value b_j. Suppose $\mathbb{A} = \{a_1, a_2\}$, $\mathbb{B} = \{b_1, b_2, b_3\}$, and that the joint probability distribution is

$$[\phi_{ij}] = \begin{bmatrix} 0.12 & 0.08 & 0.20 \\ 0.20 & 0.10 & 0.30 \end{bmatrix}.$$

Then it follows from (2.15) and (2.16) that

$$\boldsymbol{\phi}_{\mathcal{X}} = [0.4 \ \ 0.6] \text{ and } \boldsymbol{\phi}_{\mathcal{Y}} = [0.32 \ \ 0.18 \ \ 0.50].$$

It can be readily checked that the condition (2.26) holds when $j = 3$ but not when $j = 1$ or $j = 2$. Hence the variables $\mathcal{X}$ and $\mathcal{Y}$ are not independent.

Lemma 2.15 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Then $\mathcal{X}$ and $\mathcal{Y}$ are independent if and only if*

$$\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}} = \boldsymbol{\phi}_{\mathcal{X}}, \ \forall b_j \in Y. \tag{2.26}$$

There is an apparent asymmetry in the statement of Lemma 2.15. It appears as though we should say '$\mathcal{X}$ is independent of $\mathcal{Y}$ if (2.26) holds' as opposed to '$\mathcal{X}$ and $\mathcal{Y}$ are independent if (2.26) holds.' It is left as an exercise to show that (2.26) is equivalent to the statement

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} = \boldsymbol{\phi}_{\mathcal{Y}}, \ \forall a_i \in X. \tag{2.27}$$

Lemma 2.16 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Then $\mathcal{X}$ and $\mathcal{Y}$ are independent if and only if the matrix*

$$\Phi := \begin{bmatrix} \phi_{11} & \ldots & \phi_{1m} \\ \vdots & \vdots & \vdots \\ \phi_{n1} & \ldots & \phi_{nm} \end{bmatrix}$$

has rank one.

The proof is easy and is left as an exercise.

In the preceding discussion, there is nothing special about having *two* random variables – we can have any finite number of them. We can also condition the probability distribution on multiple events, and the results are consistent. To illustrate, suppose $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are random variables assuming values the finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$, $\mathbb{B} = \{b_1, \ldots, b_m\}$, $\mathbb{C} = \{c_1, \ldots, c_l\}$

respectively. Let P_θ denote their joint probability measure and $\boldsymbol{\theta} = [\theta_{ijk}] \in \mathbb{S}_{nml}$ their joint probability distribution. Then

$$\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j \& \mathcal{Z} = c_k\} = \frac{\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j \& \mathcal{Z} = c_k\}}{\Pr\{\mathcal{Y} = b_j \& \mathcal{Z} = c_k\}}. \tag{2.28}$$

In the shorthand notation introduced earlier, this becomes

$$\theta_{\{a_i | b_j \& c_k\}} = \left[\frac{\theta_{ijk}}{\sum_{i'=1}^{n} \theta_{i'jk}}, i = 1, \ldots, n \right] \in \mathbb{S}_n. \tag{2.29}$$

When there are three random variables, the 'law of iterated conditioning' applies, namely:

$$\boldsymbol{\theta}_{\{\mathcal{X}|\mathcal{Y}=b_j \& \mathcal{Z}=b_k\}} = \boldsymbol{\theta}_{\{\{\mathcal{X}\&\mathcal{Y}|\mathcal{Z}=c_k\}|\mathcal{Y}=b_j\}}. \tag{2.30}$$

In other words, in order to compute the conditional distribution of $\mathcal{X}$ given that $\mathcal{Y} = b_j$ and $\mathcal{Z} = c_k$, we can think of two distinct approaches. First, we can directly apply (2.28). Second, we can begin by computing the joint conditional distribution of $\mathcal{X}\&\mathcal{Y}$ given that $\mathcal{Z} = c_k$, and then condition this distribution of $\mathcal{Y} = b_j$. Both approaches give the same answer.

The proof of (2.30) is straightforward. To begin with, we have

$$\boldsymbol{\theta}_{\{\mathcal{X}\&\mathcal{Y}|\mathcal{Z}=c_k\}} = \left[\frac{\theta_{ijk}}{\sum_{i'=1}^{n} \sum_{j'=1}^{m} \theta_{i'j'k}}, i = 1, \ldots, n, j = 1, \ldots, m \right] \in \mathbb{S}_{nm}. \tag{2.31}$$

To make this formula less messy, let us define

$$\phi_k := \sum_{i'=1}^{n} \sum_{j'=1}^{m} \theta_{i'j'k}.$$

Then

$$\boldsymbol{\theta}_{\{\mathcal{X}\&\mathcal{Y}|\mathcal{Z}=c_k\}} = \left[\frac{\theta_{ijk}}{\phi_k}, i = 1, \ldots, n, j = 1, \ldots, m \right].$$

If we now condition this joint distribution of $\mathcal{X}\&\mathcal{Y}$ on $\mathcal{Y} = b_j$, we get

$$\begin{aligned} \boldsymbol{\theta}_{\{\{\mathcal{X}\&\mathcal{Y}|\mathcal{Z}=c_k\}|\mathcal{Y}=b_j\}} &= \left[\frac{\theta_{ijk}/\phi_k}{\sum_{i'=1}^{n} \theta_{i'jk}/\phi_k}, \mathbf{i} = 1, \ldots, n \right] \\ &= \left[\frac{\theta_{ijk}}{\sum_{i'=1}^{n} \theta_{i'jk}}, i = 1, \ldots, n \right], \end{aligned}$$

which is the same as (2.29).

From Definition 2.11, the following observation follows readily.

The next notion introduced is conditional independence, which is very important in the case of hidden Markov processes, which are a central theme of this book.

Definition 2.17 *Suppose $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_m\}$, $\mathbb{B} = \{b_1, \ldots, b_m\}$ and $\mathbb{C} = \{c_1, \ldots, c_l\}$ respectively. Then we say that $\mathcal{X}, \mathcal{Y}$ are* **conditionally independent given** *$\mathcal{Z}$ if, for all $c_k \in \mathbb{C}, b_j \in \mathbb{B}, a_i \in \mathbb{A}$, we have*

$$\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j | \mathcal{Z} = c_k\} = \Pr\{\mathcal{X} = a_i | \mathcal{Z} = c_k\} \cdot \Pr\{\mathcal{Y} = b_j | \mathcal{Z} = c_k\} \tag{2.32}$$

Example 2.8 Consider three random variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, each assuming values in $\{0, 1\}$. Suppose the joint distribution of the three variables is given by

$$\begin{bmatrix} \phi_{000} & \phi_{001} \\ \phi_{010} & \phi_{011} \end{bmatrix} = \begin{bmatrix} 0.018 & 0.056 \\ 0.042 & 0.084 \end{bmatrix}, \begin{bmatrix} \phi_{100} & \phi_{101} \\ \phi_{110} & \phi_{111} \end{bmatrix} = \begin{bmatrix} 0.096 & 0.192 \\ 0.224 & 0.288 \end{bmatrix},$$

where ϕ_{ijk} denotes $\Pr\{\mathcal{X} = i \& \mathcal{Y} = j \& \mathcal{Z} = k\}$. It is now shown that $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$. This is achieved by verifying (2.32).

From the given data, we can compute the joint distributions of $\mathcal{X}\&\mathcal{Z}$, and of $\mathcal{Y}\&\mathcal{Z}$. This gives

$$\begin{array}{c} \mathcal{X}\&\mathcal{Z} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} \quad 0 \quad & \quad 1 \quad \end{array} \\ \begin{bmatrix} \phi_{000} + \phi_{010} & \phi_{001} + \phi_{011} \\ \phi_{100} + \phi_{110} & \phi_{101} + \phi_{111} \end{bmatrix} \end{array} = \begin{array}{c} \mathcal{X}\&\mathcal{Z} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.06 & 0.14 \\ 0.32 & 0.48 \end{bmatrix} \end{array}$$

Hence

$$\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Z}=0\}} = \frac{1}{0.38} \begin{bmatrix} 0.06 \\ 0.32 \end{bmatrix}, \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Z}=1\}} = \frac{1}{0.62} \begin{bmatrix} 0.14 \\ 0.48 \end{bmatrix}.$$

An entirely similar computation yields that

$$\begin{array}{c} \mathcal{Y}\&\mathcal{Z} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} \quad 0 \quad & \quad 1 \quad \end{array} \\ \begin{bmatrix} \phi_{000} + \phi_{100} & \phi_{001} + \phi_{101} \\ \phi_{010} + \phi_{110} & \phi_{011} + \phi_{111} \end{bmatrix} \end{array} = \begin{array}{c} \mathcal{Y}\&\mathcal{Z} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.114 & 0.248 \\ 0.266 & 0.372 \end{bmatrix} \end{array}$$

Hence

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{Z}=0\}} = \frac{1}{0.38}[0.114 \;\; 0.266] = [0.3 \;\; 0.7],$$

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{Z}=1\}} = \frac{1}{0.62}[0.248 \;\; 0.372] = [0.4 \;\; 0.6].$$

Next, let us compute the joint distribution of $\mathcal{X}\&\mathcal{Y}$ conditioned on $\mathcal{Z}$. From either of the above computations, it is clear that the marginal distribution of $\mathcal{Z}$ is given by

$$\boldsymbol{\phi}_{\mathcal{Z}} = [0.38 \;\; 0.62].$$

Therefore the joint distribution of $\mathcal{X}\&\mathcal{Y}$ conditioned on $\mathcal{Z}$ can be computed using (2.31). This gives

$$\begin{array}{c} \mathcal{X}\&\mathcal{Y}|\mathcal{Z}=0 \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} \quad 0 \quad & \quad 1 \quad \end{array} \\ \begin{bmatrix} 0.018/0.38 & 0.042/0.38 \\ 0.096/0.38 & 0.224/0.38 \end{bmatrix} \end{array} = \frac{1}{0.38} \begin{bmatrix} 0.06 \\ 0.32 \end{bmatrix} [0.3 \;\; 0.7]$$

$$= \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Z}=0\}} \times \boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{Z}=0\}}.$$

Similarly,

$$\begin{array}{c} \mathcal{X}\&\mathcal{Y}|\mathcal{Z}=1 \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} \quad 0 \quad & \quad 1 \quad \end{array} \\ \begin{bmatrix} 0.056/0.62 & 0.084/0.62 \\ 0.192/0.62 & 0.288/0.62 \end{bmatrix} \end{array} = \frac{1}{0.62} \begin{bmatrix} 0.14 \\ 0.48 \end{bmatrix} [0.4 \;\; 0.6]$$

$$= \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Z}=1\}} \times \boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{Z}=1\}}.$$

Hence $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$.

Note that, if we are not 'given $\mathcal{Z}$,' then $\mathcal{X}$ and $\mathcal{Y}$ are *not* independent. From earlier discussion, it follows that the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$ is given by

$$\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j\} = \sum_{k=1}^{l} \Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j | \mathcal{Z} = c_k\} \cdot \Pr\{\mathcal{Z} = c_k\}.$$

So if we were to write the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$ in a matrix, then it would be

$$\begin{bmatrix} 0.018 & 0.042 \\ 0.096 & 0.224 \end{bmatrix} + \begin{bmatrix} 0.056 & 0.084 \\ 0.192 & 0.288 \end{bmatrix} = \begin{bmatrix} 0.074 & 0.126 \\ 0.288 & 0.512 \end{bmatrix},$$

where the rows correspond to the values of $\mathcal{Y}$ and the columns correspond to the values of $\mathcal{X}$. Since this matrix does not have rank one, $\mathcal{X}$ and $\mathcal{Y}$ are not independent. The point is that a convex combination of rank one matrices need not be of rank one.

Once we have the notion of a conditional distribution, the notion of conditional expected value is natural. Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in $\mathbb{A}$ and $\mathbb{B}$ respectively, and suppose $f : \mathbb{A} \rightarrow \mathbb{R}$ is some real-valued function. Let $\boldsymbol{\phi}$ denote the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$. Then the 'unconditional' expected value of f is denoted by $E[f, P_{\phi,\mathcal{X}}]$ or less cumbersomely by $E[f, \boldsymbol{\phi}_{\mathcal{X}}]$, and is defined as

$$E[f, \boldsymbol{\phi}_{\mathcal{X}}] = \sum_{i=1}^{n} f(a_i)(\boldsymbol{\phi}_{\mathcal{X}})_i.$$

The 'conditional' expected value of f is denoted by $E[f, P_{\phi,\{\mathcal{X}|\mathcal{Y}=b_j\}}]$ or less cumbersomely by $E[f, \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}]$, and is defined as

$$E[f, \boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}] = \sum_{i=1}^{n} f(a_i)\phi_{\{a_i|b_j\}}.$$

We conclude this subsection by introducing another notion called the conditional expectation of a random variable. The dual usage of the adjective 'conditional' is a source of endless confusion to students. The conditional *expected value* of a random variable (or a function of a random variable) is a *real number*, whereas the conditional *expectation* of a random variable is another *random variable.* Unfortunately, this dual usage is too firmly entrenched in the probability literature for the present author to deviate from it.

Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$, $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Now suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$ is a function of both $\mathcal{X}$ and $\mathcal{Y}$. One can ask: What is the best approximation of $h(\mathcal{X}, \mathcal{Y})$ in terms of a function of $\mathcal{X}$ alone? In other words, we seek a function $f : \mathbb{A} \rightarrow \mathbb{R}$ such that f best approximates h. A natural error criterion is the 'least-squares error,' namely:

$$J(f) = E[f - h, P_\phi] = \sum_{i=1}^{n} \sum_{j=1}^{m} (f_i - h_{ij})^2 \phi_{ij},$$

where we use the shorthand $f_i = f(a_i), h_{ij} = h(a_i, b_j)$. The choice of f that minimizes J is easy to compute. Note that

$$\frac{\partial J}{\partial f_i} = 2 \sum_{j=1}^{m} (f_i - h_{ij}) \phi_{ij}.$$

Hence the optimal choice of f_i is obtained by setting this partial derivative to zero, that is,

$$f_i = \frac{\sum_{j=1}^{m} h_{ij}\phi_{ij}}{\sum_{j=1}^{m} \phi_{ij}} = \frac{\sum_{j=1}^{m} h_{ij}\phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{X}})_i}. \tag{2.33}$$

Hence if we define a function $f : \mathbb{A} \to \mathbb{R}$ by $f(a_i) = f_i$, then $f(\mathcal{X})$ is the best approximation to $h(\mathcal{X}, \mathcal{Y})$ that depends on $\mathcal{X}$ alone. We formalize this idea through a definition.

Definition 2.18 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$, $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Suppose $h : \mathbb{A} \times \mathbb{B} \to \mathbb{R}$. Then the* **conditional expectation of h with respect to $\mathcal{X}$** *is the function $f : \mathbb{A} \to \mathbb{R}$ defined by (2.33), and is denoted by $h_{\mathbb{A}}$ or $h_{\mathcal{X}}$.*

In the above definition, if $(\boldsymbol{\phi}_{\mathcal{X}})_i = 0$ for some index i, then the corresponding value f_i can be assigned arbitrarily. This is because, if $(\boldsymbol{\phi}_{\mathcal{X}})_i = 0$ for some index i, then $\phi_{ij} = 0$ for all j. As a result, we can actually just drop the corresponding element a_i from the set $\mathbb{A}$ and carry on without affecting anything.

Lemma 2.19 *Suppose $h : \mathbb{A} \times \mathbb{B} \to \mathbb{R}_+$. Then $h_{\mathcal{X}} : \mathbb{A} \to \mathbb{R}_+$. Suppose $h : \mathbb{A} \times \mathbb{B} \to [\alpha, \beta]$ for some finite numbers $\alpha < \beta$. Then Then $h_{\mathcal{X}} : \mathbb{A} \to [\alpha, \beta]$.*

Proof. The first part of the lemma says that if the original function h assumes only nonnegative values, then so does its conditional expectation $h_{\mathcal{X}}$. This fact is obvious from the definition (2.33). The second part follows readily upon observing that if $h : \mathbb{A} \times \mathbb{B} \to [\alpha, \beta]$, then both $h - \alpha$ and $\beta - h$ are nonnegative-valued functions. □

A very useful property of the conditional expectation is given next.

Theorem 2.20 *Suppose $h : \mathbb{A} \times \mathbb{B} \to \mathbb{R}$ and that $g : \mathbb{A} \to \mathbb{R}$. Let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote a probability distribution on $\mathbb{A} \times \mathbb{B}$. Then*

$$E[gh, P_\phi] = E[gh_{\mathcal{X}}, P_{\phi,\mathbb{A}}]. \tag{2.34}$$

Proof. This follows from just writing out the expected value as a summation. We have

$$\begin{aligned} E[gh, P_\phi] &= \sum_{i=1}^{n} \sum_{j=1}^{m} g_i h_{ij} \phi_{ij} \\ &= \sum_{i=1}^{n} g_i \sum_{j=1}^{m} h_{ij} \phi_{ij} \\ &= \sum_{i=1}^{n} g_i (h_{\mathbb{A}})_i (\boldsymbol{\phi}_{\mathbb{A}})_i \\ &= E[gh_{\mathcal{X}}, P_{\phi,\mathbb{A}}]. \end{aligned}$$

This is the desired result. □

2.2.3 Bayes' Rule

The next result, known as **Bayes' rule**, is widely used.

Lemma 2.21 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ of cardinality n and m respectively. Then*

$$\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j\} = \frac{\Pr\{\mathcal{Y} = b_j | \mathcal{X} = a_i\} \cdot \Pr\{\mathcal{X} = a_i\}}{\Pr\{\mathcal{Y} = b_j\}}. \tag{2.35}$$

Proof. An equivalent way of writing (2.32) is:

$$\Pr\{\mathcal{X} = a_i | \mathcal{Y} = b_j\} \cdot \Pr\{\mathcal{Y} = b_j\} = \Pr\{\mathcal{Y} = b_j | \mathcal{X} = a_i\} \cdot \Pr\{\mathcal{X} = a_i\}.$$

But this statement is clearly true, since each side is equal to $\Pr\{\mathcal{X} = a_i \& \mathcal{Y} = b_j\}$. □

Example 2.9 A typical use of Bayes' rule is when we try to invert the hypothesis and conclusion, and assess the probability of the resulting statement. To illustrate, suppose there is a diagnostic test for HIV, which is accurate 98% of the time on HIV-positive patients and 99% accurate on HIV-negative patients. In other words, the probability that the test is positive when the patient has HIV is 0.98, while the probability that the test is negative when the patient does not have HIV is 0.99. We may therefore be lulled into thinking that we have a very good test at hand. But the question that really interests us is this: What is the probability that a patient who tests positive actually has HIV?

Let us introduce two random variables: $\mathcal{X}$ for a patient's actual condition, and $\mathcal{Y}$ for the outcome of a test. Thus $\mathcal{X}$ assumes values in the set $X = \{H, F\}$, where H denotes that the patient has HIV, while F denotes that the patient is free from HIV. Similarly, $\mathcal{Y}$ assumes values in the set $Y = \{P, N\}$, where P denotes that the test is positive, while N denotes that the test is negative. The data given thus far can be summarized as follows:

$$\Pr\{\mathcal{Y} = P | \mathcal{X} = H\} = 0.98, \Pr\{\mathcal{Y} = N | \mathcal{X} = F\} = 0.99. \tag{2.36}$$

But what we really want to know is the value of

$$\Pr\{\mathcal{X} = H|\mathcal{Y} = P\},$$

that is, the probability that the patient really has HIV when the test is positive.

To compute this quantity, suppose the fraction of the population that has HIV is 1%. Thus the marginal probability distribution of $\mathcal{X}$ is

$$[\Pr\{\mathcal{X} = H\} \;\; \Pr\{\mathcal{X} = F\}] = [0.01 \;\; 0.99].$$

With this information and (2.36), we can compute the joint distribution of the variables $\mathcal{X}$ and $\mathcal{Y}$. We get

$$\left[\begin{array}{cc} \phi_{\mathcal{X}=H\&\mathcal{Y}=P} & \phi_{\mathcal{X}=H\&\mathcal{Y}=N} \\ \phi_{\mathcal{X}=F\&\mathcal{Y}=P} & \phi_{\mathcal{X}=F\&\mathcal{Y}=N} \end{array}\right] = \left[\begin{array}{cc} 0.0098 & 0.0002 \\ 0.0099 & 0.9801 \end{array}\right].$$

So by adding up the two columns, we get

$$[\Pr\{\mathcal{Y} = P\} \;\; \Pr\{\mathcal{Y} = N\}] = [0.0197 \;\; 0.9803].$$

Hence, by Bayes' rule, we can compute that

$$\Pr\{\mathcal{X} = H|\mathcal{Y} = P\} = \frac{0.0098}{0.0197} \approx 0.5.$$

So actually the diagnostic is quite unreliable, because the likelihood of a patient who tests positive *not* having HIV is just about equal to the likelihood of a patient tests positive actually having HIV.

This apparent paradox is easily explained: For the sake of simplicity, assume that the test is equally accurate both with patients actually having HIV and with patients not having HIV. Let β denote the inaccuracy of the test. Thus

$$\Pr\{\mathcal{Y} = P|\mathcal{X} = H\} = \Pr\{\mathcal{Y} = N|\mathcal{X} = F\} = 1 - \beta.$$

Let α denote the fraction of the population that actually has HIV. We can carry through all of the above computations in symbolic form and obtain

$$\left[\begin{array}{cc} \phi_{\mathcal{X}=H\&\mathcal{Y}=P} & \phi_{\mathcal{X}=H\&\mathcal{Y}=N} \\ \phi_{\mathcal{X}=F\&\mathcal{Y}=P} & \phi_{\mathcal{X}=F\&\mathcal{Y}=N} \end{array}\right] = \left[\begin{array}{cc} \alpha(1-\beta) & \alpha\beta \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) \end{array}\right].$$

So

$$\Pr\{\mathcal{X} = H|\mathcal{Y} = P\} = \frac{\alpha(1-\beta)}{\alpha(1-\beta) + (1-\alpha)\beta}.$$

If, as is reasonable, both α and β are small, we can approximate both $1-\alpha$ and $1-\beta$ by 1, which leads to

$$\Pr\{\mathcal{X} = H|\mathcal{Y} = P\} \approx \frac{\alpha}{\alpha + \beta}.$$

So, unless $\beta \ll \alpha$, we get a test that is pretty useless. On the other hand, if $\beta \ll \alpha$, then $\Pr\{\mathcal{X} = H|\mathcal{Y} = P\}$ is very close to one and we have an excellent diagnostic test. The point to note is that the error of the diagnostic test must be small, not in comparison with 1, but with the likelihood of occurance of the condition that we are trying to detect.

2.2.4 MAP and Maximum Likelihood Estimates

In the previous subsections, we have discussed the issue of computing the probability distribution of one random variable, given an observation of another random variable. Now let us make the question a little more specific, and ask: What is the *most likely value* of one random variable, given an observation of another random variable. It is shown below that there are two distinct ways of formalizing this notion, and each is reasonable in its own way.

Definition 2.22 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Let $\boldsymbol{\phi}$ denote their joint distribution. Then the* **maximum a posteriori (MAP)** *estimate of $\mathcal{X}$ given an observation $\mathcal{Y} = b_j$ is the a_{i^*} such that*

$$\phi_{\{a_{i^*}|b_j\}} = \max_i \phi_{\{a_i|b_j\}}. \tag{2.37}$$

Thus the MAP estimate of $\mathcal{X}$ given an observation $\mathcal{Y} = b_j$ is the most likely value of $\mathcal{X}$ using the conditional distribution $\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}$. Since

$$\phi_{\{a_i|b_j\}} = \frac{\phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{Y}})_j},$$

and the denominator is independent of i, we can see that

$$i^* = \arg\min_i \phi_{ij}.$$

So computing the MAP estimate is very easy. Given an observation $\mathcal{Y} = b_j$, we simply scan down the j-th column of the joint distribution matrix, and pick the row i where the element ϕ_{ij} is the largest. (If there is a tie, we can use any sensible tie-breaking rule.)

The next definition gives an alternate way of defining the 'most likely' value of $\mathcal{X}$.

Definition 2.23 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $Y = \{b_1, \ldots, b_m\}$ respectively. Let $\boldsymbol{\phi}$ denote their joint distribution. Then the* **maximum likelihood estimate (MLE)** *of $\mathcal{X}$ given an observation $\mathcal{Y} = b_j$ is defined as the index i^* such that* $\Pr\{\mathcal{Y} = b_j|\mathcal{X} = a_i\}$ *is maximized when $i = i^*$.*

Thus the MLE of $\mathcal{X}$ given the observation $\mathcal{Y} = b_j$ is the choice of a_i that makes the observed value the most likely one.

The choice between MAP and MLE is essentially dictated by whether we believe that $\mathcal{X}$ 'causes' $\mathcal{Y}$, or vice versa. The joint distribution $\boldsymbol{\phi}$ is strictly neutral, and does not at all address the issue of what causes what. If we believe that $\mathcal{Y}$ causes $\mathcal{X}$, then we should believe that, following the observation $\mathcal{Y} = b_j$, the probability distribution of $\mathcal{X}$ has shifted from the marginal distribution $\boldsymbol{\phi}_{\mathcal{X}}$ to the conditional distribution $\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}$. Thus MAP is the most logical way to estimate $\mathcal{X}$. If on the other hand we believe that $\mathcal{X}$ causes $\mathcal{Y}$, we should choose the MLE of $\mathcal{X}$, because that estimate makes the observation most likely.

Example 2.10 To show that MAP and MLE can lead to diametrically opposite conclusions, consider the case where $n = m = 2$ and the joint distribution of $\mathcal{X}, \mathcal{Y}$ is given by

$$\boldsymbol{\phi} = \begin{bmatrix} 0.1 & 0.2 \\ 0.4 & 0.3 \end{bmatrix},$$

where the rows correspond to the value of $\mathcal{X}$ and the columns to the values of $\mathcal{Y}$. Suppose we observe $\mathcal{Y} = b_2$. Then, by examining the second column of $\boldsymbol{\phi}$, we see that the MAP estimate of $\mathcal{X}$ is a_2, because $\phi_{22} > \phi_{12}$. On the other hand, to compute the MLE of $\mathcal{X}$, we compute

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=x_1\}} = [1/3 \;\; 2/3], \;\; \boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=x_2\}} = [4/7 \;\; 3/7].$$

Thus b_2 is the most likely value of $\mathcal{Y}$ if $\mathcal{X} = a_1$, so the MLE of $\mathcal{X}$ given the observation $\mathcal{Y} = y_2$ is a_1.

Problem 2.9 Prove (2.22).

Problem 2.10 Show that if $\mathcal{X}, \mathcal{Y}$ are independent real-valued random variables, then their correlation coefficient is zero.

Problem 2.11 Suppose the joint distribution of two random variables $\mathcal{X}$ and $\mathcal{Y}$, each of them assuming one of the five values $\{1, 2, 3, 4, 5\}$, is as shown in the table below.

$$\Phi = \left[\begin{array}{c|ccccc} \mathcal{X} \setminus \mathcal{Y} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0.0800 & 0.0260 & 0.0280 & 0.0320 & 0.0340 \\ 2 & 0.0280 & 0.0900 & 0.0300 & 0.0270 & 0.0250 \\ 3 & 0.0260 & 0.0200 & 0.0800 & 0.0340 & 0.0400 \\ 4 & 0.0340 & 0.0300 & 0.0290 & 0.0800 & 0.0270 \\ 5 & 0.0320 & 0.0340 & 0.0330 & 0.0270 & 0.0740 \end{array} \right].$$

Compute the following:

1. The five conditional probability distributions $\boldsymbol{\phi}_{\mathcal{X}|\mathcal{Y}=n\}}$, for $n = 1, \ldots, 5$.
2. The five conditional probability distributions $\boldsymbol{\phi}_{\mathcal{Y}|\mathcal{X}=n\}}$, for $n = 1, \ldots, 5$.
3. The five conditional expected values $E[\mathcal{X}|\mathcal{Y} = n]$ for $n = 1, \ldots, 5$.
4. The five conditional expected values $E[\mathcal{Y}|\mathcal{X} = n]$ for $n = 1, \ldots, 5$.
5. The MAP estimates of $\mathcal{X}$ given that $\mathcal{Y} = n$ for $n = 1, \ldots, 5$.
6. The MAP estimates of $\mathcal{Y}$ given that $\mathcal{X} = n$ for $n = 1, \ldots, 5$.
7. The correlation coefficient $C(\mathcal{X}, \mathcal{Y})$.

Problem 2.12 Prove (2.25).

Problem 2.13 Show that (2.26) and (2.27) are equivalent conditions.

Problem 2.14 Prove Lemma 2.16.

Problem 2.15 Suppose, as in (2.33), that $h : \mathbb{A} \times \mathbb{B} \to \mathbb{R}$. For each $a_i \in \mathbb{A}$, define the function $h_{i\cdot} : \mathbb{B} \to \mathbb{R}$ by

$$h_{i\cdot} = h_{ij}.$$

Show that the expression (2.33) for the conditional expectation of h with respect to $\mathcal{X}$ can be defined as

$$(h_{\mathcal{X}})_i = E[h_{i\cdot}, \phi_{|\mathcal{X}=a_i}].$$

2.3 RANDOM VARIABLES ASSUMING INFINITELY MANY VALUES

Until now we have steadfastly restricted ourselves to random variables that assume values in a *finite* set. However, there are situations in which it is desirable to relax this assumption, and examine situations in which the range of the random variable under study is *infinite*. Within this, we make a further distinction between two situations: Where the range is a *countable* set and where the range is an *uncountable* set. Recall that a set is said to be **countable** if it can be placed in one-to-one correspondence with the set of natural numbers $\mathbb{N} = \{1, 2, \ldots\}$.[3] For example, the set of integers and the set of rational numbers are both countable sets. Next, suppose $\mathcal{M}$ is a finite set, such as $\{H, T\}$, the set of possible outcomes of a coin toss experiment, or $\{A, C, G, T\}$, the set of nucleotides. Let $\mathcal{M}^*$ denote the set of all *finite* sequences taking values in $\mathcal{M}$. Thus $\mathcal{M}^*$ consists of all sequences $\{u_1, \ldots, u_n\}$ where $u_i \in \mathcal{M}$ for all i. Then M^* is countable. But uncountably infinite sets are also relevant. For instance, if $\mathcal{M}$ is a finite set, then the set of *all* sequences (not just finite sequences) taking values in $\mathcal{M}$ is an uncountably infinite set. The set of real numbers is also uncountable.

It turns out that the method adopted thus far to define probabilities over finite sets, namely just to assign nonnegative 'weights' to each element in such a way that the weights add up to one, works perfectly well on countable sets. However, the approach breaks down when the range of the random variable is an uncountably infinite set. The great Russian mathematician A. N Kolmogorov introduced the axiomatic foundations of probability theory precisely to cope with this situation; see [71]. Though the theory is very beautiful and comprehensive, we will not be needing the more advanced theory in the present book.

[3] Some authors also include 0 in $\mathbb{N}$.

Accordingly, suppose $X = \{x_i, i \in \mathbb{N}\}$ is a countable set. Let $p_i \geq 0$ be chosen such that $\sum_{i=1}^{\infty} p_i = 1$. Then for every subset $A \subseteq X$, we can define the corresponding probability measure in analogy with (2.2), namely

$$P(A) := \sum_{i=1}^{\infty} I_A(x_i) p_i.$$

We can think of $P(A)$ as the probability $\Pr\{\mathcal{X} \in A\}$ that the random variable $\mathcal{X}$ belongs to the set A. With this definition, the 'axiomatic' properties described just after Theorem 2.3 continue hold, namely:

1. $0 \leq P(A) \leq 1$ for all subsets $A \subseteq X$.
2. If $\{A_i\}_{i \geq 1}$ is a countable collection of pairwise disjoint subsets of X, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

For a random variable $\mathcal{X}$ assuming values in the countable set X with the probability measure P defined above, we can define its mean, variance, and higher moments just as we did earlier for finite-valued random variables. Thus, if $f : X \to \mathbb{R}$ is a function, then we define

$$E[f, P] := \sum_{i=1}^{\infty} f(x_i) p_i.$$

The major potential difficulty is that, because we are dealing with an *infinite sum*, the above expected value is not guaranteed to exist. In particular, suppose we try to compute the mean value of the random variable $\mathcal{X}$ using the above definition, as

$$E[\mathcal{X}, P] = \sum_{i=1}^{\infty} x_i p_i.$$

Without loss of generality, we can renumber the x_i's in such a way that $x_i < x_{i+1}$ for all i. Since the p_i's add up to one, it is obvious that $p_i \to 0$ as $i \to \infty$. However, depending on the relationship between x_i and p_i, the above summation may or may not converge. In case the expectation of the random variable $\mathcal{X}$ is not defined, it is said to be a 'heavy-tailed' random variable. The next example illustrates one such variable.

Example 2.11 This example is sometimes referred to as the 'St. Petersberg paradox.' Suppose a gambler visits a casino where he plays a coin-tossing game. At each step, both the gambler and the casino put up equal stakes, after which the gambler predicts the outcome of the coin toss. If he calls correctly, he gets the entire stake, whereas if he calls incorrectly, the casino gets the entire stake. The game is fair, with the coin turning up heads or tails with equal probability. Moreover, each coin toss is independent of all previous coin tosses. To simplify the notation, let us suppose that the gambler always calls heads. In view of the independence assumption, this strategy has just as good a chance of winning as any other.

Now the following strategy is 'guaranteed' to fetch a positive payoff to the gambler: At each step, he merely doubles his stake. Thus, at the first step, he bets \$1. If he wins, he quits and goes home. If he loses, he bets \$2. If he loses again, he bets \$4 at the next step, and so on. The game reaches n steps only if the gambler has lost all n times, meaning that his accumulated losses are \$ $1 + 2 + \ldots 2^n = 2^{n+1} - 1$. At the $(n+1)$-st step, he bets 2^{n+1}. If he wins, his cumulative winning amount is precisely \$1, the inital bet.

One feels that there is something strange about this strategy; indeed the difficulty is that the random variable in this case is heavy-tailed and does not have a finite expectation. We can see that the game has a countable set of possible outcomes, namely $H, TH, T^2H, \ldots, T^nH, \ldots$, where T^nH denotes a sequence of n tails followed by a head. The probability of T^nH is obviously $2^{-(n+1)}$ because the coin is fair. In this case, the accumulated losses at time n are $2^{n+1} - 1$. Thus, in order to bet 2^{n+1} at the next step, the gambler must have an initial sum of $2^{n+2} - 1$. Therefore, the amount of money that the player must have to begin with, call it $\mathcal{X}$, equals $2^{n+2} - 1$ with probability $2^{-(n+1)}$. If we try to compute the expected value of this random variable, we see that

$$\sum_{i=1}^{n} (2^{n+2} - 1) \cdot 2^{-(n+1)} = \sum_{i=1}^{n} 2 - 2^{-(n+1)}.$$

Now, as $n \to \infty$, the second summation converges nicely to -1. Unfortunately, the first summation blows up. Hence, unless one has an infinite amount of money to begin with, the above 'strategy' will not work.

2.4 TAIL PROBABILITY ESTIMATES: MARKOV AND CHEBYCHEFF INEQUALITIES

In this section, we introduce two very useful inequalities known as 'Markov's inequality' and 'Chebycheff's inequality'. We use the formalism of a random variable assuming real values in order to state the result. Since the set of real numbers is uncountably infinite, the earlier approach of assigning a weight to each possible outcome does not work, and we need to adopt a different approach. What follows is a very superficial introduction to the subject, and a reader interested in a serious discussion of the subject is referred to any of the classic texts on the subject, such as [22, 19] for example.

With each real-valued random variable $\mathcal{X}$ we associate a so-called **probability distribution function (pdf)** $P_{\mathcal{X}}$, defined as follows:

$$P_{\mathcal{X}}(a) = \Pr\{\mathcal{X} \leq a\}, \ \forall a \in \mathbb{R}.$$

The pdf is monotonically nondecreasing, as is obvious from the definition; thus

$$a \leq b \ \Rightarrow \ P_{\mathcal{X}}(a) \leq P_{\mathcal{X}}(b).$$

The pdf also has a property known as 'cadlag,' which is an abbreviation of the French expression 'continué à droite, limité à gauche.' In English this means 'continuous from the right, and limit exists from the left.' In other words, the function $P_{\mathcal{X}}$ has the property that, for each real number a,

$$\lim_{x \to a^+} P_{\mathcal{X}}(x) = P_{\mathcal{X}}(a),$$

while $\lim_{x \to a^-} P_{\mathcal{X}}(x)$ exists, but may or may not equal $P_{\mathcal{X}}(a)$. Due to the monotonicity of the pdf, it is clear that

$$\lim_{x \to a^+} P_{\mathcal{X}}(x) \leq P_{\mathcal{X}}(a).$$

If the above holds with equality, then $P_{\mathcal{X}}$ is continuous at a. Otherwise it has a positive jump equal to the difference between $P_{\mathcal{X}}(a)$ and the limit on the left side.

In general the function $P_{\mathcal{X}}$ need not be differentiable, or even continuous. However, for the purposes of the present discussion, it is sufficient to consider the case where $P_{\mathcal{X}}$ is continuously differentiable everywhere, except for a countable set of points $\{x_i\}_{i=1}^{\infty}$, where the function has a jump. Thus

$$\lim_{x \to x_i^-} P_{\mathcal{X}}(x) < P_{\mathcal{X}}(x_i),$$

but $P_{\mathcal{X}}(\cdot)$ is continuously differentiable at all $x \neq x_i$. In such a case, the difference

$$P_{\mathcal{X}}(x_i) - \lim_{x \to x_i^-} P_{\mathcal{X}}(x) =: \mu_i$$

can be interpreted as the (nonzero) probability that the random variable $\mathcal{X}$ *exactly equals* x_i. For all other values of x, it is interpreted that the probability of the random variable $\mathcal{X}$ *exactly* equaling x is zero. However, if $a < b$, then the probability of the random variable $\mathcal{X}$ lying in the interval $(a, b]$ is taken as $P_{\mathcal{X}}(b) - P_{\mathcal{X}}(a)$.

To define the expected value of the random variable $\mathcal{X}$, we adapt the earlier formulation to the present situation. To simplify notation, let $P(\cdot)$ denote the pdf of $\mathcal{X}$. Then we define

$$E[\mathcal{X}, P] = \int_{-\infty}^{\infty} x P(dx),$$

where the integral is a so-called Riemann-Stiltjes integral. If P is continuously differentiable over some interval $[a, b]$, we define

$$\int_a^b f(x) P(dx) = \int_a^b f(x) \frac{dP}{dx} dx,$$

and add the term $f(x_i)\mu_i$ whenever the interval $[a, b]$ contains one of the points x_i where $P_{\mathcal{X}}$ has a jump discontinuity. As before, the existence of the expected value is not guaranteed.

Theorem 2.24 (Markov's Inequality) *Suppose $\mathcal{X}$ is a real-valued random variable with the property that $|\mathcal{X}|$ has finite expectation. Then, for every real number a, we have*

$$\Pr\{|\mathcal{X}| \geq a\} \leq \frac{E[|\mathcal{X}|, P]}{a}. \tag{2.38}$$

Proof. By definition, we have

$$\begin{aligned} E[|\mathcal{X}|, P] &= \int_{|x|<a} |x| P(dx) + \int_{|x|\geq a} |x| P(dx) \\ &\geq \int_{|x|\geq a} |x| P(dx) \\ &\geq \int_{|x|\geq a} a P(dx) \\ &= a \Pr\{|\mathcal{X}| \geq a\}. \end{aligned}$$

The desired inequality follows by dividing both sides by a. □

Corollary 2.25 *Suppose $\mathcal{X}$ is a nonnegative-valued random variable with finite expectation. Then, for every real number a, we have*

$$\Pr\{\mathcal{X} \geq a\} \leq \frac{E[\mathcal{X}, P]}{a}. \tag{2.39}$$

The proof is entirely analogous to that of Theorem 2.24.

Markov's inequality in the above form is not particularly useful. However, we get a more useful version if we examine a function of $\mathcal{X}$.

Corollary 2.26 *Suppose $\mathcal{X}$ is a real-valued random variable. Then for every $\epsilon > 0, \gamma \geq 0$, we have*

$$\Pr\{\mathcal{X} \geq \epsilon\} \leq \exp(-\gamma\epsilon) E[\exp(\gamma\mathcal{X}), P], \ \forall \gamma \geq 0, \tag{2.40}$$

provided only that $\exp(\gamma\mathcal{X})$ *has finite expectation.*

Proof. Note that, whenever $\gamma \geq 0$, the function $x \mapsto \exp(\gamma x)$ is nonnegative-valued and nondecreasing. Hence, for every $\epsilon > 0$, we have

$$\mathcal{X} \geq \epsilon \ \Leftrightarrow \ \exp(\gamma\mathcal{X}) \geq \exp(\gamma\epsilon).$$

Now apply (2.39) with $\mathcal{X}$ replaced by $\exp(\gamma\mathcal{X})$ and a replaced by $\exp(\gamma\epsilon)$. □

There is a variant of Markov's inequality for random variables that have not only finite expectation but also finite variance. As before, we define the variance of $\mathcal{X}$ as

$$\text{var}(\mathcal{X}) := E[(\mathcal{X} - E(\mathcal{X}))^2],$$

assuming it exists of course. It is common to denote the variance by σ^2, so that σ is the standard deviation. With this notation, we now state the next result.

Theorem 2.27 (Chebycheff's Inequality) *Suppose $\mathcal{X}$ is a real-valued random variable with finite expectation and variance. Then for each $\epsilon > 0$, we have*

$$\Pr\{|\mathcal{X} - E(\mathcal{X})| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \tag{2.41}$$

Proof. We reason as follows:

$$\begin{aligned}\Pr\{|\mathcal{X} - E(\mathcal{X})| \geq \epsilon\} &= \Pr\{(\mathcal{X} - E(\mathcal{X}))^2 \geq \epsilon^2\} \\ &\leq \frac{E[(\mathcal{X} - E(\mathcal{X}))^2]}{\epsilon^2} \text{ by (2.39)} \\ &= \frac{\sigma^2}{\epsilon^2}.\end{aligned}$$

□

Chapter Three

Introduction to Information Theory

In this chapter, we introduce a very important notion, called the 'entropy' of a probability distribution. One can think of entropy as the level of uncertainty associated with a random variable (or more precisely, the probability distribution of the random variable). Entropy has several useful properties, and the relevant ones are brought out here. The next concept is relative entropy, also known as the Kullback-Leibler divergence, named after the two statisticians who invented the notion. The Kullback-Leibler divergence measures the 'disparity' between two probability distributions, and has a very useful interpretation in terms of the rate at which one can learn to discriminate between the correct and an incorrect hypothesis, when there are two competing hypotheses. To lay the foundation to introduce these concepts, we begin with the notion of convexity, which has many applications that go far beyond the few that are discussed in this book.

3.1 CONVEX AND CONCAVE FUNCTIONS

In this section we introduce a very useful 'universal' concept known as convexity (and its mirror image, concavity). Though we make use of this concept in a very restricted setting (namely, to study the properties of the entropy function), the concept itself has many applications. We begin with the notion of a convex set, and then move to the notion of a convex (or concave) function.

If x, y are real numbers, and $\lambda \in [0,1]$, the number $\lambda x + (1-\lambda)y$ is called a **convex combination** of x and y. More generally, if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are n-dimensional vectors and $\lambda \in [0,1]$, the vector $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is called a **convex combination** of $\mathbf{x}$ and $\mathbf{y}$. If $\lambda \in (0,1)$ and $\mathbf{x} \neq \mathbf{y}$, then the vector $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is called a **strict convex combination** of $\mathbf{x}$ and $\mathbf{y}$.

Definition 3.1 *A subset $S \subseteq \mathbb{R}^n$ is said to be a* **convex set** *if*

$$\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in S \ \forall \lambda \in [0,1], \ \forall \mathbf{x}, \mathbf{y} \in S. \tag{3.1}$$

Thus a set S is convex if every convex combination of two elements of S once again belongs to S. In two dimensions $n = 2$, we can visualize a convex set very simply. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, then the set $\{\lambda\mathbf{x} + (1-\lambda)\mathbf{y} : \lambda \in [0,1]\}$ is the line segment joining the two vectors $\mathbf{x}$ and $\mathbf{y}$. Thus a set $S \subseteq \mathbb{R}^2$ is convex if and only if the line segment joining any two points in the set S once again

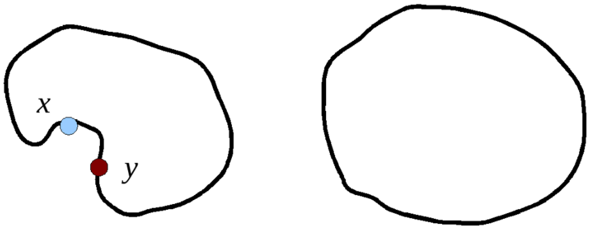


Figure 3.1 Examples of convex and nonconvex sets

belongs to the set S. As seen in Figure 3.1 below, the set on the left is not convex, because the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ does not lie entirely in S; in contrast, the set on the right is convex.

Example 3.1 The n-dimensional simplex $\mathbb{S}_n$, which is where every n-dimensional probability distribution has to 'live,' is a convex set. Thus if $\mathbf{p}, \mathbf{q}$ are n-dimensional probability distributions, then so is the convex combination $\lambda\mathbf{p} + (1-\lambda)\mathbf{q}$ for every λ in $[0,1]$.

Example 3.2 An elaboration of the previous example comes from conditional distributions. Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite set $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Recall from Section 2.2 that the marginal distributions $\boldsymbol{\phi}_{\mathcal{X}} \in \mathbb{S}_n$ and $\boldsymbol{\phi}_{\mathcal{Y}} \in \mathbb{S}_m$ are defined by

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{m} \phi_{ij}, (\boldsymbol{\phi}_{\mathcal{Y}})_j = \sum_{i=1}^{n} \phi_{ij},$$

while the m conditional distributions of $\mathcal{X}$ given the observations $\mathcal{Y} = b_j$ are given by

$$\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}} = \left[\frac{\phi_{ij}}{\sum_{i'=1}^{n} \phi_{i'j}}, i = 1, \ldots, n \right] \in \mathbb{S}_n.$$

Now it can be verified that the marginal distribution $\boldsymbol{\phi}_{\mathcal{X}}$ is a convex combination of the m conditional distributions $\boldsymbol{\phi}_{\{\mathcal{X}|\mathcal{Y}=b_j\}}, j = 1, \ldots, m$, where the weights are the components of the marginal distribution $\boldsymbol{\phi}_{\mathcal{Y}}$. This is a straight-forward calculation and is left as an exercise.

Definition 3.1 is stated for a convex combination of *two* vectors, but can be easily extended to a convex combination of any number of points. Suppose $S \subseteq \mathbb{R}^n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_m \in S$. Then a vector of the form

$$\mathbf{y} = \sum_{i=1}^{m} \lambda_i \mathbf{x}_i, \lambda_i \geq 0 \; \forall i, \sum_{i=1}^{m} \lambda_i = 1$$

is called a convex combination of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. It is easy to show, by recursively applying Definition 3.1, that if $S \subseteq \mathbb{R}^n$ is a convex set then every convex combination of any finite number of vectors in S again belongs to S.

Definition 3.2 *Suppose $S \subseteq \mathbb{R}^n$ is a convex set and $f : S \to \mathbb{R}$. We say that the function f is* **convex** *if*

$$f[\lambda \mathbf{x} + (1-\lambda)\mathbf{y}] \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \ \forall \lambda \in [0,1], \ \forall \mathbf{x}, \mathbf{y} \in S. \tag{3.2}$$

We say that the function f is **strictly convex** *if*

$$f[\lambda \mathbf{x} + (1-\lambda)\mathbf{y}] < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \ \forall \lambda \in (0,1), \ \forall \mathbf{x}, \mathbf{y} \in S, \mathbf{x} \neq \mathbf{y}. \tag{3.3}$$

We say that the function f is **concave** *if*

$$f[\lambda \mathbf{x} + (1-\lambda)\mathbf{y}] \geq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \ \forall \lambda \in [0,1], \ \forall \mathbf{x}, \mathbf{y} \in S. \tag{3.4}$$

Finally, we say that the function f is **strictly concave** *if*

$$f[\lambda \mathbf{x} + (1-\lambda)\mathbf{y}] > \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \ \forall \lambda \in (0,1), \ \forall \mathbf{x}, \mathbf{y} \in S, \mathbf{x} \neq \mathbf{y}. \tag{3.5}$$

Equations (3.2) through (3.5) are stated for a convex combination of *two* vectors $\mathbf{x}$ and $\mathbf{y}$. But we can make repeated use of these equations and prove the following facts. If f is a convex function mapping a convex set S into $\mathbb{R}$, and $\mathbf{x}_1, \ldots, \mathbf{x}_m \in S$, then

$$f\left(\sum_{i=1}^{m} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{m} \lambda_i f(\mathbf{x}_i), \text{ whenever } [\lambda_1 \ldots \lambda_m] =: \boldsymbol{\lambda} \in \mathbb{S}_m.$$

Analogous inequalities are valid for concave, strictly convex, and strictly concave functions.

The above definitions are all algebraic. But in the case where S is an interval $[a, b]$ in the real line (finite or infinite), the various inequalities can be given a simple pictorial interpretation. Suppose we plot the graph of the function f. This consists of all pairs $(x, f(x))$ as x varies over the interval $[a, b]$. Suppose $(x, f(x))$ and $(y, f(y))$ are two points on the graph. Then the straight line joining these two points is called the 'chord' of the graph. (We can assume that the two points are distinct, because otherwise the various inequalities (3.2) through (3.5) all become trivial.) Equation (3.2) states that for any two points $x, y \in [a, b]$, the chord joining the two points $(x, f(x))$ and $(y, f(y))$ lies *above* the graph of the function $(z, f(z))$ whenever z lies between x and y. Equation (3.4) says exactly the opposite: It says that the chord joining the two points $(x, f(x))$ and $(y, f(y))$ lies *below* the graph of the function $(z, f(z))$ whenever z lies between x and y. Equation (3.3) says that, not only does the chord joining the two points $(x, f(x))$ and $(y, f(y))$ lie *above* the graph of the function $(z, f(z))$ whenever z lies between x and y, but in fact the chord does not even touch the graph, except at the two extreme points $(x, f(x))$ and $(y, f(y))$. Equation (3.5) says the opposite of the above. Finally, observe that f is (strictly) convex if and only if $-f$ is (strictly) concave.

Figure 3.1 depicts the interpretation of the definition of convexity. This figure depicts the fact that, for all z belonging to the chord connecting x and y, the value $f(z)$ lies below the chord value. The same figure also *suggests* that, if the chord is extended beyond the two original points x and y, then the value $f(z)$ actually lies *above* the chord value. This intuition is indeed correct, and not just in one dimension either!

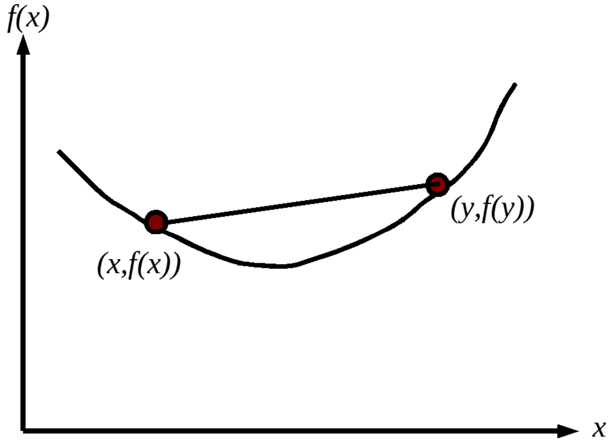


Figure 3.2 Graph below chord interpretation of a convex function

Lemma 3.3 *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is convex, and suppose* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *with* $\mathbf{x} \neq \mathbf{y}$. *Then, for every* $\lambda < 0$ *and every* $\lambda > 1$, *we have that*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}). \tag{3.6}$$

Proof. We begin with the case $\lambda < 0$. Let $\lambda = -\alpha$ where $\alpha > 0$, and define $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Then simple algebra shows that

$$\mathbf{y} = \frac{\alpha}{1+\alpha}\mathbf{x} + \frac{1}{1+\alpha}\mathbf{z},$$

so that $\mathbf{y}$ is a convex combination of $\mathbf{x}$ and $\mathbf{z}$. Now the convexity of $f(\cdot)$ implies that

$$f(\mathbf{y}) \leq \frac{\alpha}{1+\alpha}f(\mathbf{x}) + \frac{1}{1+\alpha}f(\mathbf{z}),$$

which can be rearranged as

$$f(\mathbf{z}) \geq -\alpha f(\mathbf{x}) + (1+\alpha)f(\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}).$$

The case where $\lambda > 1$ is handled entirely similarly by interchanging the roles of $\mathbf{x}$ and $\mathbf{y}$. □

It can be shown that, for all practical purposes, a convex function (and thus a concave function) has to be continuous. But if a function is not merely continuous but also differentiable, then it is possible to give alternate characterizations of convexity (and of course concavity) that are more interesting.

Lemma 3.4 *Suppose* $f : [a,b] \to \mathbb{R}$ *is convex and is continuously differentiable on* (a,b). *Then*

$$f(y) \geq f(x) + f'(x)(y-x), \ \forall x \in (a,b), y \in [a,b], y \neq x. \tag{3.7}$$

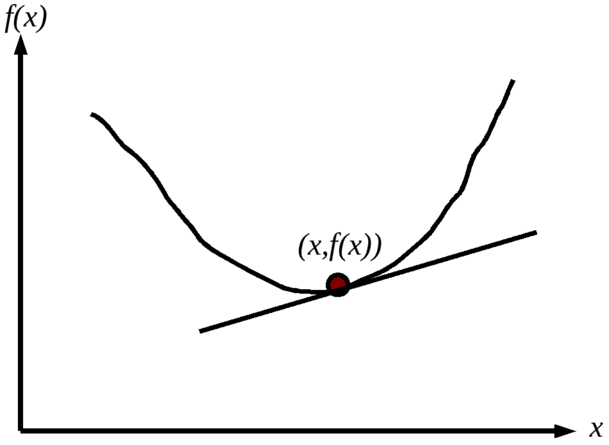


Figure 3.3 The graph above tangent property of a convex function

If f is concave on $[a, b]$ and continuously differentiable on (a, b), then

$$f(y) \leq f(x) + f'(x)(y - x),\ \forall x \in (a, b), y \in [a, b], y \neq x. \qquad (3.8)$$

If f is strictly convex on $[a, b]$ and continuously differentiable on (a, b), then

$$f(y) > f(x) + f'(x)(y - x),\ \forall x \in (a, b), y \in [a, b], y \neq x. \qquad (3.9)$$

If f is strictly concave on $[a, b]$ and continuously differentiable on (a, b), then

$$f(y) < f(x) + f'(x)(y - x),\ \forall x \in (a, b), y \in [a, b], y \neq x. \qquad (3.10)$$

We do not give the proof of this or other such lemmas, as they are beyond the scope of the book. Instead, the reader is referred to the classic text of Rockafellar [93]. But we give instead the interpretations of the various inequalities above. Suppose f is continuously differentiable on (a, b). Then for every $x \in (a, b)$, the function $y \mapsto f(x)+f'(x)(y-x)$ is the tangent to the graph of f at the point $(x, f(x))$. Thus (3.7) says that for a convex function, the tangent lies below the graph. This is to be contrasted with (3.2), which says that the chord lies above the graph. Equation (3.9) says that if the function is strictly convex, then not only does the tangent lie below the graph, but the tangent touches the graph only at the single point $(x, f(x))$. The interpretations of the other two inequalities are entirely similar. Figure 3.1 depicts the 'graph above the tangent' property of a convex function, which is to be contrasted with the 'graph below the chord' property depicted in Figure 3.1.

If the function is in fact *twice* continuously differentiable, then we can give yet another set of characterizations of the various forms of convexity.

Lemma 3.5 *Suppose $f : [a, b] \to \mathbb{R}$ is twice continuously differentiable on (a, b). Then*

1. *f is convex if and only if $f''(x) \geq 0$ for all $x \in (a,b)$.*

2. *f is concave if and only if $f''(x) \leq 0$ for all $x \in (a,b)$.*

3. *f is strictly convex if and only if $f''(x) > 0$ for all $x \in (a,b)$.*

4. *f is strictly convex if and only if $f''(x) < 0$ for all $x \in (a,b)$.*

This lemma is also found in [93].

We now study two very specific functions that are very relevant to information theory.

Example 3.3 Consider the function $f(x) = \log x$, defined on $(0, \infty)$.[1] Since

$$f'(x) = 1/x,\ f''(x) = -1/x^2 < 0\ \forall x \in (0, \infty),$$

it follows that $\log x$ is strictly concave on $(0, \infty)$. As a result, if we substitute $x = 1$ in (3.10), we get

$$\log y < y - 1,\ \forall y > 0, y \neq 1. \tag{3.11}$$

Example 3.4 The function

$$h(p) = p \log(1/p) = -p \log p, p \in [0,1] \tag{3.12}$$

plays a very central role in information theory. For $p \in (0,1]$ the function is well-defined and can be differentiated as many times as one wishes. When $p = 0$ we can define $h(0) = 0$, since it is easy to verify using L'Hôpital's rule that $h(p) \to 0$ as $p \to 0^+$. Note that

$$h'(p) = -\log p - 1, h''(p) = -1/p < 0\ \forall p \in (0,1).$$

Thus $h(\cdot)$ is strictly concave on $[0,1]$ (and indeed on $(0, \infty)$, though values of p larger than one have no relevance to information theory). In particular, $h(p) > 0\ \forall p \in (0,1)$, and $h(0) = h(1) = 0$.

Now we present a very useful result, known as **Jensen's inequality**.

Theorem 3.6 *Suppose $\mathcal{X}$ is a random variable assuming one of n real values $x_1, \ldots, x_n$ belonging an open interval (a,b), with probabilities $\mu_1, \ldots, \mu_n$. Suppose $f : (a,b) \to \mathbb{R}$ is convex. Then*

$$f(E[\mathcal{X}, P_\mu]) \leq E[f(\mathcal{X}), P_\mu] \tag{3.13}$$

Proof. The proof is a ready consequence of the definition of convexity. By definition, we have

$$E[\mathcal{X}, P_\mu] = \sum_{i=1}^{n} \mu_i x_i,$$

[1] Here and elsewhere log denotes the natural logarithm, while lg denotes the binary logarithm, that is, logarithm to the base 2.

$$f(E[\mathcal{X}, P_\mu]) = f\left(\sum_{i=1}^{n} \mu_i x_i\right) \leq \sum_{i=1}^{n} \mu_i f(x_i) = E[f(\mathcal{X}), P_\mu],$$

which is the desired conclusion. □

The above theorem statement and proof don't really do justice to Jensen's inequality, which actually holds in a far more abstract setting than the above.

Problem 3.1 Given a function $f : S \to \mathbb{R}$ where S is a convex subset of some Euclidean space $\mathbb{R}^d$, its epigraph is denoted by $\mathrm{epi}(f)$ and is defined by

$$\mathrm{epi}(f) = \{(\mathbf{x}, y) : \mathbf{x} \in S, y \in \mathbb{R} \text{ and } y \geq f(\mathbf{x})\}.$$

Show that f is a convex *function* if and only if $\mathrm{epi}(f)$ is a convex *set* in $\mathbb{R}^{d+1}$. State and prove the analog of this statement for concave functions.

Problem 3.2 Suppose S is a convex set and $f : S \to \mathbb{R}$. Show that f is a convex function if and only if $-f$ is a concave function.

Problem 3.3 Suppose S is a convex subset of some Euclidean space $\mathbb{R}^d$ and $f : S \to \mathbb{R}$ is convex, while $g : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing. In other words, $\alpha \leq \beta$ implies that $g(\alpha) \leq g(\beta)$. Show that the function $\mathbf{x} \mapsto g(f(\mathbf{x}))$ is convex. Here the symbol $\mapsto$ is read as 'mapsto', and $\mathbf{x} \mapsto g(f(\mathbf{x}))$ means the function that associates $g(f(\mathbf{x}))$ with each $\mathbf{x} \in S$. State and prove analogous statements for strictly convex, concave, and strictly concave functions.

Problem 3.4 Using Lemma 3.5, show that the function $h(u) = -u \log u$ is a strictly concave function of u.

Problem 3.5 Consider the following three probability distributions:

$$\boldsymbol{\phi} = [0.2 \ \ 0.3 \ \ 0.5], \boldsymbol{\psi} = [0.4 \ \ 0.5 \ \ 0.1], \boldsymbol{\theta} = [0.3 \ \ 0.4 \ \ 0.3].$$

Can you express any one of these three distributions as a convex combination of the other two? Justify your answer.

3.2 ENTROPY

The notion of entropy is very central to information theory. In this section we introduce this concept and derive several of its properties.

3.2.1 Definition of Entropy

Definition 3.7 *Suppose $\boldsymbol{\mu} \in \mathbb{S}_n$ is an n-dimensional probability distribution. The* **entropy** *of the distribution is denoted by $H(\boldsymbol{\mu})$ and is defined by*

$$H(\boldsymbol{\mu}) := \sum_{i=1}^{n} \mu_i \log(1/\mu_i) = -\sum_{i=1}^{n} \mu_i \log \mu_i. \tag{3.14}$$

Note that we can also write

$$H(\boldsymbol{\mu}) = \sum_{i=1}^{n} h(\mu_i),$$

where $h(\cdot)$ is the function defined in Example 3.4.

In (3.14), we are using the natural logarithm. It is of course possible to replace log by lg and take the logarithm to the base two. Some authors try to specify which logarithm is used by saying that the entropy is measured in 'bits' if lg is used, and 'nats' if log is used. Clearly the two numbers will always differ by the constant factor $\log 2$, so it does not really matter which base is used for the logarithm, so long as one is consistent.

If $\mathcal{X}$ is a random variable assuming values in a set $\mathbb{A} = \{a_1, \ldots, a_n\}$, and $\boldsymbol{\mu}$ is the probability distribution of $\mathcal{X}$, then we can also refer to $H(\boldsymbol{\mu})$ as $H(\mathcal{X})$, the entropy of the random variable $\mathcal{X}$, rather than as the entropy of the *probability distribution* of the random variable $\mathcal{X}$. It is helpful to be able to switch back and forth between the two usages, and the two notations $H(\boldsymbol{\mu})$ and $H(\mathcal{X})$. However, the reader is cautioned that, strictly speaking, entropy is a property of probability distributions, and not of the random variables associated with those probability distributions. For example, consider two random variables: the outcome of tossing a fair coin, and drawing a ball from a box containing two identical balls of different colours. The underlying random variables are in some sense 'different,' but the entropies are the same.

3.2.2 Properties of the Entropy Function

Theorem 3.8 *We have the following properties of the entropy function.*

1. $H(\boldsymbol{\mu}) \geq 0$ *for all probability distributions* $\boldsymbol{\mu} \in \mathbb{S}_n$.

2. $H(\boldsymbol{\mu}) = 0$ *if and only if* $\boldsymbol{\mu}$ *is a degenerate distribution, i.e., there is an index* i *such that* $\mu_i = 1$ *and* $\mu_j = 0$ *for all* $j \neq i$.

3. $H(\boldsymbol{\mu}) \leq \log n \; \forall \boldsymbol{\mu} \in \mathbb{S}_n$, *with equality if and only if* $\boldsymbol{\mu}$ *is the uniform distribution, that is,* $\mu_i = 1/n$ *for all indices* i.

Proof. By definition we have

$$H(\boldsymbol{\mu}) = \sum_{i=1}^{n} h(\mu_i),$$

where the function $h(\cdot)$ is defined in (3.12). Since $h(\mu_i) \geq 0$ for all $\mu_i \in [0,1]$, it follows that $H(\boldsymbol{\mu}) \geq 0$. This proves the first statement. Moreover, $H(\boldsymbol{\mu}) = 0$ if and only if $h(\mu_i) = 0$ for all indices i, that is, if and only if $\mu_i = 0$ or 1 for every index i. But since the μ_i's must add up to one, we see that $H(\boldsymbol{\mu}) = 0$ if and only if all components of $\boldsymbol{\mu}$ are zero except for one component which must equal one. This proves the second statement. To prove the third statement, suppose $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathcal{S}_n$. Then, since

$\mu_i \in [0,1]\ \forall i$ and $\sum_{i=1}^n \mu_i = 1$, it follows that

$$\frac{1}{n} = \sum_{i=1}^n \frac{\mu_i}{n}$$

is a convex combination of the numbers μ_1 through μ_n (with equal weights $1/n$). Now recall that the function $h(\cdot)$ in Example 3.4 is strictly concave. As a result it follows that

$$h(1/n) = h\left(\frac{1}{n}\sum_{i=1}^n \mu_i\right) \geq \frac{1}{n}\sum_{i=1}^n h(\mu_i),$$

with equality if and only if all μ_i are equal (and thus equal $1/n$). Since $h(1/n) = (1/n)\log n$, the above equation can be rewritten as

$$H(\boldsymbol{\mu}) = \sum_{i=1}^n h(\mu_i) \leq nh(1/n) = \log n,$$

with equality holding if and only if all μ_i are equal to $1/n$. □

The motivation for the next theorem is that we can view the entropy function $H(\cdot)$ as a function mapping the convex set $\mathbb{S}_n$ of n-dimensional probability distributions into the set $\mathbb{R}_+$ of nonnegative numbers. The theorem states that the entropy function is strictly concave.

Theorem 3.9 *For each integer n, the function $H(\cdot) : \mathbb{S}_n \to \mathbb{R}_+$ is strictly concave. Thus if $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$ are probability distributions and $\lambda \in (0,1)$, then*

$$H[\lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{\nu}] \geq \lambda H(\boldsymbol{\mu}) + (1-\lambda)H(\boldsymbol{\nu}). \tag{3.15}$$

Moreover, equality holds if and only if $\boldsymbol{\mu} = \boldsymbol{\nu}$.

Proof. Recall that the function $h(\cdot)$ defined in (3.12) is strictly concave. From the definition of $H(\cdot)$ it follows that

$$\begin{aligned} H[\lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{\nu}] &= \sum_{i=1}^n h(\lambda\mu_i + (1-\lambda)\nu_i) \\ &\geq \sum_{i=1}^n \lambda h(\mu_i) + (1-\lambda)h(\nu_i) \\ &= \lambda\sum_{i=1}^n h(\mu_i) + (1-\lambda)\sum_{i=1}^n h(\nu_i) \\ &= \lambda H(\boldsymbol{\mu}) + (1-\lambda)H(\boldsymbol{\nu}). \end{aligned}$$

Moreover, if $\mu_i \neq \nu_i$ for even a single index i, then the corresponding inequality in the second step becomes strict. □

3.2.3 Conditional Entropy

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. In Section 2.2 we studied the notions of the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$, as well as the conditional distribution of $\mathcal{X}$ given an observed value of $\mathcal{Y}$. Specifically, let $\boldsymbol{\phi}$ denote the joint distribution of $(\mathcal{X}, \mathcal{Y})$. Then $\boldsymbol{\phi} \in \mathbb{S}_{nm}$, but we can represent $\boldsymbol{\phi}$ as a doubly indexed set of numbers, as follows:

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_{11} & \dots & \phi_{1m} \\ \vdots & \vdots & \vdots \\ \phi_{n1} & \dots & \phi_{nm} \end{bmatrix}.$$

We also defined the 'marginal distributions' $\boldsymbol{\phi}_{\mathcal{X}}$ on X and $\boldsymbol{\phi}_{\mathcal{Y}}$ on Y as follows:

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{m} \phi_{ij}, i = 1, \dots, n, \text{ and } (\boldsymbol{\phi}_{\mathcal{Y}})_j = \sum_{i=1}^{n} \phi_{ij}, j = 1, \dots, m.$$

Given an observation $\mathcal{X} = a_j$, we defined the conditional probability distribution $\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}}$ as[2]

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} = [\phi_{\{b_1|a_i\}} \cdots \phi_{\{b_m|a_i\}}],$$

where

$$\phi_{\{b_j|a_i\}} := \frac{\phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{X}})_i} = \frac{\phi_{ij}}{\sum_{j'=1}^{m} \phi_{ij'}}.$$

Then $\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} \in \mathbb{S}_m \ \forall i$. All of the above is just a reprise of previously discussed material for the convenience of the reader.

Definition 3.10 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$ and $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Let $\boldsymbol{\phi}$ denote their joint probability distribution and let $\boldsymbol{\phi}_{\mathcal{X}}$ and $\boldsymbol{\phi}_{\mathcal{Y}}$ denote the marginal distributions. Then the quantity*

$$H(\mathcal{Y}|\mathcal{X}) := \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i H(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}})$$

is called the **conditional entropy** *of the random variable $\mathcal{Y}$ with respect to the random variable $\mathcal{X}$.*

Theorem 3.11 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$ and $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Then*

$$H(\mathcal{Y}) \geq H(\mathcal{Y}|\mathcal{X}), \tag{3.16}$$

or equivalently

$$H(\boldsymbol{\phi}_{\mathcal{Y}}) \geq \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i H(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}}). \tag{3.17}$$

[2] Just for variety's sake we are using the conditional distribution of $\mathcal{Y}$ given $\mathcal{X}$, whereas in Chapter 2 we used the conditional distribution of $\mathcal{X}$ given $\mathcal{Y}$.

Proof. It is obvious from the various formulas that

$$\phi_{\mathcal{Y}} = \sum_{i=1}^{n} (\phi_{\mathcal{X}})_i \cdot \phi_{\{\mathcal{Y}|\mathcal{X}=a_i\}}. \tag{3.18}$$

In other words, the marginal probability distribution $\phi_{\mathcal{Y}}$ of the random variable $\mathcal{Y}$ is a *convex combination* of the various conditional probability distributions $\phi_{\{\mathcal{Y}|\mathcal{X}=a_i\}}$, weighted by the probabilities $(\phi_{\mathcal{X}})_i$. Thus the desired inequality (3.17) follows from Theorem 3.9 and repeated application of (3.15). □

The quantity $H(\phi_{\{\mathcal{Y}|\mathcal{X}=x_i\}})$ is the entropy of the conditional probability of $\mathcal{Y}$, after we have observed that the value of $\mathcal{X}$ is a_i. One way of interpreting Theorem 3.11 is that, *on average*, the conditional entropy of $\mathcal{Y}$ following an observation of another variable $\mathcal{X}$ is no larger than the unconditional entropy $H(\mathcal{Y})$. But this statement applies *only* 'on average.' It is quite possible that *for some specific observations*, this entropy is in fact higher than the 'unconditional entropy' $H(\phi_{\mathcal{Y}})$; see Example 3.5 below. However, Theorem 3.11 states that *on average*, one cannot be worse off by making an observation of $\mathcal{X}$ than by not observing $\mathcal{X}$.

Example 3.5 Suppose $|X| = |Y| = 2$, and that the joint probability distribution ϕ equals

$$\phi = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix},$$

where the rows correspond to the values of $\mathcal{X}$ and the columns to values of $\mathcal{Y}$. Thus

$$\phi_{\mathcal{X}} = [0.4 \ \ 0.6], \ \phi_{\mathcal{Y}} = [0.3 \ \ 0.7].$$

So, if we know nothing about $\mathcal{X}$, we get

$$H(\mathcal{Y}) = H(\phi_{\mathcal{Y}}) = -0.3 \log 0.3 - 0.7 \log 0.7 \approx$$

Now suppose we measure $\mathcal{X}$ and it turns out that $\mathcal{X} = a_1$. Then

$$\phi_{\{\mathcal{Y}|\mathcal{X}=a_1\}} = [0.1/0.4 \ \ 0.3/0.4] = [1/4 \ \ 3/4].$$

In this case we have

$$H(\phi_{\{\mathcal{Y}|\mathcal{X}=a_2\}}) = (1/4)\log 4 + (3/4)\log(4/3) \approx$$

which is *lower* than the unconditional entropy. On the other hand,

$$\phi_{\{\mathcal{Y}|\mathcal{X}=a_2\}} = [1/3 \ \ 2/3],$$

and

$$H(\phi_{\{\mathcal{Y}|\mathcal{X}=a_2\}}) \approx$$

which is *higher* than the unconditional entropy. Finally

$$H(\mathcal{Y}|\mathcal{X}) = (\phi_{\mathcal{X}})_1 H(\phi_{\{\mathcal{Y}|\mathcal{X}=a_1\}}) + (\phi_{\mathcal{X}})_2 H(\phi_{\{\mathcal{Y}|\mathcal{X}=a_2\}}) \approx$$

which is *lower than* $H(\mathcal{Y}) = H(\phi_{\mathcal{Y}})$.

Theorem 3.12 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Let $\boldsymbol{\phi}$ denote their joint probability distribution and let $\boldsymbol{\phi}_{\mathcal{X}}$ and $\boldsymbol{\phi}_{\mathcal{Y}}$ denote the marginal distributions. Then*

$$H((\mathcal{X}, \mathcal{Y})) = H(\mathcal{X}) + H(\mathcal{Y}|\mathcal{X}). \tag{3.19}$$

In the above theorem statement, we have used the convention of associating entropy with a random variable rather than with its probability distribution. Thus (3.19) says that the entropy of the *joint* random variable $(\mathcal{X}, \mathcal{Y})$ is the sum of the entropy of $\mathcal{X}$ by itself, and the conditional entropy of $\mathcal{Y}$ given $\mathcal{X}$.

Proof. The desired equality (3.19) is equivalent to

$$H(\mathcal{Y}|\mathcal{X}) = H((\mathcal{X}, \mathcal{Y})) - H(\mathcal{X}). \tag{3.20}$$

To establish this relation, let us compute the right side, observing that

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{n} \phi_{ij} \; \forall i, \phi_{ij} = (\boldsymbol{\phi}_{\mathcal{X}})_i \phi_{b_j|a_i} \; \forall i, j.$$

Thus

$$\begin{aligned} H((\mathcal{X}, \mathcal{Y})) - H(\mathcal{X}) &= -\sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log \phi_{ij} + \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i \log(\boldsymbol{\phi}_{\mathcal{X}})_i \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log \phi_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log(\boldsymbol{\phi}_{\mathcal{X}})_i \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log \left(\frac{\phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{X}})_i} \right) \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} (\boldsymbol{\phi}_{\mathcal{X}})_i \phi_{b_j|a_i} \log \phi_{b_j|a_i} \\ &= -\sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i \sum_{j=1}^{m} \phi_{b_j|a_i} \log \phi_{b_j|a_i} \\ &= \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i H(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}}) = H(\mathcal{Y}|\mathcal{X}). \end{aligned}$$

This establishes (3.20) and completes the proof. □

Corollary 3.13 *Let all notation be as in Theorem 3.12. Then*

$$H((\mathcal{X}, \mathcal{Y})) = H(\mathcal{X}) + H(\mathcal{Y})$$

if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent.

Proof. The statement $H((\mathcal{X}, \mathcal{Y})) = H(\mathcal{X}) + H(\mathcal{Y})$ is equivalent to $H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{Y})$. Now we know from (3.18) that $\boldsymbol{\phi}_{\mathcal{Y}}$ is a convex combination of the n

conditional probability distributions $\phi_{\{\mathcal{Y}|\mathcal{X}=a_i\}}$. Since the entropy function $H(\cdot)$ is strictly concave, we have that

$$H(\mathcal{Y}) = H(\phi_{\mathcal{Y}}) \geq \sum_{i=1}^{n} (\phi_{\mathcal{X}})_i H(\phi_{\{\mathcal{Y}|\mathcal{X}=a_i\}}) = H(\mathcal{Y}|\mathcal{X}),$$

with equality if and only if

$$\phi_{\{\mathcal{Y}|\mathcal{X}=a_i\}} = \phi_{\mathcal{Y}} \; \forall i,$$

which is the same as saying that $\mathcal{X}$ and $\mathcal{Y}$ are independent. □

Corollary 3.14 *Let $\mathcal{X}, \mathcal{Y}$ be as in Theorem 3.12. Then*

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

Proof. From the definition it is obvious that $H((\mathcal{X}, \mathcal{Y})) = H((\mathcal{Y}, \mathcal{X}))$. The conclusion now follows from Corollary 3.13 by interchanging $\mathcal{X}$ and $\mathcal{Y}$. □

Definition 3.15 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Let ϕ denote their joint probability distribution and let $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ denote the marginal distributions. Then the quantity*

$$I(\mathcal{X}, \mathcal{Y}) := H(\mathcal{X}) + H(\mathcal{Y}) - H((\mathcal{X}, \mathcal{Y})) = H(\phi_{\mathcal{X}}) + H(\phi_{\mathcal{Y}}) - H(\phi)$$

is called the **mutual information** *between $\mathcal{X}$ and $\mathcal{Y}$.*

From the above definition, it is clear that the mutual information is 'symmetric', that is, $I(\mathcal{X}, \mathcal{Y}) = I(\mathcal{Y}, \mathcal{X})$. However, as the next lemma shows, there are other, equivalent, definitions that do not appear to be symmetric.

Lemma 3.16 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \ldots, a_n\}$ and $\mathbb{B} = \{b_1, \ldots, b_m\}$ respectively. Then*

$$I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}).$$

Thus $I(\mathcal{X}, \mathcal{Y}) \geq 0$, and $I(\mathcal{X}, \mathcal{Y}) = 0$ if and only if $\mathcal{X}, \mathcal{Y}$ are independent.

Proof. From Theorem 3.12, we have that

$$H(\mathcal{Y}|\mathcal{X}) = H((\mathcal{X}, \mathcal{Y})) - H(\mathcal{X}), H(\mathcal{X}|\mathcal{Y}) = H((\mathcal{X}, \mathcal{Y})) - H(\mathcal{Y}).$$

So

$$\begin{aligned} I(\mathcal{X}, \mathcal{Y}) &= H(\mathcal{X}) + H(\mathcal{Y}) - H((\mathcal{X}, \mathcal{Y})) \\ &= H(\mathcal{X}) + H(\mathcal{Y}) - [H(\mathcal{Y}|\mathcal{X}) + H(\mathcal{X})] \\ &= H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}), \end{aligned}$$

and similarly for the other equation. The second statement is obvious. □

3.2.4 Uniqueness of the Entropy Function

Up to now we have established a few key properties of the entropy function. First, as shown in Property 3 in Theorem 3.8, for each integer n the entropy function $H(\boldsymbol{\phi})$ is maximized when $\boldsymbol{\phi}$ is the *uniform distribution* in $\mathbb{S}_n$. Second, as shown in Theorem 3.12, the entropy of a joint distribution is given by (3.19), so that $H(\mathcal{X}, \mathcal{Y})) = H(\mathcal{X}) + H(\mathcal{Y}|\mathcal{X})$. In a remarkable paper, the Russian probabilist A. N. Khinchin showed that, with one technical assumption, *the only way* to define a continuous entropy function that satisfies the above two properties is via (3.14). An English translation of this paper can be found as a part of [70]. We reproduce that proof below, with some changes in notation.

Theorem 3.17 *Suppose $\{f_n\}$ is a family of functions such that $f_n : \mathbb{S}_n \to [0, \infty)$, satisfying the following properties:*

1. *$f_n(\cdot)$ is continuous on $\mathbb{S}_n$ for each n.*

2. *For each n, it is the case that*

$$f_n(\boldsymbol{\phi}) \leq f_n(\mathbf{u}_n) \ \forall \boldsymbol{\phi} \in \mathbb{S}_n,$$

   where $\mathbf{u}_n$ denotes the uniform probability distribution on n entries, that is, $\mathbf{u}_n = (1/n, \ldots, 1/n)$.

3. *For each n and each $\boldsymbol{\phi} \in \mathbb{S}_n$, define $\bar{\boldsymbol{\phi}} \in \mathbb{S}_{n+1}$ as*

$$\bar{\boldsymbol{\phi}} = (\boldsymbol{\phi}, 0).$$

   Then

$$f_{n+1}(\bar{\boldsymbol{\phi}}) = f_n(\boldsymbol{\phi}). \tag{3.21}$$

4. *The function f_n satisfies a property analogous to (3.19). To amplify, suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Then*

$$f_{nm}(\boldsymbol{\phi}) = \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i f_m(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=i\}}) + f_n(\boldsymbol{\phi}_{\mathcal{X}}). \tag{3.22}$$

With these assumptions, there exists a constant λ, independent of n, such that

$$f_n(\boldsymbol{\phi}) = \lambda H(\boldsymbol{\phi}), \ \forall \boldsymbol{\phi} \in \mathbb{S}_n, \ \forall n \geq 1,$$

where $H(\cdot)$ is defined in (3.14).

Remarks:

1. Suppose $\boldsymbol{\phi} \in \mathbb{S}_n$. Then $\bar{\boldsymbol{\phi}} \in \mathbb{S}_{n+1}$ can be thought of as a probability distribution on $n + 1$ elements, except that the additional element has zero probability, that is, its occurence is an 'impossible' event.

So (3.21) states that adding one impossible event to any probability distribution does not change the value of its 'entropy-like' function. Note that, by repeated application of (3.21), we can readily establish the following property. Suppose $\boldsymbol{\phi} \in \mathbb{S}_n$ and let l be any integer. Then

$$f_{n+l}((\boldsymbol{\phi}, \mathbf{0}_l)) = f_n(\boldsymbol{\phi}). \tag{3.23}$$

2. A ready consequence of (3.22) is that if $\mathcal{X}, \mathcal{Y}$ are independent random variables, so that $\phi_{ij} = (\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}_{\mathcal{Y}})_j$ for all i, j, then

$$f_{nm}(\boldsymbol{\phi}) = f_n(\boldsymbol{\phi}_{\mathcal{X}}) + f_m(\boldsymbol{\phi}_{\mathcal{Y}}). \tag{3.24}$$

3. Thus Theorem 3.17 states that *the only* continuous function that satisfies the technical condition (3.21) along with the analogs of Property 3 in Theorem 3.8 and (3.19) is the function $H(\cdot)$ defined in (3.14), or some scalar multiple thereof.

Proof. For convenience, define

$$L(n) = f_n(\mathbf{u}_n),$$

where as before $\mathbf{u}_n$ denotes the uniform distribution on n elements. By Property 3, we have that

$$f_{n+1}((\bar{\mathbf{u}}_n)) = f_n(\mathbf{u}_n) = L(n).$$

Next, by Property 2, it follows that

$$f_{n+1}((\bar{\mathbf{u}}_n)) \leq f_{n+1}(\mathbf{u}_{n+1}) = L(n+1).$$

Combining these two inequalities leads to the conclusion

$$L(n) \leq L(n+1).$$

Thus $\{L(n)\}$ is a nondecreasing sequence as a function of n.

Next, suppose $n = lm$, and consider $u_n = \mathbf{u}_{lm}$. Since $1/n = (1/l) \cdot (1/m)$, we can equate $\mathbf{u}_n$ with the joint distribution of $\mathcal{X}, \mathcal{Y}$, where $\mathcal{X}$ is uniformly distributed over $\{1, \ldots, l\}$, $\mathcal{Y}$ is uniformly distributed over $\{1, \ldots, m\}$, and $\mathcal{X}, \mathcal{Y}$ are independent. Thus it follows from (3.24) applied to this situation that

$$f_{lm}(\mathbf{u}_{lm}) = f_l(\mathbf{u}_l) + f_m(\mathbf{u}_m),$$

or equivalently

$$L(lm) = L(l) + L(m) \ \forall l, m.$$

In particular

$$L(m^2) = 2L(m), \text{ and}$$

$$L(m^{s+1}) = L(m^s) + L(m) = (s+1)L(m), \tag{3.25}$$

where the last step follows by induction.

Next, suppose l, m, n are arbitrary integers ≥ 2, and determine a unique integer s according to

$$m^s \leq l^n < m^{s+1}.$$

This implies that

$$s \log m \leq n \log l \leq (s+1) \log m, \text{ or}$$

$$\frac{s}{n} \leq \frac{\log l}{\log m} \leq \frac{s}{n} + \frac{1}{n}. \tag{3.26}$$

Now since $L(\cdot)$ is a nondecreasing function, it follows that

$$L(m^s) \leq L(l^n) \leq L(m^{s+1}).$$

Applying (3.25) now shows that

$$sL(m) \leq nL(l) \leq (s+1)L(m),$$

or

$$\frac{s}{n} \leq \frac{L(l)}{L(m)} \leq \frac{s}{n} + \frac{1}{n}. \tag{3.27}$$

Combining (3.26) and (3.27) shows that

$$\left| \frac{\log l}{\log m} - \frac{L(l)}{L(m)} \right| \leq \frac{1}{n}.$$

Since the left side of the inequality does not contain n, we can let $n \to \infty$, which shows that

$$\frac{\log l}{\log m} = \frac{L(l)}{L(m)},$$

or

$$L(m) = \frac{L(l)}{l} \log m, \ \forall l, m.$$

However, since l, m are arbitrary, we can define

$$\lambda := \frac{L(2)}{2},$$

and conclude that

$$L(m) = \lambda \log m, \ \forall m. \tag{3.28}$$

Next, suppose $\boldsymbol{\mu} \in \mathbb{S}_n$ contains only rational numbers. Specifically, suppose $\mu_i = g_i/g$ for all i, where g_i, g are all integers. It is clear that $\sum_{i=1}^n g_i = g$ since $\boldsymbol{\mu} \in \mathbb{S}_n$. Let m denote the largest number among the g_i, that is, $m := \max_i g_i$. We now define a joint distribution $\boldsymbol{\phi}$ on $\{1, \ldots, n\} \times \{1, \ldots, m\}$ that is closely related to $\boldsymbol{\mu}$. Define

$$\phi_{ij} := \begin{cases} 1/g, & 1 \leq j \leq g_i, \\ 0, & j > g_i. \end{cases}$$

We can think of $\boldsymbol{\phi}$ as the joint distribution of $(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ is a random variable taking values in $\{1, \ldots, n\}$ and $\mathcal{Y}$ is a random variable taking values in $\{1, \ldots, m\}$. From this distribution, two facts are readily apparent. First, the marginal distribution $\boldsymbol{\phi}_{\mathcal{X}}$ is given by

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{m} \phi_{ij} = g_i/g, \; \forall i.$$

Thus $\mathcal{X}$ has the distribution $\boldsymbol{\mu}$ with which we started. Second, to compute the conditional distribution $\boldsymbol{\phi}_{\mathcal{Y}}$, observe that

$$\Pr\{\mathcal{Y} = j | \mathcal{X} = i\} = \frac{\Pr\{\mathcal{Y} = j \& \mathcal{X} = i\}}{\Pr\{\mathcal{X} = i\}} = \begin{cases} 1/g_i, & 1 \leq j \leq g_i, \\ 0, & j > g_i. \end{cases}$$

In other words,

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=i\}} = (\mathbf{u}_{g_i}, \mathbf{0}_{m-g_i}).$$

Therefore, by Property 3 and its consequence (3.23), we get

$$f_m(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=i\}}) = f_{g_i}(\mathbf{u}_{g_i}) = L_{g_i}(g_i) = \lambda \log g_i.$$

Now the distribution $\boldsymbol{\phi}$ itself, viewed on a set of nm elements, consists of the uniform distribution $\mathbf{u}_g$ augmented by $nm - g$ zeros. Thus by Property 3 again, we get

$$f_{nm}(\boldsymbol{\phi}) = f_g(\mathbf{u}_g) = \lambda \log g.$$

To complete this step, we apply Property 4 to $\boldsymbol{\phi}$ and recall that $\boldsymbol{\phi}_{\mathcal{X}} = \boldsymbol{\mu}$, the original distribution. Then it follows from (3.22) that

$$\lambda \log g = \sum_{i=1}^{n} \frac{g_i}{g} \lambda \log g_i + f_n(\boldsymbol{\mu}),$$

$$\begin{aligned} -f_n(\boldsymbol{\mu}) &= \lambda \sum_{i=1}^{n} \left[\frac{g_i}{g} \log g_i \right] - \lambda \log g \\ &= \lambda \sum_{i=1}^{n} \left[\frac{g_i}{g} \log g_i - \frac{g_i}{g} \log g \right] \text{ since } \sum_i g_i = g \\ &= \lambda \sum_{i=1}^{n} \frac{g_i}{g} \log \frac{g_i}{g} \\ &= -\lambda H(\boldsymbol{\mu}), \end{aligned}$$

where $H(\cdot)$ is defined in (3.14).

Thus it has been shown that $f_n(\boldsymbol{\mu}) = \lambda H(\boldsymbol{\mu})$ whenever every entry of $\boldsymbol{\mu} \in \mathbb{S}_n$ is a rational number. To complete the proof, observe that every real number can be approximated arbitrarily closely by a rational number, and that f_n is continuous on $\mathbb{S}_n$, for each n. So we conclude that

$$f_n(\boldsymbol{\mu}) = \lambda H(\boldsymbol{\mu}) \; \forall \boldsymbol{\mu} \in \mathbb{S}_n, \; \forall n,$$

which is the desired conclusion. □

3.3 RELATIVE ENTROPY AND THE KULLBACK-LEIBLER DIVERGENCE

In this section we introduce a very important notion known as relative entropy. It is also called the Kullback-Leibler divergence after the two persons who invented this notion.

Definition 3.18 *If $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$ are probability vectors with n components each, define $\boldsymbol{\mu} \ll \boldsymbol{\nu}$ or $\boldsymbol{\nu} \gg \boldsymbol{\mu}$ (read as '$\boldsymbol{\mu}$ is dominated by $\boldsymbol{\nu}$,' or '$\boldsymbol{\nu}$ dominates $\boldsymbol{\mu}$') if $\nu_i = 0 \;\Rightarrow\; \mu_i = 0$. Then*

$$H(\boldsymbol{\mu} \| \boldsymbol{\nu}) := \sum_{i=1}^{n} \mu_i \log(\mu_i / \nu_i) \tag{3.29}$$

is called the **relative entropy** *of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\nu}$, or the* **Kullback-Leibler (K-L) divergence** *of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\nu}$.*

Note that if $\nu_i = 0$ for some index i, the assumption that $\boldsymbol{\mu} \ll \boldsymbol{\nu}$ guarantees that μ_i also equals zero. So in defining the K-L divergence, we take $0 \log(0/0)$ to equal 0. If $\boldsymbol{\mu} \not\ll \boldsymbol{\nu}$, then there exists an index i such that $\nu_i = 0$ but $\mu_i \neq 0$. In this case we can take $H(\boldsymbol{\mu} \| \boldsymbol{\nu}) = \infty$.

As is customary in the information theory community, we use the same symbol H for both the 'absolute' entropy of one distribution as defined in (3.14), as well as the 'relative' entropy between two distributions. This dual usage should not cause confusion because the number of arguments of the H function should make it clear which usage is meant.

Actually, the quantity defined by Kullback and Leibler in their original paper [76] would correspond to $H(\boldsymbol{\mu} \| \boldsymbol{\nu}) + H(\boldsymbol{\nu} \| \boldsymbol{\mu})$, which would result in a quantity that is symmetric in $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. However, subsequent researchers have recognized the advantages of defining the divergence as in (3.29).

From the definition (3.29), it is not even clear that the K-L divergence is nonnegative. Since both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are probability distributions, it is clear that the components of both vectors add up to one. Hence, if $\boldsymbol{\mu} \neq \boldsymbol{\nu}$, then for some i the ratio μ_i/ν_i exceeds one and for other i it is less than one. So $\log(\mu_i/\nu_i)$ is positive for some i and negative for other i. Why then should the divergence be nonnegative? The question is answered next.

Theorem 3.19 *Suppose $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$ with $\boldsymbol{\mu} \ll \boldsymbol{\nu}$. Then $H(\boldsymbol{\mu} \| \boldsymbol{\nu}) \geq 0$, and $H(\boldsymbol{\mu} \| \boldsymbol{\nu}) > 0$ if $\boldsymbol{\mu} \neq \boldsymbol{\nu}$.*

Proof. Observe that log is a strictly concave function, whence $-\log$ is a strictly convex function. Therefore

$$\begin{aligned} H(\boldsymbol{\mu}\|\boldsymbol{\nu}) &= \sum_{i=1}^{n} \mu_i \log \frac{\mu_i}{\nu_i} \\ &= -\sum_{i=1}^{n} \mu_i \log \frac{\nu_i}{\mu_i} \\ &\geq -\log \left(\sum_{i=1}^{n} \mu_i \frac{\nu_i}{\mu_i} \right) \\ &= -\log \left(\sum_{i=1}^{n} \nu_i \right) \\ &= -\log(1) = 0. \end{aligned}$$

To prove the second part of the conclusion, note that if $\boldsymbol{\mu} \neq \boldsymbol{\nu}$, then there exist at least two indices i, j such that $\mu_i \neq \nu_i$, $\mu_j \neq \nu_j$. In this case, the quantity

$$\sum_{i=1}^{n} \mu_i \frac{\nu_i}{\mu_i}$$

is a *nontrivial convex combination* of the number $\nu_1/\mu_1, \ldots, \nu_n/\mu_n$. Thus the inequality in the above becomes strict, because $\log(\cdot)$ is a strictly concave function. □

A ready corollary of the above argument is the following.

Corollary 3.20 *Suppose $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$ are probability distributions with n components, and we define the 'loss' function*

$$J(\boldsymbol{\mu}, \boldsymbol{\nu}) := \sum_{i=1}^{n} \mu_i \log(1/\nu_i). \tag{3.30}$$

Then, viewed as a function of $\boldsymbol{\nu}$, the loss function assumes its minimum value when $\boldsymbol{\nu} = \boldsymbol{\mu}$, and the minimum value is $H(\boldsymbol{\mu})$. In other words,

$$J(\boldsymbol{\mu}, \boldsymbol{\nu}) \geq J(\boldsymbol{\mu}, \boldsymbol{\mu}) = H(\boldsymbol{\mu}) \ \forall \boldsymbol{\nu} \in \mathbb{S}_n. \tag{3.31}$$

Moreover, equality holds if and only if $\boldsymbol{\nu} = \boldsymbol{\mu}$.

Proof. Routine algebra shows that

$$H(\boldsymbol{\mu}\|\boldsymbol{\nu}) = J(\boldsymbol{\mu}, \boldsymbol{\nu}) - J(\boldsymbol{\mu}, \boldsymbol{\mu}).$$

Since we already know that $H(\boldsymbol{\mu}\|\boldsymbol{\nu}) \geq 0$ and that $H(\boldsymbol{\mu}\|\boldsymbol{\nu}) = 0$ if and only if $\boldsymbol{\mu} = \boldsymbol{\nu}$, it follows that $J(\boldsymbol{\mu}, \boldsymbol{\nu}) \geq J(\boldsymbol{\mu}, \boldsymbol{\mu})$ with equality if and only if $\boldsymbol{\nu} = \boldsymbol{\mu}$. It is routine to verify that $J(\boldsymbol{\mu}, \boldsymbol{\mu}) = H(\boldsymbol{\mu})$. □

The proof of Theorem 3.19 makes it clear that if we define the 'divergence' between two probability distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$ by replacing the logarithm function by *any* strictly convex function $\eta(\cdot)$ such that $\eta(1) = 0$,[3] the conclusions of Theorem 3.19 would still hold. In other words, if $\eta(\cdot)$ is strictly convex and satisfies $\eta(1) = 0$, and we define

$$D_\eta(\boldsymbol{\mu}\|\boldsymbol{\nu}) := -\sum_{i=1}^{n} \mu_i \eta(\nu_i/\mu_i),$$

then we would have $D_\eta(\boldsymbol{\mu}\|\boldsymbol{\nu}) \geq 0 \; \forall \boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$, and $D_\eta(\boldsymbol{\mu}\|\boldsymbol{\nu}) = 0$ if and only if $\boldsymbol{\mu} = \boldsymbol{\nu}$. The paper [79] contains a survey of the various 'divergences' that can be obtained for different choices of the function η, and their properties. However, Corollary 3.20 depends crucially on the fact that $\log(\nu_i/\mu_i) = \log \nu_i - \log \mu_i$ and thus does not hold for more general 'divergences.'

The K-L divergence has a very straight-forward interpretation in terms of the likelihood of misclassifying an observation. Suppose $\mathcal{X}$ is a random variable taking values in the set $\mathbb{A} = \{a_1, \ldots, a_n\}$. Suppose we are told that the random variable $\mathcal{X}$ has the distribution $\boldsymbol{\mu}$ or $\boldsymbol{\nu}$, and our challenge is to decide between the two competing hypotheses (namely '$\mathcal{X}$ is distributed according to $\boldsymbol{\mu}$' and '$\mathcal{X}$ is distributed according to $\boldsymbol{\nu}$'). For this purpose, we make a series of independent observations of $\mathcal{X}$, resulting in a 'sample path' $u_1^l := (u_1, u_2, \ldots, u_l)$ of length l, where each u_k belongs to $\mathbb{A}$.

In order to discriminate between the two competing hypotheses, we compute the **likelihood** of the observed sample path under each of the two hypotheses. Now, given the observation u_1^l of length l, let l_i denote the number of occurences of a_i in the sample path. Then the likelihood of this sample path, in case the underlying distribution is $\boldsymbol{\mu}$, equals

$$L_{\boldsymbol{\mu}}(u_1^l) = \prod_{i=1}^{n} \mu_i^{l_i}.$$

Similarly, the likelihood of this sample path in case the underlying distribution is $\boldsymbol{\nu}$ equals

$$L_{\boldsymbol{\nu}}(u_1^l) = \prod_{i=1}^{n} \nu_i^{l_i}.$$

We choose the hypothesis that is more likely. Thus, we say that the underlying distribution is $\boldsymbol{\mu}$ if $L_{\boldsymbol{\mu}}(u_1^l) > L_{\boldsymbol{\nu}}(u_1^l)$, and that the underlying distribution is $\boldsymbol{\nu}$ if $L_{\boldsymbol{\mu}}(u_1^l) < L_{\boldsymbol{\nu}}(u_1^l)$, For the time being, let us not worry about a 'tie,' because the probability of a tie approaches zero as $l \to \infty$. Now, instead of comparing $L_{\boldsymbol{\mu}}(u_1^l)$ and $L_{\boldsymbol{\nu}}(u_1^l)$, it is more customary to compute the **log likelihood ratio** $\log(L_{\boldsymbol{\mu}}(u_1^l)/L_{\boldsymbol{\nu}}(u_1^l))$. There are many reasons for this. From the standpoint of analysis, taking the logarithm turns products into sums, and thus the log likelihood function is often easier to analyze. From the standpoint of computation, the true likelihoods $L_{\boldsymbol{\mu}}(\mathbf{u})$ and $L_{\boldsymbol{\nu}}(\mathbf{u})$ often approach zero at a geometric rate as $l \to \infty$. So if l is at all a large

[3] This by itself is not a limitation, because if $\eta(\cdot)$ is a strictly convex function, so is $\eta(\cdot) - \eta(1)$.

number, then $L_{\boldsymbol{\mu}}(u_1^l)$ and $L_{\boldsymbol{\nu}}(u_1^l)$ rapidly fall below the machine zero, thus making computation difficult.

Now we can compute that

$$\log(L_{\boldsymbol{\mu}}(u_1^l)/L_{\boldsymbol{\nu}}(u_1^l)) = \sum_{i=1}^{n} l_i \log(\mu_i/\nu_i).$$

However, the indices l_i are themselves random numbers, as they correspond to the number of times that the outcome a_i appears in the sample path. Suppose now that 'the truth' is $\boldsymbol{\mu}$. In other words, the underlying probability distribution really is $\boldsymbol{\mu}$. Then the expected value of l_i equals $\mu_i l$. Thus the expected value of the log likelihood ratio, when 'the truth' is $\boldsymbol{\mu}$, equals

$$E[\log(L_{\boldsymbol{\mu}}(u_1^l)/L_{\boldsymbol{\nu}}(u_1^l)), P_\mu] = l \sum_{i=1}^{n} \mu_i \log(\mu_i/\nu_i) = lH(\boldsymbol{\mu}\|\boldsymbol{\nu}).$$

In other words, the expected value of the log likelihood ratio equals $H(\boldsymbol{\mu}\|\boldsymbol{\nu})$ multiplied by the length of the observation. For this reason, we can also interpret $H(\boldsymbol{\mu}\|\boldsymbol{\nu})$ as the 'per sample' contribution to the log likelihood ratio. This computation shows us two things. First, if we are trying to choose between two competing hypotheses where one of them is the 'truth' and the other is not, then 'in the long run' we will choose the truth. Moreover, 'the farther' the other competing hypothesis is from the 'truth,' the more quickly the log-likelihood classifier will zero in on the 'truth.'

We shall return to this argument again in Chapter 7, when we discuss large deviation theory.

This interpretation can be extended to a more general situation. The interpretation of the K-L divergence given above assumes that one of the two competing hypotheses is in fact the truth. What if this assumption does not hold? Suppose the samples are generated by an underlying probability distribution $\boldsymbol{\mu}$, and the competing hypotheses are that the distribution is $\boldsymbol{\nu}$ or $\boldsymbol{\theta}$. Then

$$\begin{aligned} E[\log(L_{\boldsymbol{\nu}}(u_1^l)/L_{\boldsymbol{\theta}}(u_1^l)), P_\mu] &= l \sum_{i=1}^{n} \mu_i \log(\nu_i/\theta_i) \\ &= \sum_{i=1}^{n} \mu_i \log\left(\frac{\nu_i}{\mu_i}\right) - \sum_{i=1}^{n} \mu_i \log\left(\frac{\theta_i}{\mu_i}\right) \\ &= H(\boldsymbol{\mu}\|\boldsymbol{\theta}) - H(\boldsymbol{\mu}\|\boldsymbol{\nu}). \end{aligned}$$

Hence, in the long run, the maximum likelihood classifier would choose $\boldsymbol{\nu}$ if $H(\boldsymbol{\mu}\|\boldsymbol{\nu}) < H(\boldsymbol{\mu}\|\boldsymbol{\theta})$, and choose $\boldsymbol{\theta}$ if the inequality is reversed. In other words, the K-L divergence induces a partial order on the set of probability distributions. Given a fixed $\boldsymbol{\mu} \in \mathbb{S}_n$, all the other $\boldsymbol{\nu} \in \mathbb{S}_n$ can be ranked in terms of the divergence $H(\boldsymbol{\mu}\|\boldsymbol{\nu})$. If we try to choose between two competing hypotheses $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ when the 'truth' is $\boldsymbol{\mu}$, in the long run we would choose the one that is 'closer' to the 'truth.'

Unfortunately the K-L divergence is *not* symmetric in general. That is, in general

$$H(\boldsymbol{\mu}\|\boldsymbol{\nu}) \neq H(\boldsymbol{\nu}\|\boldsymbol{\mu}).$$

Moreover, in general

$$H(\boldsymbol{\mu}\|\boldsymbol{\nu}) + H(\boldsymbol{\nu}\|\boldsymbol{\theta}) \not\geq H(\boldsymbol{\mu}\|\boldsymbol{\theta}).$$

Hence we cannot use the K-L divergence to define any kind of a 'metric' distance between probability measures. Nevertheless, the K-L divergence is very useful, because it quantifies the probability of misclassification using the log-likelihood criterion. Again, it is shown in [79] that if the log function is replaced by some other strictly convex function, then it is possible for the resulting 'divergence' to satisfy the triangle inequality. However, all the other nice features such as the interpretation in terms of the log-likelihood function would be lost.

In Definition 3.15, we introduced the notion of 'mutual information' between two random variables. It is now shown that the mutual information can also be defined in terms of the K-L divergence.

Theorem 3.21 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A}, \mathbb{B}$ respectively, and let $\boldsymbol{\phi}$ denote their joint distribution. Then the mutual information $I(\mathcal{X}, \mathcal{Y})$ is given by*

$$I(\mathcal{X}, \mathcal{Y}) = H(\boldsymbol{\phi}\|\boldsymbol{\phi}_{\mathcal{X}} \times \boldsymbol{\phi}_{\mathcal{Y}}).$$

Proof. By definition we have that

$$I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H((\mathcal{X}, \mathcal{Y})).$$

Now substitute directly in terms of $\boldsymbol{\phi}$, and observe that

$$(\boldsymbol{\phi}_{\mathcal{X}})_i = \sum_{j=1}^{m} \phi_{ij}, (\boldsymbol{\phi}_{\mathcal{Y}})_j = \sum_{i=1}^{n} \phi_{ij}.$$

This gives

$$\begin{aligned} I(\mathcal{X}, \mathcal{Y}) &= -\sum_{i=1}^{n} \left[\sum_{j=1}^{m} \phi_{ij} \right] \log(\boldsymbol{\phi}_{\mathcal{X}})_i - \sum_{j=1}^{m} \left[\sum_{i=1}^{n} \phi_{ij} \right] \log(\boldsymbol{\phi}_{\mathcal{Y}})_j \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log \phi_{ij} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{ij} \log \left[\frac{\phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{X}})_i (\boldsymbol{\phi}_{\mathcal{Y}})_j} \right] \\ &= H(\boldsymbol{\phi}\|\boldsymbol{\phi}_{\mathcal{X}} \times \boldsymbol{\phi}_{\mathcal{Y}}). \end{aligned}$$

□

The above line of reasoning is extended in the next result, which is sometimes referred to as the 'information inequality'.

Theorem 3.22 *Suppose $\mathbb{A}, \mathbb{B}$ are finite sets, that $\boldsymbol{\mu}_{\mathbb{A}}, \boldsymbol{\mu}_{\mathbb{B}}$ are distributions on $\mathbb{A}, \mathbb{B}$ respectively, and that $\boldsymbol{\nu}$ is a distribution on the product set $\mathbb{A} \times \mathbb{B}$. Then*

$$H(\boldsymbol{\nu}\|\boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}) \geq H(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}}) + H(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}}). \tag{3.32}$$

with equality if and only if $\boldsymbol{\nu}$ is a product distribution, i.e. $\boldsymbol{\nu} = \boldsymbol{\nu}_{\mathbb{A}} \times \boldsymbol{\nu}_{\mathbb{B}}$.

Proof. Let us write the K-L divergence in terms of the loss function, and note that $H(\boldsymbol{\nu}) \leq H(\boldsymbol{\nu}_{\mathbb{A}}) + H(\boldsymbol{\nu}_{\mathbb{B}})$. This gives

$$\begin{aligned} H(\boldsymbol{\nu} \| \boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}) &= J(\boldsymbol{\nu} \| \boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}) - H(\boldsymbol{\nu}) \\ &\geq J(\boldsymbol{\nu} \| \boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}) - H(\boldsymbol{\nu}_{\mathbb{A}}) - H(\boldsymbol{\nu}_{\mathbb{B}}) \\ &= J(\boldsymbol{\nu}_{\mathbb{A}} \| \boldsymbol{\mu}_{\mathbb{A}}) + J(\boldsymbol{\nu}_{\mathbb{B}} \| \boldsymbol{\mu}_{\mathbb{B}}) - H(\boldsymbol{\nu}_{\mathbb{A}}) - H(\boldsymbol{\nu}_{\mathbb{B}}) \\ &= H(\boldsymbol{\nu}_{\mathbb{A}} \| \boldsymbol{\mu}_{\mathbb{A}}) + H(\boldsymbol{\nu}_{\mathbb{B}} \| \boldsymbol{\mu}_{\mathbb{B}}). \end{aligned}$$

Here we make use of the easily proved fact

$$J(\boldsymbol{\nu} \| \boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}) = J(\boldsymbol{\nu}_{\mathbb{A}} \| \boldsymbol{\mu}_{\mathbb{A}}) + J(\boldsymbol{\nu}_{\mathbb{B}} \| \boldsymbol{\mu}_{\mathbb{B}}).$$

In order for the inequality to be an equality, we must have $H(\boldsymbol{\nu}) = H(\boldsymbol{\nu}_{\mathbb{A}}) + H(\boldsymbol{\nu}_{\mathbb{B}})$, which is the case if and only if $\boldsymbol{\nu}$ is a product distribution, i.e. $\boldsymbol{\nu} = \boldsymbol{\nu}_{\mathbb{A}} \times \boldsymbol{\nu}_{\mathbb{B}}$. □

The next theorem gives a nice expression for the K-L divergence between two probability distributions of joint random variables. In [26], p. 24, this result is referred to as the "chain rule" for the K-L divergence.

Theorem 3.23 *Suppose $\boldsymbol{\phi}, \boldsymbol{\theta}$ are probability distributions on a product set $\mathbb{A} \times \mathbb{B}$, where $|\mathbb{A}| = n$ and $|\mathbb{B}| = m$. Suppose $\boldsymbol{\phi} \ll \boldsymbol{\theta}$. Let $\boldsymbol{\phi}_{\mathcal{X}}, \boldsymbol{\theta}_{\mathcal{X}}$ denote the marginal distributions on $\mathbb{A}$. Then*

$$H(\boldsymbol{\phi} \| \boldsymbol{\theta}) = H(\boldsymbol{\phi}_{\mathcal{X}} \| \boldsymbol{\theta}_{\mathcal{X}}) + \sum_{i=1}^{n} (\boldsymbol{\phi}_{\mathcal{X}})_i H(\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} \| \boldsymbol{\theta}_{\{\mathcal{Y}|\mathcal{X}=a_i\}}). \tag{3.33}$$

Proof. To simplify the notation, let us use the symbols

$$f_i := \sum_{j=1}^{m} \phi_{ij}, 1 \leq i \leq n, \mathbf{f} := [f_1 \ldots f_n] \in \mathbb{S}_n,$$

$$g_i := \sum_{j=1}^{m} \theta_{ij}, 1 \leq i \leq n, \mathbf{g} := [g_1 \ldots g_n] \in \mathbb{S}_n.$$

Thus both $\mathbf{f}$ and $\mathbf{g}$ are probability distributions on $\mathbb{A}$. Next, for each i between 1 and n, define

$$c_{ij} := \frac{\phi_{ij}}{f_i}, d_{ij} := \frac{\theta_{ij}}{g_i}, 1 \leq j \leq m.$$

$$\mathbf{c}_i := [c_{i1} \ldots c_{im}], \mathbf{d}_i := [d_{i1} \ldots d_{im}].$$

Then clearly

$$\boldsymbol{\phi}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} = \mathbf{c}_i, \text{ and } \boldsymbol{\theta}_{\{\mathcal{Y}|\mathcal{X}=a_i\}} = \mathbf{d}_i.$$

As per the conventions established earlier, if $f_i = 0$ for some i, we take $\mathbf{c}_i = \boldsymbol{\phi}_{\mathcal{Y}}$. In other words, when conditioned on the 'impossible' event $\mathcal{X} = a_i$, the conditional distribution of $\mathcal{Y}$ is taken as its marginal distribution. Similar remarks apply in case $g_i = 0$ for some i. Next let us study the dominance between these conditional distributions. If $g_i = 0$ for some index i, then it follows that $\theta_{ij} = 0 \ \forall j$. Since $\boldsymbol{\phi} \ll \boldsymbol{\theta}$, this implies that $\phi_{ij} = 0 \ \forall j$, i.e., that $f_i = 0$. In other words, the dominance condition $\boldsymbol{\phi} \ll \boldsymbol{\theta}$ implies that $\boldsymbol{\phi}_{\mathcal{X}} \ll \boldsymbol{\theta}_{\mathcal{X}}$. The same dominance condition $\boldsymbol{\phi} \ll \boldsymbol{\theta}$ also implies that $\mathbf{c}_i \ll \mathbf{d}_i \ \forall i$.

Now, in terms of the new symbols, the desired conclusion (3.33) can be rewritten as

$$H(\boldsymbol{\phi}\|\boldsymbol{\theta}) = H(\mathbf{f}\|\mathbf{g}) + \sum_{i=1}^{n} f_i H(\mathbf{c}_i\|\mathbf{d}_i).$$

This relationship can be established simply by expanding $H(\boldsymbol{\phi}\|\boldsymbol{\theta})$. Note that

$$\begin{aligned} H(\boldsymbol{\phi}\|\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij} \log\left(\frac{\phi_{ij}}{\theta_{ij}}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij} \log\left(\frac{f_i c_{ij}}{g_i d_{ij}}\right) \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{m} \phi_{ij}\right) \log\frac{f_i}{g_i} + \sum_{i=1}^{n}\sum_{j=1}^{m} f_i c_{ij} \log\left(\frac{c_{ij}}{d_{ij}}\right) \\ &= H(\mathbf{f}\|\mathbf{g}) + \sum_{i=1}^{n} f_i H(\mathbf{c}_i\|\mathbf{d}_i). \end{aligned}$$

This is the desired relationship. □

We conclude this chapter with a couple of useful properties of the K-L divergence.

Theorem 3.24 *The K-L divergence $H(\boldsymbol{\phi}\|\boldsymbol{\theta})$ is jointly convex in $\boldsymbol{\phi}, \boldsymbol{\theta}$; that is, if $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{S}_n$ with $\boldsymbol{\phi}_1 \ll \boldsymbol{\theta}_1$, $\boldsymbol{\phi}_2 \ll \boldsymbol{\theta}_2$ and $\lambda \in (0,1)$, we have*

$$H(\lambda\boldsymbol{\phi}_1 + (1-\lambda)\boldsymbol{\phi}_2\|\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) \leq \lambda H(\boldsymbol{\phi}_1\|\boldsymbol{\theta}_1) + (1-\lambda)H(\boldsymbol{\phi}_2\|\boldsymbol{\theta}_2). \quad (3.34)$$

The proof of Theorem 3.24 depends on an auxiliary lemma called the 'log sum inequality' that is of independent interest.

Lemma 3.25 *Suppose $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m$ are nonnegative, and that at least one α_i and at least one β_j are positive. Assume further that $\beta_i = 0 \Rightarrow \alpha_i = 0$. Then*

$$\sum_{i=1}^{m} \alpha_i \log\left(\frac{\alpha_i}{\beta_i}\right) \geq \left(\sum_{i=1}^{m} \alpha_i\right) \cdot \log\left(\frac{\sum_{i=1}^{m} \alpha_i}{\sum_{j=1}^{m} \beta_j}\right). \quad (3.35)$$

Remark: Note that neither the α_i's nor the β_i's need to add up to one in order for (3.35) to hold.

Proof. (of Lemma 3.25) Define

$$A = \sum_{i=1}^{n} \alpha_i, B = \sum_{i=1}^{n} \beta_i, a_i = \frac{\alpha_i}{A}, b_i = \frac{\beta_i}{B}, i = 1, \ldots, n.$$

Then the vectors $\mathbf{a} = [a_1 \ldots a_n], \mathbf{b} = [b_1 \ldots b_n] \in \mathbb{S}_n$. Thus it follows from Theorem 3.19 that

$$\sum_{i=1}^{n} a_i \log \left(\frac{a_i}{b_i} \right) \geq 0.$$

Now let us substitute $\alpha_i = Aa_i, \beta_i = Bb_i$ for all i. This leads to

$$\begin{aligned}
\sum_{i=1}^{m} \alpha_i \log \left(\frac{\alpha_i}{\beta_i} \right) &= A \left[\sum_{i=1}^{n} a_i \log \left(\frac{a_i}{b_i} \right) + \sum_{i=1}^{n} a_i \log \frac{A}{B} \right] \\
&= A \left[\sum_{i=1}^{n} a_i \log \left(\frac{a_i}{b_i} \right) + \log \frac{A}{B} \right] \text{ because } \sum_{i=1}^{n} a_i = 1 \\
&\geq A \log \frac{A}{B}.
\end{aligned}$$

This is precisely (3.35). □

Proof. (of Theorem 3.24): The left side of (3.34) is the sum over $i = 1, \ldots, n$ of the term

$$g_i := [\lambda \phi_{1i} + (1 - \lambda) \phi_{2i}] \cdot \log \left(\frac{\lambda \phi_{1i} + (1 - \lambda) \phi_{2i}}{\lambda \theta_{1i} + (1 - \lambda) \theta_{2i}} \right).$$

Now apply the log sum inequality with $m = 2$ and

$$\alpha_{1i} = \lambda \phi_{1i}, a_{2i} = (1 - \lambda) \phi_{2i}, \beta_{1i} = \lambda \theta_{1i}, b_{2i} = (1 - \lambda) \theta_{2i}.$$

Note that since $\boldsymbol{\phi}_1 \ll \boldsymbol{\theta}_1$, $\boldsymbol{\phi}_2 \ll \boldsymbol{\theta}_2$, it follows that $\beta_{ji} = 0 \;\Rightarrow\; \alpha_{ji} = 0$ for $j = 1, 2$. Then it follows from (3.35) that, for each i, we have

$$\begin{aligned}
g_i &= (\alpha_{1i} + \alpha_{2i}) \log \frac{\alpha_{1i} + \alpha_{2i}}{\beta_{1i} + \beta_{2i}} \\
&\leq \alpha_{1i} \log \frac{\alpha_{1i}}{\beta_{1i}} + \alpha_{2i} \log \frac{\alpha_{2i}}{\beta_{2i}} \\
&= \lambda \phi_{1i} \log \frac{\phi_{1i}}{\theta_{1i}} + (1 - \lambda) \phi_{2i} \log \frac{\phi_{2i}}{\theta_{2i}}.
\end{aligned}$$

Summing over $i = 1, \ldots, n$ shows that

$$\sum_{i=1}^{n} g_i \leq \lambda H(\boldsymbol{\phi}_1 \| \boldsymbol{\theta}_1) + (1 - \lambda) H(\boldsymbol{\phi}_2 \| \boldsymbol{\theta}_2),$$

which is the desired inequality. □

The last result of this chapter relates the total variation metric and the Kullback-Leibler divergence.

Theorem 3.26 *Suppose $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathbb{S}_n$, and that $\boldsymbol{\mu} \ll \boldsymbol{\nu}$. Then*

$$\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq [(1/2)H(\boldsymbol{\mu}\|\boldsymbol{\nu})]^{1/2}. \tag{3.36}$$

Proof. Let us view $\boldsymbol{\mu}, \boldsymbol{\nu}$ as distributions on some finite set $\mathbb{A}$. Define $\mathbb{A}_+ := \{i : \mu_i \geq \nu_i\}$, and $\mathbb{A}_- := \{i : \mu_i < \nu_i\}$. Then together $\mathbb{A}_+, \mathbb{A}_-$ partition $\mathbb{A}$. Next, let us define

$$p := \sum_{i \in \mathbb{A}_+} \mu_i, q := \sum_{i \in \mathbb{A}_+} \nu_i,$$

and observe that $p \geq q$ by the manner in which we have defined the two sets. Moreover, it follows from (2.12) that

$$\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) = p - q.$$

Next, by definition we have

$$\begin{aligned} H(\boldsymbol{\mu}\|\boldsymbol{\nu}) &= \sum_{i \in \mathbb{A}_+} \mu_i \log \left(\frac{\mu_i}{\nu_i} \right) + \sum_{i \in \mathbb{A}_-} \mu_i \log \left(\frac{\mu_i}{\nu_i} \right) \\ &\geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \end{aligned}$$

The last step follows from the log sum inequality because

$$\sum_{i \in \mathbb{A}_+} \mu_i \log \left(\frac{\mu_i}{\nu_i} \right) \geq \left(\sum_{i \in \mathbb{A}_+} \mu_i \right) \log \left(\frac{\sum_{i \in \mathbb{A}_+} \mu_i}{\sum_{i \in \mathbb{A}_+} \nu_i} \right) = p \log \frac{p}{q},$$

and similarly for the other term.

Thus the proof of the inequality (3.36) can be achieved by studying only distributions on a set of cardinality two. Define $\boldsymbol{\phi} = (p, 1-p), \boldsymbol{\psi} = (q, 1-q)$ and suppose $p \geq q$. The objective is to prove that

$$p - q \leq [(1/2)H(\boldsymbol{\phi}\|\boldsymbol{\psi})]^{1/2},$$

or equivalently that

$$2(p-q)^2 \leq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \ \forall q \in (0, p), \ \forall p \in (0, 1). \tag{3.37}$$

To prove (3.37), define

$$f(q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - 2(p-q)^2,$$

viewed as a function of q for a fixed p. Then the objective is to show that $f(q) \geq 0$ for all $q \in (0, p]$. Clearly $f(p) = 0$. Moreover,

$$\begin{aligned} f'(q) &= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \\ &= \frac{-p(1-q) + (1-p)q + 4(p-q)q(1-q)}{q(1-q)} \\ &= -(p-q)\frac{4q^2 - 4q + 1}{q(1-q)} \\ &= -(p-q)\frac{(2q-1)^2}{q(1-q)}. \end{aligned}$$

Hence $f'(p) = 0$. Moreover, since $(2q-1)^2 \geq 0, q(1-q) \geq 0$ always, it follows that $f'(q) \leq 0 \; \forall q < p$. Hence $f(q) \geq 0 \; \forall q < p$. □

The result above is universally known as 'Pinsker's inequality', but it would be fairer to credit also Csiszar; see [29]. It turns out that if we seek a bound of the form

$$H(\boldsymbol{\mu} \| \boldsymbol{\nu}) \geq C[\rho(\boldsymbol{\mu}, \boldsymbol{\nu})]^2,$$

then $C = 2$ is the best possible constant. More generally, suppose we seek a bound of the form

$$H(\boldsymbol{\mu} \| \boldsymbol{\nu}) \geq \sum_{i=1}^{\infty} C_{2i} V^{2i},$$

where $V = \rho(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then it is shown in [104] that

$$H(\boldsymbol{\mu} \| \boldsymbol{\nu}) \geq \frac{1}{2} V^2 + \frac{1}{36} V^4 + \frac{1}{270} V^6 + \frac{221}{340200} V^8.$$

The same paper also gives a general methodology for extending the power series further, if one has the patience.

A couple of final comments. First, since the total variation metric is symmetric while the K-L divergence is not, we can also write (3.36) as

$$\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \max\{[(1/2)H(\boldsymbol{\mu} \| \boldsymbol{\nu})]^{1/2}, [(1/2)H(\boldsymbol{\nu} \| \boldsymbol{\mu})]^{1/2}\}. \tag{3.38}$$

Second, there is no *lower bound* for the total variation metric in terms of the K-L divergence. Indeed, there cannot be. Just choose $\boldsymbol{\nu}, \boldsymbol{\mu}$ such that for some index i we have $\nu_i = 0$ but μ_i is very small but positive. Then $\rho(\boldsymbol{\nu}, \boldsymbol{\mu})$ is small but $H(\boldsymbol{\mu} \| \boldsymbol{\nu})$ is infinite. So there cannot exist a lower bound analogous to (3.38).

Chapter Four

Nonnegative Matrices

In this chapter, the focus is on nonnegative matrices. They are relevant in the study of Markov processes (see Chapter 5), because the state transition matrix of a Markov process is a special kind of nonnegative matrix, known as a stochastic matrix.[1] However, it turns out that practically all of the useful properties of a stochastic matrix also hold for the more general class of nonnegative matrices. Hence it is desirable to present the theory in the more general setting, and then specialize to Markov processes. The reader may find the more general results useful in some other context.

Nonnegative matrices have two very useful properties. First, through symmetric row and column permutations, every nonnegative matrix can be put into a corresponding "canonical form" that is essentially unique. Second, the eigenvalues of a nonnegative matrix, subject to some additional conditions, have a very special structure. The two sections of the present chapter are devoted respectively to these two properties.

4.1 CANONICAL FORM FOR NONNEGATIVE MATRICES

In this section it is shown that to every nonnegative matrix A there corresponds an essentially unique canonical form C. Moreover, A can be transformed into C via symmetric row and column permutations.

4.1.1 Basic Version of the Canonical Form

Let $\mathbb{R}_+$ denote the set $[0, \infty)$ of nonnegative numbers. Suppose $A \in \mathbb{R}_+^{n \times n}$ is a given nonnegative matrix. It turns out that the canonical form associated with A depends only on which elements of A are positive, but does not otherwise depend on the *size* of the positive elements. Thus, to capture the *location* of the positive elements of A, we defind the associated **incidence matrix** T corresponding to A as follows: $T \in \{0, 1\}^{n \times n}$, and

$$t_{ij} = \begin{cases} 1 & \text{if} \quad a_{ij} > 0, \\ 0 & \text{if} \quad a_{ij} = 0. \end{cases}$$

Since A has n rows and columns, we can think of $\mathcal{N} := \{1, \ldots, n\}$ as the set of nodes of a directed graph, and place an edge from node i to node j if and only if $a_{ij} > 0$ (or equivalently, $t_{ij} = 1$). A **path** of length l from node

[1] Both of these terms are defined in Chapter 5.

i to node j is a sequence of pairs $\{(n_0, n_1), \ldots, (n_{l-1}, n_l)\}$, where $n_0 = i$, $n_l = j$, and in addition $a_{n_s, n_{s+1}} > 0$ for all $s = 0, \ldots, l-1$. Let $a_{ij}^{(l)}$ denote the ij-th element of A^l. From the matrix multiplication formula

$$a_{ij}^{(l)} = \sum_{s_1=1}^{n} \cdots \sum_{s_{l-1}=1}^{n} a_{is_1} a_{s_1 s_2} \cdots a_{s_{l-1} j}$$

it is obvious that $a_{ij}^{(l)} > 0$ if and only if there exists a path of length l from i to j.

At this point the reader may wonder why we don't define a path from i to j as a sequence *of length* $\leq n$. Clearly, in any sequence of length $\geq n+1$, there must exist at least one cycle (a path whose starting and end nodes are the same). This cycle can be removed without affecting the existence of a path from i to j. In short, there exists a path from i to $j \neq i$ if and only if there exists a path of length $\leq n-1$ from i to j. There exists a path from i back to i if and only if there exists a path from i to itself of length $\leq n$. So why then do we not restrict the length of the path to be $\leq n$?

The answer is found in the sentence at the end of the next to previous paragraph. Given a matrix A, we wish to study the pattern of nonzero (or positive) elements of successive power A^l for *all* values of l, not just when $l \leq n$. Note that the statement "$a_{ij}^{(l)} > 0$ if and only if there exists a path of length l from i to j" is valid for *every* value of l, even if $l \geq n+1$.

We say that a node i **leads to** another node j if there exists a path of *positive* length from i to j. We write $i \to j$ if i leads to j. In particular, i leads to i if and only if there is a cycle from node i to itself. Thus it is quite possible for a node *not* to lead to itself.

A node i is said to be **inessential** if there exists a j (of necessity, not equal to i) such that $i \to j$, but $j \not\to i$. Otherwise, i is said to be **essential**. Note that if i is essential, i leads to j implies j leads to i. By convention, if a node i does not lead to any other node j, then i is taken to be inessential. For instance, if row i of the matrix A is identically zero, then i is inessential.

With the above definitions, we can divide the set of nodes $\mathcal{N} = \{1, \ldots, n\}$ into two disjoint sets: $\mathcal{I}$ denoting the set of inessential nodes, and $\mathcal{E}$ denoting the set of essential nodes.

Now it is claimed that if $i \in \mathcal{E}$ and $i \to j \in \mathcal{N}$, then $j \in \mathcal{E}$. In other words, an essential node can lead only to another essential node. The proof of this claim makes use of the property that the relation $\to$ is transitive, or in other words, $i \to j$ and $j \to k$ implies that $i \to k$. The transitivity of $\to$ is clear from the definition of a path, but we give below an algebraic proof. If $i \to j$, then there is an integer l such that $a_{ij}^{(l)} > 0$. Similarly, if $j \to k$, then there exists another integer r such that $a_{jk}^{(r)} > 0$. Now from the formula for matrix multiplication, it follows that

$$a_{ik}^{(l+r)} = \sum_{s=1}^{n} a_{is}^{(l)} a_{sk}^{(r)} \geq a_{ij}^{(l)} a_{jk}^{(r)} > 0.$$

Thus $i \to k$. Coming back to the main issue, we are given that $i \in \mathcal{E}$ and that $i \to j \in \mathcal{N}$; we wish to show that $j \in \mathcal{E}$. For this purpose, we must show that if $j \to k \in \mathcal{N}$, then $k \to j$. Accordingly, suppose that $j \to k$. Since $i \to j$ and $j \to k$, the transitivity of $\to$ implies that $i \to k$. Next, since $i \in \mathcal{E}$, it follows that $k \to i$. Invoking once again the transivity property, we conclude from $k \to i$ and $i \to j$ that $k \to j$. Since k is arbitrary other than that $j \to k$, this shows that $j \in \mathcal{E}$.

We have just shown that if $i \in \mathcal{E}$ and $i \to j$, then $j \in \mathcal{E}$. As a consequence, if $i \in \mathcal{E}$ and $j \in \mathcal{I}$, then $i \not\to j$. In particular, if $i \in \mathcal{E}$ and $j \in \mathcal{I}$, then $a_{ij} = 0$. Otherwise, if $a_{ij} > 0$, then $i \to j$ (because there is a path of length one), which is a contradiction. We can now summarize this observation through a very preliminary version of the canonical form.

Lemma 4.1 *Renumber the rows and columns of A in such a way that all the nodes in $\mathcal{E}$ come first, followed by all the nodes in $\mathcal{I}$. Let Π denote the permutation matrix corresponding to the renumbering. Then*

$$\Pi^{-1} A \Pi = \begin{array}{cc} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} P & \mathbf{0} \\ R & Q \end{array} \right] \end{array}. \tag{4.1}$$

Example 4.1 Suppose a 10×10 matrix A has the following structure, where $\times$ denotes a positive element. This notation is chosen deliberately to highlight the fact that the actual values of the elements of A don't matter, only whether they are positive or zero. Note that the incidence matrix corresponding to A can be obtained simply by changing all occurences of $\times$ to 1.

$$A = \left[\begin{array}{cccccccccc} 0 & \times & 0 & 0 & \times & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & \times & 0 & 0 & 0 & \times & 0 & \times & 0 \\ 0 & \times & 0 & \times & 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 \end{array} \right].$$

With this matrix we can associate a directed graph where there is an edge from node i to another node j if and only if $a_{ij} > 0$.

Let us compute the **reachability matrix** M of this graph, where we write

$m_{ij} = +$ if $i \to j$ and $m_{ij} = \cdot$ if $i \not\to j$. Thus

$$M = \begin{bmatrix} \cdot & + & + & + & + & + & + & + & + & + \\ \cdot & + & + & + & + & + & + & + & + & + \\ \cdot & + & + & + & + & + & + & + & + & + \\ \cdot & \cdot & \cdot & + & \cdot & + & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & + & \cdot & + & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & + & \cdot & + & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & + & \cdot & + & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & + & \cdot & + & \cdot & + & + & + \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & + & + \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & + & + \end{bmatrix}.$$

From the reachability matrix we can see that nodes $1, 2, 3$ are inessential, whereas nodes 4 through 10 are essential. So if we simply make a cyclic permuation and shift 1, 2, 3 to the end, the matrix A now looks like

$$\Pi^{-1}A\Pi = \begin{array}{c|ccccccc|ccc|} & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 \\ 4 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & \times & 0 & 0 & \times & 0 & 0 & 0 & \times & 0 \\ 2 & 0 & 0 & 0 & \times & 0 & \times & 0 & 0 & 0 & \times \\ 3 & \times & 0 & \times & \times & 0 & 0 & 0 & 0 & \times & 0 \end{array}$$

Thus the triangular structure of A is brought out clearly. As we develop the theory further, we shall return to this example and refine further the matrix on the right side.

Example 4.2 The purpose of this example is to demonstrate the possibility that *every* node can be inessential. Suppose

$$A = \begin{bmatrix} 0 & \times \\ 0 & 0 \end{bmatrix}.$$

Then it is easy to see that both nodes are inessential.

The next lemma shows when such a phenomenon can occur.

Lemma 4.2 *Suppose $A \in \mathbb{R}_+^{n \times n}$ and that every row of A contains at least one positive element. Then there exists at least one essential node.*

Proof. Suppose by way of contradiction that every node is inessential. Recall that there are two ways in which a node i can be inessential: (i) i does not lead to any other node, or (ii) there exists a $j \in \mathcal{N}$ such that $i \to j$ but $j \not\to i$. By assumption, no row of A is identically zero. Hence every node i

leads to at least one other node j (which could be the same as i). Hence the first possibility is ruled out. Now choose a node $i_0 \in \mathcal{N}$ arbitrarily. Since i_0 is inessential, there exists a node $i_1 \in \mathcal{N}$ such that $i_0 \to i_1$ but $i_1 \not\to i_0$. Clearly this implies that $i_1 \neq i_0$. Now i_1 is also inessential. Hence there exists an $i_2 \in \mathcal{N}$ such that $i_1 \to i_2$ but $i_2 \not\to i_1$. This implies, *inter alia*, that $i_2 \neq i_1$ and that $i_2 \neq i_0$. It is obvious that $i_2 \neq i_1$. If $i_2 = i_0$ then $i_2 = i_0 \to i_1$, which is a contradiction. Now repeat the procedure. At step l, we have $l+1$ nodes $i_0, i_1, \ldots, i_l$ such that

$$i_0 \to i_1 \to \cdots \to i_{l-1} \to i_l, \text{ but } i_l \not\to i_{l-1}.$$

We claim that this implies that i_l is not equal to any of $i_0, \ldots, i_{l-1}$; otherwise, the transitivity property of $\to$ implies that we have $i_l \to i_{l-1}$, which is a contradiction. Hence all the $l+1$ nodes are distinct. After we repeat the process n times, we will supposedly have $n+1$ distinct nodes $i_0, i_1, \ldots, i_n$. But this is impossible since there are only n nodes. Hence it is not possible for every node to be inessential. □

Lemma 4.3 *Suppose $A \in \mathbb{R}_+^{n \times n}$ and that every row of A contains at least one positive element. Then every inessential node (if any) leads to an essential node.*

Proof. As before let $\mathcal{E}$ denote the set of essential nodes and $\mathcal{I}$ the set of inessential nodes. From Lemma 4.2, we know that $\mathcal{E}$ is nonempty. If $\mathcal{I}$ is nonempty, the lemma states that for all $i \in \mathcal{I}$ there exists a $j \in \mathcal{E}$ such that $i \to j$. Accordingly, suppose $i \in \mathcal{I}$ is arbitrary. Since every row of A contains at least one positive element, node i leads to at least one other node. Since node i is inessential, there exists a $j \in \mathcal{N}$ such that $i \to j$ but $j \not\to i$. If $j \in \mathcal{E}$ the claim is established, so suppose $j \in \mathcal{I}$. So there exists a $k \in \mathcal{N}$ such that $j \to k$ but $k \not\to j$. Clearly $k \neq j$ and also $k \neq i$ since $i \to j$. If $k \in \mathcal{E}$, then $i \to j \to k$ and the claim is established, so suppose $k \in \mathcal{I}$. Proceeding as before, we can choose $l \in \mathcal{N}$ such that $k \to l$ but $l \not\to k$. Clearly $l \neq i, j, k$. If $l \in \mathcal{E}$ then $i \to j \to k \to l$ implies $i \to l \in \mathcal{E}$. Proceeding in this fashion, we will construct a sequence of nodes, all of them distinct, and all belonging to $\mathcal{I}$. Since $\mathcal{I}$ is a finite set, sooner or later this process must stop with some node $l \in \mathcal{E}$. Hence $i \to l \in \mathcal{E}$. Then it follows from Lemma 4.3 that there exists a $j \in \mathcal{E}$ such that $a_{ij} > 0$. □

These two lemmas lead to a slightly refined version of the canonical form.

Lemma 4.4 *Suppose $A \in \mathbb{R}_+^{n \times n}$ and that every row of A contains at least one positive element. Then in the canonical form (4.1), the set $\mathcal{E}$ is nonempty. Moreover, if the set $\mathcal{I}$ is also nonempty, then*

1. *The matrix R contains at least one positive element, i.e., $R \neq \mathbf{0}$.*

2. *Let $m = |\mathcal{I}|$, the number of inessential nodes. Express A^m in the form*

$$\Pi^{-1} A^m \Pi = \begin{array}{cc} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} P^m & \mathbf{0} \\ R^{(m)} & Q^m \end{array} \right] \end{array}. \tag{4.2}$$

Then each row of $R^{(m)}$ contains at least one nonzero element.

Proof. The fact that $\mathcal{E}$ is nonempty follows from Lemma 4.2. If $r_{ij} = 0 \; \forall i \in \mathcal{I}, j \in \mathcal{E}$, then $i \not\to j$ whenever $i \in \mathcal{I}, j \in \mathcal{E}$. We know from Lemma 4.3 that this cannot be. This establishes the first statement.

The proof of the second statement is based on Lemmas 4.1 and 4.3. Note that, since $\Pi^{-1} A \Pi$ is block triangular, the diagonal blocks of $\Pi^{-1} A^m \Pi$ are indeed P^m and Q^m respectively. However, the off-diagonal block is a complicated function of P and Q, so we denote it by $R^{(m)}$. Let $i \in \mathcal{I}$ be arbitrary. Then from Lemma 4.3, there exists a $j \in \mathcal{E}$ such that $i \to j$. It can be assumed without loss of generality that the path from i to j does not contain any other essential nodes. If the path from i to j does indeed pass through another $j' \in \mathcal{E}$, then we can replace j by j'. So suppose $i \to j$, and that the path from i to j does not pass through any other essential node; thus the path must pass through only inessential nodes. Since there are only m inessential nodes, this implies that the path from i to j must have length no larger than m. Denote this length by $l(i,j)$. Next, since $j \in \mathcal{E}$, there exists a $j' \in \mathcal{E}$ such that $a_{jj'} > 0$. This is because, if $a_{jj'} = 0$ for all j', then clearly the node j does not lead to any other node, and by definition j would be inessential. Thus there exists a j' such that $a_{jj'} > 0$. Now, it is obvious that $j \to j'$, and we already know from Lemma 4.1 that an essential node can lead only to another essential node. Thus there exists a $j' \in \mathcal{E}$ such that $a_{jj'} > 0$. Concatenating the path of length $l(i,j)$ from i to j with the path of length one from j to j', we get a path of length $l(i,j)+1$ from $i \in \mathcal{I}$ to $j' \in \mathcal{E}$. Repeating this process $m - l(i,j)$ times gives a path from $i \in \mathcal{I}$ to some $k \in \mathcal{E}$ that has length exactly equal to m. Thus, in the matrix $R^{(m)}$, the ik-th entry is positive. Since i is arbitrary, this establishes the second statement. □

4.1.2 Irreducible Matrices

In order to proceed beyond the basic canonical form (4.1), we introduce the notion of irreducibility.

Definition 4.5 *A matrix $A \in \mathbb{R}_+^{n \times n}$ is said to be* **reducible** *if there exists a partition of $\mathcal{N}$ into disjoint nonempty sets $\mathcal{I}$ and $\mathcal{J}$ such that, through a symmetric permutation of rows and columns, A can be brought into the form*

$$\Pi^{-1} A \Pi = \begin{array}{c c} & \begin{array}{cc} \mathcal{I} & \mathcal{J} \end{array} \\ \begin{array}{c} \mathcal{I} \\ \mathcal{J} \end{array} & \left[\begin{array}{cc} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{array} \right] \end{array}, \tag{4.3}$$

where Π is a permutation matrix. If this is not possible, then A is said to be **irreducible***.*

From the above definition, we can give an equivalent charactertization of irreducibility: $A \in \mathbb{R}_+^{n \times n}$ is irreducible if and only if, for every partition of $\mathcal{N}$ into disjoint nonempty sets $\mathcal{I}$ and $\mathcal{J}$, there exist $i \in \mathcal{I}$ and $j \in \mathcal{J}$ such that $a_{ij} > 0$.

The next result brings out the importance of irreducibility.

Theorem 4.6 *For a given $A \in \mathbb{R}_+^{n \times n}$, the following statements are equivalent:*

(i) A is irreducible.

(ii) Every $i \in \mathcal{N}$ leads to every other $j \in \mathcal{N}$.

Proof. **(ii) $\Rightarrow$ (i).** Actually we prove the contrapositive, namely: If (i) is false, then (ii) is false. Accordingly, suppose (i) is false and that A is reducible. Put A in the form (4.3). Then it is easy to see that

$$\Pi^{-1} A^l \Pi = \begin{array}{ccc} & \mathcal{I} & \mathcal{J} \\ \mathcal{I} & A^l_{11} & \mathbf{0} \\ \mathcal{J} & A^l_{21} & A^l_{22} \end{array} .$$

Hence $a_{ij}^{(l)} = 0 \; \forall i \in \mathcal{I}, j \in \mathcal{J}, \; \forall l$. In turn this implies that $i \not\to j$ whenever $i \in \mathcal{I}, j \in \mathcal{J}$. Hence (ii) is false.

(i) $\Rightarrow$ (ii). Recall the alternate characterization of irreducibility given after Definition 4.5. Choose $i \in \mathcal{N}$ arbitrarily. It is shown that i leads to every $j \in \mathcal{N}$. To begin the argument, let $l = 1$, $i_1 = i$, and consider the partition $\mathcal{I}_1 = \{i_1\}$, $\mathcal{J}_1 = \mathcal{N} \setminus \mathcal{I}_1$. Since A is irreducible, there exists an $i_2 \in \mathcal{J}_1$ such that $a_{i_1 i_2} > 0$. So $i_1 \to i_2$. Next, let $l = 2$, and consider the partition $\mathcal{I}_2 = \{i_1, i_2\}$ and $\mathcal{J}_2 = \mathcal{N} \setminus \mathcal{I}_2$. Since A is irreducible, there exists an $i_3 \in \mathcal{J}_2$ such that $a_{i_s i_3} > 0$ for either $s = 1$ or $s = 2$. If $s = 1$, then $a_{i_1 i_3} > 0$ implies that $i_1 \to i_3$. If $s = 2$, then $i_2 \to i_3$. Since $i_1 \to i_2$, the transitivity of $\to$ implies that $i_1 \to i_3$. In either case we can conclude that $i_1 \to i_3$. Repeat the argument. At step l we would have identified l *distinct* nodes $i_1, \ldots, i_l$ such that $i_1 \to i_s$ for $s = 2, \ldots, l$. Now consider the partition $\mathcal{I}_l = \{i_1, \ldots, i_l\}$, $\mathcal{J}_l = \mathcal{N} \setminus \mathcal{I}_l$. Since A is irreducible, there exists an $i_{l+1} \in \mathcal{J}_l$ such that $a_{i_s i_{l+1}} > 0$ for some $s = 1, \ldots, l$. Since $i_1 \to i_s$ for $s = 2, \ldots, l$, this implies that $i_1 \to i_{l+1}$. Notice that at each step we pick up yet another *distinct* node. Hence, after $l = n - 1$ steps, we conclude that $i \to j$ for all $j \neq i$, which is (i). □

Corollary 4.7 *For a given $A \in R_+^{n \times n}$, the following statements are equivalent:*

(i) A is irreducible.

(ii) For each $i, j \in \mathcal{N}$, there exists an integer l such that $a_{ij}^{(l)} > 0$.

(iii) Define

$$B := \sum_{l=1}^{n-1} A^l .$$

Then $b_{ij} > 0 \; \forall i, j \in \mathcal{N}, i \neq j$. In other words, all off-diagonal elements of B are positive

Proof. **(i) $\Rightarrow$ (ii).** The only difference between the present Statement (ii) and Statement (ii) of Theorem 4.6 is that there is no restriction here that $i \neq j$. Suppose A is irreducible. Then, as shown in Theorem 4.6, for every $i, j \in \mathcal{N}, i \neq j$, we have that $i \to j$. By the same logic, $j \to i$ also. Thus $i \to i \; \forall i \in \mathcal{N}$.

(ii) $\Rightarrow$ (i). Suppose (ii) is true. Then in particular $i \to j$ whenever $i, j \in \mathcal{N}, i \neq j$. So from Theorem 4.6, A is irreducible.

(iii) $\Leftrightarrow$ (ii). Statement (iii) is equivalent to: For each $i, j \in \mathcal{N}, i \neq j$, there exists an $l \leq n-1$ such that $a_{ij}^{(l)} > 0$. This is the same as: For each $i, j \in \mathcal{N}, i \neq j$, there exists a path *of length* $\leq n-1$ from i to j. Clearly (iii) implies (ii). To see that (ii) implies (iii), observe that if there is a path of any length l from i to $j \neq i$, and if $l \geq n$, then the path must include a cycle (a path starting and ending at the same node). This cycle can be removed from the path without affecting the reachability of j from i. Hence it can be assumed that $l \leq n-1$, which is (iii). □

4.1.3 Final Version of Canonical Form

With the aid of Theorem 4.6, we can refine the basic canonical form introduced in Lemma 4.1.

We begin by reprising some definitions from Subsection 4.1.1. Given a matrix $A \in \mathbb{R}_+^{n \times n}$, we divided the nodes in $\mathcal{N} = \{1, \ldots, n\}$ into two disjoint sets: The set $\mathcal{E}$ of essential nodes, and the set $\mathcal{I}$ of inessential nodes. We also showed that $a_{ij} = 0$ whenever $i \in \mathcal{E}$ and $j \in \mathcal{I}$, leading to the canonical form (4.1). The objective of the present subsection is to refine further the structure of the canonical form.

For this purpose, observe that the binary relation $\to$ ("leads to") is an *equivalence relation* on the set $\mathcal{E}$. In order to be an equivalence relation, $\to$ must satisfy the following three conditions:

(a) $i \to i$ for all $i \in \mathcal{E}$. (Reflexivity)

(b) $i \to j$ implies that $j \to i$. (Symmetry)

(c) $i \to j, j \to k$ together imply that $i \to k$.

Let us now verify each of these conditions in succession.

Suppose $i \in \mathcal{E}$. Recall the convention that if i does not lead to any node, then i is taken to be inessential. Hence $i \in \mathcal{E}$ implies that there exists some $j \in \mathcal{N}$ (which could be i) such that $i \to j$. By the definition of an essential node, this in turn implies that $j \to i$. Now the transitivity of $\to$ implies that $i \to i$. Hence $\to$ is reflexive. Next, suppose $i \to j \neq i$. Since $i \in \mathcal{E}$, this implies that $j \to i$. Hence $\to$ is symmetric. Finally, the transitivity of $\to$ has already been established in Subsection 4.1.1. Therefore $\to$ is an equivalence relation on $\mathcal{E}$.

Note that $\to$ need not be an equivalence relation on $\mathcal{N}$, the set of *all* nodes. In particular, only transitivity is guaranteed, and $\to$ need not be either reflexive or symmetric. If we look at the matrix A of Example 4.1, we see that node 1 is not reachable from itself. Moreover, node 2 can be reached from node 1, but node 1 cannot be reached from node 2.

Since $\rightarrow$ is an equivalence relation on $\mathcal{E}$, we can partition $\mathcal{E}$ into its disjoint equivalence classes under $\rightarrow$. Let s denote the number of these equivalence classes, and let $\mathcal{E}_1, \ldots, \mathcal{E}_s$ denote the equivalence classes. Hence if $i, j \in \mathcal{E}$ belong to disjoint equivalence classes, then it follows that $i \not\rightarrow j$ and $j \not\rightarrow i$. In particular, $a_{ij} = 0$ whenever i, j belong to disjoint equivalence classes. Now let us permute the elements of $\mathcal{E}$ in such a way that all elements of $\mathcal{E}_1$ come first, followed by those of $\mathcal{E}_2$, and ending finally with the elements of $\mathcal{E}_s$. Note that the ordering of the equivalence classes themselves, as well as the ordering of the elements within a particular equivalence class, can both be arbitrary. With this permutation, the matrix P in (4.1) looks like

$$\Pi_P^{-1} P \Pi_P = \begin{array}{cccc} & \mathcal{E}_1 & \ldots & \mathcal{E}_s \\ \mathcal{E}_1 & P_1 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}_s & \mathbf{0} & \ldots & P_s \end{array} .$$

Moreover, each matrix P_l has the property that every node in $\mathcal{E}_l$ leads to every other node in $\mathcal{E}_l$. Hence, by Theorem 4.6, it follows that each matrix P_l is *irreducible.*

We can also refine the structure of the matrix Q in (4.1). Unfortunately, the results are not so elegant as with P. Let us begin with the set $\mathcal{I}$ of inessential nodes, and partition it into $\mathcal{I}_1, \ldots, \mathcal{I}_r$ such that each $\mathcal{I}_l$ is a **communicating class**, that is, each $\mathcal{I}_l$ has the property that if $i, j \in \mathcal{I}_l$ then $i \rightarrow j$ and also $j \rightarrow i$. From the definition, it is clear that if $i, j \in \mathcal{I}$ belong to disjoint communicating classes, then either $i \rightarrow j$ or $j \rightarrow i$, but not both (or perhaps neither). Hence the communicating classes can be numbered in such a way that if $i \in \mathcal{I}_{r_1}$ and $j \in \mathcal{I}_{r_2}$ and $r_1 < r_2$, then $i \not\rightarrow j$. Note that the ordering of these communicating classes need not be unique. With this renumbering, the matrix Q in (4.1) becomes *block-triangular.*

These observations can be captured in the following:

Theorem 4.8 *Given a matrix $A \in \mathbb{R}_+^{n \times n}$, identify the set $\mathcal{E}$ of essential nodes and the set $\mathcal{I}$ of inessential nodes. Let s denote the number of equivalence classes of $\mathcal{E}$ under $\rightarrow$, and let r denote the number of distinct communicating classes of $\mathcal{I}$ under $\rightarrow$. Then there exists a permutation matrix Π over $\mathcal{N}$ such that*

$$\Pi^{-1} A \Pi = \begin{array}{ccc} & \mathcal{E} & \mathcal{I} \\ \mathcal{E} & P & \mathbf{0} \\ \mathcal{I} & R & Q \end{array} . \tag{4.4}$$

Moreover, P and Q have the following special forms:

$$P = \begin{array}{cccc} & \mathcal{E}_1 & \ldots & \mathcal{E}_s \\ \mathcal{E}_1 & P_1 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}_s & \mathbf{0} & \ldots & P_s \end{array} , \tag{4.5}$$

$$Q = \begin{array}{c} \\ \mathcal{I}_1 \\ \mathcal{I}_2 \\ \vdots \\ \mathcal{I}_r \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{I}_1 & \mathcal{I}_2 & \ldots & \mathcal{I}_r \end{array} \\ \left[\begin{array}{cccc} Q_{11} & \mathbf{0} & \ldots & \mathbf{0} \\ Q_{21} & Q_{22} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{r1} & Q_{r2} & \ldots & Q_{rr} \end{array} \right] \end{array} . \tag{4.6}$$

If every row of A contains a positive element, then the set $\mathcal{E}$ is nonempty. Moreover, if the set $\mathcal{I}$ is also nonempty, then every row of the matrix R contains at least one positive element.

Example 4.3 Let us return to the matrix A of Example 4.1. We have already seen that the essential nodes are $\mathcal{E} = \{4, \ldots, 10\}$ and the inessential nodes are $\mathcal{I} = \{1, 2, 3\}$. To complete the canonical form, we need to do two things. First, we need to identify the equivalence classes of $\mathcal{E}$ under $\rightarrow$. Second, we need to identify the communicating classes of $\mathcal{I}$ under $\rightarrow$. From the reachability matrix, we can see that there are three equivalence classes of $\mathcal{E}$, namely $\mathcal{E}_1 = \{4, 6, 8\}$, $\mathcal{E}_2 = \{5, 7\}$, and $\mathcal{E}_3 = \{9, 10\}$. Note that the ordering of these sets is arbitrary. Next, the reachability matrix also shows that there are two communicating classes in $\mathcal{I}$, namely $\mathcal{I}_1 = \{2, 3\}$ and $\mathcal{I}_2 = \{1\}$. Here the ordering is *not* arbitrary if we wish Q to have a block-triangular structure. Hence the matrix A can be permuted into the following canonical form:

$$\Pi^{-1} A \Pi = \begin{array}{c|ccc|cc|cc|cc|c|}
 & 4 & 6 & 8 & 5 & 7 & 9 & 10 & 2 & 3 & 1 \\ \hline
4 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
8 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
5 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
9 & 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 \\
10 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ \hline
2 & 0 & 0 & 0 & 0 & \times & \times & 0 & 0 & \times & 0 \\
3 & \times & \times & 0 & 0 & \times & 0 & 0 & \times & 0 & 0 \\ \hline
1 & 0 & 0 & \times & \times & 0 & 0 & 0 & \times & 0 & 0
\end{array}$$

4.1.4 Irreducibility, Aperiodicity and Primitivity

In this subsection we go beyond studying merely irreducible matrices, and introduce two more notions, namely: the period of an irreducible matrix, and primitive matrices. An important tool in this theory is the greatest common divisor (g.c.d.) of a set of integers. The g.c.d. of a *finite* set of integers is the stuff of high school arithmetic, but here we study the g.c.d. of an *infinite* set of integers.

In the sequel, $\mathbb{Z}$ denotes the set of integers, and $\mathbb{Z}_+$ the set of nonnegative integers. Throughout we use the notation "$a|b$" to denote the fact that $a \in \mathbb{Z}$

divides $b \in \mathbb{Z}$, i.e., there exists another integer $c \in \mathbb{Z}$ such that $b = ac$. Note that $a|b$ if and only if the nonnegative integer $|a|$ divides $|b|$.

Suppose $b_1, \ldots, b_m \in \mathbb{Z}$. Then their **greatest common divisor (g.c.d.)** is an integer a that satisfies two conditions:

(i) $a|b_i$ for all i.

(ii) If $c|b_i$ for all i, then $c|a$.

The g.c.d. is unique except for its sign. Thus if a is a g.c.d. of $\{b_1, \ldots, b_m\}$, then $-a$ is the only other g.c.d. By convention, here we always take the g.c.d. to be a positive number.

The following elementary fact is well-known. Suppose that $b_1, \ldots, b_m \in \mathbb{Z}_+$ and let $a \in \mathbb{Z}_+$ denote their g.c.d. Then there exist $c_1, \ldots, c_m \in \mathbb{Z}$ such that

$$a = \sum_{i=1}^{m} c_i b_i.$$

Note that the c_i need not be nonnegative; they will just be integers, and in general some of them will be negative. Since each b_i is a positive multiple of a, an easy consequence of this is that every *sufficiently large* multiple of a can be expressed as a *nonnegative* "linear combination" of the b_i's. More precisely, there exists an integer l_0 such that, whenever $l \geq l_0$, we can find *nonnegative* constants $c_1, \ldots, c_m \in \mathbb{Z}_+$ such that

$$la = \sum_{i=1}^{m} c_i b_i.$$

Just how large is "large enough"? It is shown in [42] that $l_0 \leq b_0^2$, where $b_0 := \max_i b_i$.

All of this is fine for a *finite* set of integers. What about an *infinite* set of integers? Suppose $1 \leq b_1 < b_2 < \ldots$ is a countable collection of integers. We can define their g.c.d. in terms of the same two conditions (i) and (ii) above. Notice that the definition of a g.c.d. is unambiguous even when there are infinitely many integers. Define a_i to be the g.c.d. of $\{b_1, \ldots, b_i\}$. Then $1 \leq a_{i+1} \leq a_i$. So the sequence $\{a_i\}$ is bounded below and therefore converges to some value, call it a. Moreover, since each a_i is an *integer*, there exists an integer i_0 such that $a_i = a \; \forall i \geq i_0$. These observations can be summarized as follows: Every set of integers $\{b_1, b_2, \ldots\}$, ordered such that $1 \leq b_1 < b_2 < \ldots$, has a g.c.d., call it a. Moreover, there exists an integer i_0 such that a is the g.c.d. of $\{b_1, \ldots, b_{i_0}\}$. Every sufficiently large multiple of a can be represented as a nonnegative linear combination of $b_1, \ldots, b_{i_0}$.

The next important concept is that of the period of a primitive matrix, given in Definition 4.13. We lead up to this important definition in stages.

Lemma 4.9 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible. Then for each $i \in \mathcal{N}$, there exists an integer $r > 0$ such that $a_{ii}^{(r)} > 0$.*

Proof. Given $i \in \mathcal{N}$, choose an arbitrary $j \neq i$. Since A is irreducible, Theorem 4.6 implies that $i \to j$ and $j \to i$ Now the transitivity of $\to$ implies that $i \to i$. Hence $a_{ii}^{(r)} > 0$ for some integer $r > 0$. □

Definition 4.10 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible, and let $i \in \mathcal{N}$. The* **period** *of i is defined as the g.c.d. of all integers r such that $a_{ii}^{(r)} > 0$.*

Example 4.4 Suppose A has the form

$$A = \begin{bmatrix} 0 & \times & 0 \\ \times & 0 & \times \\ \times & 0 & 0 \end{bmatrix}.$$

Then there are two cycles from node 1 to itself, namely: $1 \to 2 \to 1$ (length $= 2$) and $1 \to 2 \to 3 \to 1$ (length $= 3$). Hence $a_{ii}^{(2)} > 0$ and $a_{ii}^{(3)} > 0$. Since the g.c.d. of 2 and 3 is one, the period of node 1 is 1.

Example 4.5 Suppose A has the form

$$A = \begin{bmatrix} 0 & \times & 0 & 0 \\ \times & 0 & 0 & \times \\ \times & 0 & 0 & 0 \\ 0 & 0 & \times & 0 \end{bmatrix}.$$

Then there are cycles from node 1 to itself of length 2 ($1 \to 2 \to 1$) and length 4 ($1 \to 2 \to 4 \to 3 \to 1$). All cycles from node 1 to itself are concatenations of these two cycles. Hence every cycle from node 1 to itself has even length. Therefore the period of node 1 is 2.

Now consider node 4. There is no cycle of length 2 from node 4 to itself, but there are cycles of length 4 ($4 \to 3 \to 1 \to 2 \to 4$) and length 6 ($4 \to 3 \to 1 \to 2 \to 1 \to 2 \to 4$). The g.c.d. of 4 and 6 is also 2, so the period of node 4 is also 2.

We shall see shortly that this is not a coincidence, but is a general property of irreducible matrices. Incidentally, this example also shows the rationale behind permitting the cycles to have lengths larger than the size of the matrix.

Theorem 4.11 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible. Then every node in $\mathcal{N}$ has the same period.*

Proof. For a node i, define

$$S_i := \{r : a_{ii}^{(r)} > 0\}.$$

In words, S_i consists of the lengths of all cycles from node i to itself. Let p_i denote the period of node i. Then by definition p_i is the g.c.d. of all integers in the set S_i. In particular, $p_i | r \; \forall r \in S_i$.

Choose any $j \neq i$. Since A is irreducible, we have from Corollary 4.7 that $i \to j$ and $j \to i$. So there exist integers l and m such that $a_{ij}^{(l)} > 0$ and $a_{ji}^{(m)} > 0$. As a consequence,

$$a_{ii}^{(l+m)} = \sum_{s=1}^{n} a_{is}^{(l)} a_{si}^{(m)} \geq a_{ij}^{(l)} a_{ji}^{(m)} > 0.$$

So $l + m \in S_i$ and as a result $p_i | (l + m)$.

Now let $r \in S_j$ be arbitrary. Then by definition $a_{jj}^{(r)} > 0$. Therefore

$$a_{ii}^{(l+m+r)} = \sum_{s=1}^{n} \sum_{t=1}^{n} a_{is}^{(l)} a_{st}^{(r)} a_{ti}^{(m)} \geq a_{ij}^{(l)} a_{jj}^{(r)} a_{ji}^{(m)} > 0.$$

So $l + m + r \in S_i$, and $p_i | (l + m + r)$. Since it has already been shown that $p_i | (l + m)$, we conclude that $p_i | r$. But since r is an *arbitrary* element of S_j, it follows that p_i divides *every* element of S_j, and therefore $p_i | p_j$ (where p_j is the period of node j).

However, i and j are arbitrary nodes, so their roles can be interchanged throughout, leading to the conclusion that $p_j | p_i$. This shows that $p_i = p_j$ for all i, j. □

Theorem 4.11 shows that we can speak of "the period of an irreducible matrix" without specifying which node we are speaking about.

Definition 4.12 *An irreducible matrix is said to be* **aperiodic** *if its period is one.*

Definition 4.13 *A matrix $A \in \mathbb{R}_+^{n \times n}$ is said to be* **primitive** *if there exists an integer m such that $A^m > \mathbf{0}$.*

The next theorem is one of the key results in the theory of nonnegative matrices.

Theorem 4.14 *Given a matrix $A \in \mathbb{R}_+^{n \times n}$, the following statements are equivalent:*

(i) A is irreducible and aperiodic.

(ii) A is primitive.

Proof. We begin with a simple observation. Consider two statements:

1. There exists an integer m such that $A^m > \mathbf{0}$ (that is, A is primitive as defined in Definition 4.13).

2. There exists an integer l_0 such that $A^l > \mathbf{0} \; \forall l \geq l_0$.

We claim that both statements are equivalent. It is clear that Statement 2 implies Statement 1; just take $m = l_0$. To prove the implication in the opposite direction, suppose Statement 1 is true. Then clearly no row of A

can be identically zero. (If a row of A is identically zero, then the same row of A^m would continue to be identically zero for all values of m.) Suppose $a_{ij}^{(m)} > 0 \; \forall i, j \in \mathcal{N}$. Then

$$a_{ij}^{(m+1)} = \sum_{s=1}^{n} a_{is} a_{sj}^{(m)} > 0 \; \forall i, j,$$

since $a_{is} > 0$ for at least one value of s for each given i, and $a_{sj}^{(m)} > 0$ for every s, j. This shows that $A^{m+1} > \mathbf{0}$. Now repeat with m replaced by $m + 1$, and use induction.

(ii) $\Rightarrow$ (i). Suppose A is primitive. Then A has to be irreducible. If A has the form

$$\Pi^{-1} A \Pi = \begin{bmatrix} \times & \mathbf{0} \\ \times & \times \end{bmatrix}$$

after symmetric permutation of rows and columns, then A^m has the form

$$\Pi^{-1} A^m \Pi = \begin{bmatrix} \times & \mathbf{0} \\ \times & \times \end{bmatrix}$$

for every value of m. So a reducible matrix cannot be primitive. Second, suppose A is irreducible but has period $p > 1$. Then, by the definition of the period, it follows that $a_{ii}^{(l)} = 0$ whenever l is not a multiple of p. Hence Statement 2 above is false and A cannot be primitive.

(i) $\Rightarrow$ (ii). Suppose A is irreducible and aperiodic. Fix a value $i \in \mathcal{N}$. Let $d_1, d_2, \ldots$ denote the lengths of all the cycles from i to itself. By the definition of the period, the g.c.d. of all these lengths is one. Hence the g.c.d. of a finite subset of these lengths is also one. Let $d_1, \ldots, d_s$ denote the finite subset of the cycle lengths whose g.c.d. is one. Then, as discussed at the beginning of this subsection, *every* sufficiently large integer r can be expressed as a nonnegative linear combination of the form $r = \sum_{t=1}^{s} \mu_t d_t$. Clearly, if there are cycles of lengths $d_1, \ldots, d_s$ from node i to itself, then there is a cycle of length $\sum_{t=1}^{s} \mu_t d_t$ for every set of nonnegative integers $\mu_1, \ldots, \mu_s$. (Follow the cycle of length d_t μ_t times, and do this for $t = 1, \ldots, s$.) So the conclusion is that there exists an integer r_i such that, for every integer $r \geq r_i$, there is a cycle of length r from node i to itself. Note that the smallest such integer r_i may depend on i. Now define $r^* := \max\{r_1, \ldots, r_n\}$. Then by the manner in which in which r^* has been chosen, it follows that

$$a_{ii}^{(r)} > 0, \; \forall i \in \mathcal{N}, \; \forall r \geq r^*.$$

Now it is shown that $A^{r^*+n-1} > \mathbf{0}$; this is enough to show that A is primitive. To show that $A^{r^*+n-1} > \mathbf{0}$, we must show that for each $i, j \in \mathcal{N}$, there exists a path of length *exactly equal* to $r^* + n$ from node i to node j. It can be assumed that $i \neq j$, since $a_{ii}^{(r^*+n-1)}$ is already known to be positive for all i. Since A is irreducible, we have from Theorem 4.6 that every i leads to every other j. Choose a path of length $\mu(i, j)$ from node i to node j. Without loss of generality, it can be assumed that $\mu(i,j) \leq n-1$. Now by the characterization of the integer r^*, there exists a path of length $r^*+n-1-\mu(i,j)$ from node i to itself. If this is concatenated with a path of length $\mu(i,j)$ from node i to node j, we will get a path of length r^*+n-1 from node i to node j. Hence $A^{r^*+n-1} > \mathbf{0}$ and A is primitive. □

Theorem 4.14 shows that, if A is irreducible and aperiodic, then $A^l > \mathbf{0}$ for all sufficiently large l. Now we answer the question of how large l needs to be.

Theorem 4.15 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible and aperiodic. Define*

$$m_0(A) := \min\{m : A^m > \mathbf{0}\}.$$

Next, define

$$\mu(n) := \max_{A \in \mathbb{R}_+^{n \times n}} m_0(A),$$

where it is understood that the maximum is taken only over the set of irreducible and aperiodic matrices. Then

$$(n-2)(n-1) \leq \mu(n) \leq 3n^2 + n - 1. \tag{4.7}$$

Proof. As a preliminary first step, we repeat the trick from the proof of Theorem 4.14. Define

$$l_0(A) := \min\{l^* : A^l > \mathbf{0}\ \forall l \geq l^*\}.$$

Then it is claimed that $l_0(A) = m_0(A)$. It is obvious that $m_0(A) \leq l_0(A)$, since by definition $A^{l_0(A)} > \mathbf{0}$. To prove the converse, observe as before that if A is irreducible, then no row of A can be identically zero. Let $m = m_0(A)$. Then by assumption $a_{ij}^{(m)} > 0\ \forall i,j$. So

$$a_{ij}^{(m+1)} = \sum_{s=1}^{n} a_{is} a_{sj}^{(m)} > 0\ \forall i,j,$$

since $a_{is} > 0$ for some s and $a_{sj}^{(m)} > 0$ for every s. Repeating this argument with m replaced by $m+1, m+2$ etc. shows that $A^l > \mathbf{0}\ \forall l \geq m_0(A)$. Hence $l_0(A) \leq m_0(A)$.

First we establish the lower bound for $\mu(n)$. Note that the bound is trivial if $n = 1$ or 2, since the left side equals zero in either case. So suppose $n \geq 3$, and define $A \in \mathbb{R}_+^{n \times n}$ to be a cyclic permutation matrix on $\mathcal{N} = \{1, \ldots, n\}$ with one extra positive element, as follows:

$$a_{1,2} = 1, a_{2,3} = 1, \ldots, a_{i,i+1} = 1 \text{ for } i = 1, \ldots, n-1, a_{n,1} = 1,$$

and in addition $a_{n-1,1} = 1$. All other elements of A are zero. To illustrate, if $n = 5$, then

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now there are two cycles of length $n-1$ and n respectively from node 1 to itself; these are $1 \to 2 \to \cdots \to n{-}1 \to 1$ and $1 \to 2 \to \cdots \to n{-}1 \to n \to 1$. There are no other cycles of length $\leq n-1$. Hence all cycles from node 1 to itself are have lengths of the form $\alpha(n-1)+\beta n$ for nonnegative integers α and β. It is easy to verify that $(n-2)(n-1)$ is the smallest integer l_0 with the property that *every* integer $l \geq l_0$ can be expressed in the form $\alpha(n{-}1){+}\beta n$ for nonnegative integers α and β. Hence $l_0(A) = (n{-}2)(n{-}1)$, whence $\mu_0(A) \geq (n-2)(n-1)$.

The proof of the upper bound requires a result that is not proven here; see [1, 2], Lemma 2.3. It follows from this lemma that if A is irreducible and aperiodic, then for *every* integer $s \geq 3n^2$, there exits a cycle of length s from node i to itself for every i. Now we can repeat the argument in the proof of Theorem 4.14 to show that for every integer $l \geq 3n^2+n-1$, there exists a path of length l from node i to node j, for every i, j. Hence $l_0(A) \leq 3n^2+n-1$, for every irreducible and aperiodic A. □

4.1.5 Canonical Form for Periodic Irreducible Matrices

Up to now we have been studying *aperiodic* irreducible matrices. In this section, the focus is on irreducible matrices whose period is greater than one.

Lemma 4.16 *Suppose* $A \in \mathbb{R}_+^{n\times n}$ *is irreducible and has period* $p \geq 2$. *Then for each* $i, j \in \mathcal{N} = \{1, \ldots, n\}$, *there exists a unique integer* $r(i,j) \in \{0, \ldots, p-1\}$ *such that the length of every path from node* i *to node* j *equals* $r(i,j) \mod p$.

Proof. Consider two distinct paths from node i to node j, of lengths l_1 and l_2 respectively. The desired conclusion is that $l_1 = l_2 \mod p$, or equivalently, that $p|(l_1-l_2)$. Choose some path of length m from node j to node i. Such a path exists because A is irreducible. By concatenating this path with the two paths from node i to node j, we get two cycles from node i to itself, of lengths l_1+m and l_2+m respectively. Now, since p is the period of A, it divides the length of each cycle. Thus $p|(l_1+m)$ and $p|(l_2+m)$. Hence p divides the difference, i.e., $p|[(l_1+m)-(l_2+m)]$, or $p|(l_1-l_2)$. □

Lemma 4.17 *Suppose* $A \in \mathbb{R}_+^{n\times n}$ *is irreducible and has period* $p \geq 2$. *Define the integer* $r(\cdot,\cdot)$ *as in Lemma 4.16. Then for all* $i, j, k \in \mathcal{N}$, *we have that*

$$r(i,k) = [r(i,j)+r(j,k)] \mod p.$$

Proof. Choose a path of length l from node i to node j, and of length m from node j to node k. By concatenating the two paths, we get a path of length $l+m$ from node i to node k. Now Lemma 4.16 implies that $r(i,k) = (l+m) \mod p$, irrespective of how the paths are chosen. But clearly

$$\begin{aligned} r(i,k) &= (l+m) \mod p = [l \mod p + m \mod p] \mod p \\ &= [r(i,j)+r(j,k)] \mod p. \end{aligned}$$

This is the desired conclusion. □

The above two lemmas suggest a way of partitioning the nodes in $\mathcal{N}$. Choose an arbitrary node $i \in \mathcal{N}$. Define

$$\mathcal{C}_s := \{j \in \mathcal{N} : r(i,j) = s\}, s = 1, \ldots, p-1.$$

THus $\mathcal{C}_s$ consists of all nodes reachable from node i by paths of length $s \mod p$. Finally, define

$$\mathcal{C}_0 := \mathcal{N} \setminus \left[\cup_{s=1}^{p-1} \mathcal{C}_s \right].$$

Clearly $i \in \mathcal{C}_0$, since every cycle from node i to itself has length $0 \mod p$, and as a result $i \not\in \mathcal{C}_s$ for $s = 1, \ldots, p-1$. But $\mathcal{C}_0$ could contain other nodes as well. (See Example 4.6 below.) In fact, $\mathcal{C}_0$ consists of all nodes that are reachable from node i by paths whose lengths are multiples of p. Now the following observation is a ready consequence of Lemma 4.17.

Lemma 4.18 *Partition $\mathcal{N}$ into disjoint sets $\mathcal{C}_0, \ldots, \mathcal{C}_{p-1}$ as above. Then, for $j_1 \in \mathcal{C}_{s_1}$ and $j_2 \in \mathcal{C}_{s_2}$, we have*

$$r(j_1, j_2) = (s_2 - s_1) \mod p.$$

Proof. Note that

$$j_l \in \mathcal{C}_{s_l} \Rightarrow r(i, s_l) = s_l \mod p, \text{ for } l = 1, 2.$$

Hence, from Lemma 4.17,

$$r(j_1, j_2) = [r(i, j_2) - r(i, j_1)] \mod p = (s_2 - s_1) \mod p.$$

This is the desired conclusion. □

Now we are in a position to state the main result of this section.

Theorem 4.19 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible and has period $p \geq 2$. Partition $\mathcal{N}$ into disjoint sets $\mathcal{C}_0, \ldots, \mathcal{C}_{p-1}$ as above. Then, by a symmetric permutation of rows and columns, A can be put in the form*

$$\Pi^{-1} A \Pi = \begin{array}{c} \\ \mathcal{C}_0 \\ \mathcal{C}_1 \\ \mathcal{C}_2 \\ \vdots \\ \mathcal{C}_{p-2} \\ \mathcal{C}_{p-1} \end{array} \begin{array}{c} \begin{array}{cccccc} \mathcal{C}_0 & \mathcal{C}_1 & \mathcal{C}_2 & \ldots & \mathcal{C}_{p-2} & \mathcal{C}_{p-1} \end{array} \\ \left[\begin{array}{cccccc} \mathbf{0} & A_{01} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{12} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & A_{p-2,p-1} \\ A_{p-1,0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \end{array} \right] \end{array} =: B, \tag{4.8}$$

where the matrix $A_{s,s+1}$ has dimension $|\mathcal{C}_s| \times |\mathcal{C}_{s+1}|$. This canonical form is unique to within (i) a cyclic permutation of the classes $\mathcal{C}_0, \ldots, \mathcal{C}_{p-1}$, and (ii) an arbitrary permutation of the indices within each class. Moreover, each of the p cyclic products

$$M_0 := A_{01} A_{12} \cdots A_{p-1,0},$$

$$M_1 := A_{12} \cdots A_{p-1,0} A_{01}, \ldots$$

$$M_{p-1} := A_{p-1,0} A_{01} \cdots A_{p-2,p-1} \tag{4.9}$$

is primitive.

Proof. It readily follows from Lemma 4.18 that if $i \in \mathcal{C}_s$ and $j \in \mathcal{C}_t$, then $a_{ij} = 0$ unless $t = (s+1) \mod p$. This is because if $a_{ij} > 0$, then there exists a path of length one from node i to node j, which implies that $r(i,j) = 1$, which in turn implies that $t = (s+1) \mod p$. This shows that the permuted matrix $\Pi^{-1} A \Pi$ has the block-cyclic form shown in (4.8).

It remains only to show that each of the matrices in (4.9) is primitive. Note that the matrix B defined in (4.8) has a nonzero matrix only in block $(i, i+1)$ for $i = 0, \ldots, p-1$. (Here and elsewhere in this proof, all indices exceeding $p-1$ should be replaced by their values $\mod p$.) Hence, for every integer $m \geq 1$, B^m has a nonzero matrix only in blocks $(i, i+m)$ for $i = 0, \ldots, p-1$. In particular, B^p is block-diagonal and equals

$$B^p = \begin{array}{c c} & \begin{array}{ccc} \mathcal{C}_0 & \ldots & \mathcal{C}_{p-1} \end{array} \\ \begin{array}{c} \mathcal{C}_0 \\ \vdots \\ \mathcal{C}_{p-1} \end{array} & \left[\begin{array}{ccc} M_0 & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & M_{p-1} \end{array} \right] \end{array} .$$

So for every integer l, B^{pl} equals

$$B^{pl} = (B^p)^l = \begin{array}{c c} & \begin{array}{ccc} \mathcal{C}_0 & \ldots & \mathcal{C}_{p-1} \end{array} \\ \begin{array}{c} \mathcal{C}_0 \\ \vdots \\ \mathcal{C}_{p-1} \end{array} & \left[\begin{array}{ccc} M_0^l & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & M_{p-1}^l \end{array} \right] \end{array} .$$

The assertion that each M_i is primitive is equivalent to the statement that each of the diagonal blocks of B^{pl} is a strictly positive matrix for all sufficiently large values of l. Hence the assertion that each M_i is primitive is equivalent to the following statement: There exists an integer l_0 such that, for all $l \geq l_0$, there exists a path of length lp from node i to node j whenever they both belong to the same class $\mathcal{C}_s$, for some $s \in \{0, \ldots, p-1\}$. Accordingly, suppose i, j belong to the same class $\mathcal{C}_s$. By the definition of these classes, it follows that there is a path from node i to node j whose length is a multiple of p; call it $m(i,j)p$. By the definition of the period, the g.c.d. of the lengths of all cycles from node i to itself is p. Hence there exists an integer $l_0(i)$ such that there exist cycles of length lp from node i to itself for all $l \geq l_0(i)$. Hence there exist paths of length lp from node i to node j for all $l \geq l_0(i) + m(i,j)$. Since there are only finitely many i, j, the quantity

$$l_0 := \max_{i,j} \{ l_0(i) + m(i,j) \}$$

is finite. Moreover, for all $l \geq l_0$, there exist paths of length lp from node i to node j whenever i, j belong to the same class $\mathcal{C}_s$, This shows that each matrix M_i is primitive. □

Example 4.6 Suppose A is a 9×9 matrix of the form

$$A = \begin{array}{cc} & \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} & \left[\begin{array}{ccccccccc} 0 & \times & 0 & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & \times & 0 & 0 & \times & 0 & \times & 0 \\ 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 & 0 \\ 0 & \times & 0 & 0 & 0 & 0 & \times & 0 & 0 \\ \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & 0 & 0 & 0 & \times & 0 \\ \times & 0 & 0 & 0 & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

It can be verified that A has period 3. The classes (starting from node 1) are:

$$\mathcal{C}_0 = \{1, 5\}, \mathcal{C}_1 = \{2, 4, 7, 9\}, \mathcal{C}_2 = \{3, 6, 8\}.$$

After permuting rows and columns accordingly, the matrix A can be put in the form

$$\Pi^{-1} A \Pi = \begin{array}{cc} & \begin{array}{ccccccccc} 1 & 5 & 2 & 4 & 7 & 9 & 3 & 6 & 8 \end{array} \\ \begin{array}{c} 1 \\ 5 \\ 2 \\ 4 \\ 7 \\ 9 \\ 3 \\ 6 \\ 8 \end{array} & \left[\begin{array}{cc|cccc|ccc} 0 & 0 & \times & \times & \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & \times & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & \times & 0 & 0 \\ \hline 0 & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

4.2 PERRON-FROBENIUS THEORY

In this section, we present various theorems about primitive and irreducible matrices. The first such theorems are due to Perron [87] and Frobenius [50, 51, 52]. Perron's original paper was for *positive* matrices, while Frobenius extended the theory to *nonnegative* matrices. The paper by Wielandt [117] was very influential in that most subsequent expositions of the theory follow his approach. But the theory continues to be known by the names of the two originators. A substantial generalization of the theory results when nonnegative matrices are replaced by matrices that leave a given "cone" invariant. For an exposition of this approach, see [17]. Much of the discussion in this section follows [99], Chapter 1.

Throughout this section, we write $\mathbf{x} \geq \mathbf{0}$ to indicate that every component of a (row or column) vector $\mathbf{x}$ is nonnegative, and $\mathbf{x} > \mathbf{0}$ to indicate that every component of $\mathbf{x}$ is positive. Similarly, $A \geq \mathbf{0}$ ($A > \mathbf{0}$) indicates that every component of the matrix A is nonnegative (positive). Expressions such as $\mathbf{x} \geq \mathbf{y}$ or $B < A$ are self-explanatory. Also, given an arbitrary row vector $x \in \mathbb{C}^n$, the symbol $\mathbf{x}_+$ denotes the vector $[|x_1| \ldots |x_n|] \in \mathbb{R}^n_+$. Note that $\mathbf{x}_+$ always belongs to $\mathbb{R}^n_+$. The notation A_+ is defined analogously. Finally, as in Chapter 2, the symbol $\mathbb{S}_n$ denotes the n-dimensional simplex, i.e.,

$$\mathbb{S}_n = \{\mathbf{x} \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}.$$

Given a matrix A, the **spectrum** of A consists of all eigenvalues of A and is denoted by spec(A). If spec$(A) = \{\lambda_1, \ldots, \lambda_n\}$, then

$$\rho(A) := \max\{|\lambda_i| : \lambda_i \in \text{spec}(A)\}$$

is called the **spectral radius** of A. Note that spec(A) can contain complex numbers, but $\rho(A)$ is always real and nonnegative.

4.2.1 Perron-Frobenius Theorem for Primitive Matrices

In this section, we state and prove the principal result for primitive matrices. In the next subsection, it shown that very similar results hold for irreducible matrices even if they are not primitive.

We begin with an upper bound for the spectral radius of an arbitray matrix.

Lemma 4.20 *Given a matrix $A \in \mathbb{R}^{n \times n}$, define*

$$r(A; \mathbf{x}) := \min_{j \in \mathcal{N}} \frac{(\mathbf{x}A_+)_j}{x_j} \; \forall \mathbf{x} \in \mathbb{S}_n,$$

and

$$r(A) := \max_{\mathbf{x} \in \mathbb{S}_n} r(A; \mathbf{x}). \tag{4.10}$$

Then $\rho(A) \leq r(A)$.

Proof. Note that in the definition of $r(A; \mathbf{x})$, we take the ratio $(\mathbf{x}A_+)_j / x_j$ to be ∞ if $x_j = 0$. However, since $\mathbf{x} \in \mathbb{S}_n$, clearly $x_j \neq 0$ for at least one index $j \in \mathcal{N}$. Hence $r(A; \mathbf{x})$ is finite for all $\mathbf{x} \in \mathbb{S}_n$. An equivalent and alternate definition of $r(A; \mathbf{x})$ is:

$$r(A; \mathbf{x}) := \max\{\lambda \in \mathbb{R}_+ : \mathbf{x}A_+ \geq \lambda \mathbf{x}\}.$$

This definition brings out clearly the fact that, for a fixed matrix A, the map $\mathbf{x} \mapsto r(A; \mathbf{x})$ is *upper semi-continuous.*[2] Suppose $\lambda_i \to \lambda_0, \mathbf{x}_i \to \mathbf{x}_0, A_+\mathbf{x}_i \geq \lambda_i \mathbf{x}_i$ where $\lambda_i, \lambda_0 \in \mathbb{R}_+$ and $\mathbf{x}_i, \mathbf{x}_0 \in \mathbb{S}_n$. Then it is obvious that $A_+\mathbf{x}_0 \geq \lambda_0 \mathbf{x}_0$, whence $r(A; \mathbf{x}_0) \geq \lambda_0$. In particular, if $\mathbf{x}_i \to \mathbf{x}_0$ and $r_0 = \limsup_{i \to \infty} r(A; \mathbf{x}_i)$, then $r(A; \mathbf{x}_0) \geq r_0$. Since $\mathbb{S}_n$ is a compact subset of $\mathbb{R}^n$, $r(A; \cdot)$ attains its maximum on $\mathbb{S}_n$.

[2] The reader is referred to [95] for various concepts and results from real analysis, such as compactness, semi-continuity etc.

To show that $\rho(A) \leq r(A)$, let $\lambda \in \mathbb{C}$ be an arbitrary eigenvalue of A, and let $\mathbf{z} \in \mathbb{C}^n$ be an associated eigenvector. Without loss of generality, we can scale $\mathbf{z}$ such that $\mathbf{z}_+ \in \mathbb{S}_n$. Now

$$\lambda z_j = \sum_{i=1}^{n} z_i a_{ij} \ \forall j \in \mathcal{N},$$

$$|\lambda||z_j| = \left| \sum_{i=1}^{n} z_i a_{ij} \right| \leq \sum_{i=1}^{n} |z_i||a_{ij}| \ \forall j \in \mathcal{N},$$

$$|\lambda| \leq \min_{j \in \mathcal{N}} \frac{(\mathbf{z}_+ A_+)_j}{(\mathbf{z}_+)_j} \leq r(A).$$

Hence $\rho(A) \leq r(A)$. □

Now we come to the Perron-Frobenius theorem for primitive matrices.

Theorem 4.21 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is primitive. Then*

(i) $\rho(A)$ is an eigenvalue of A.

(ii) There exists a row eigenvector $\mathbf{v} > \mathbf{0}$ of A corresponding to the eigenvalue $\rho(A)$.

(iii) If $B \in \mathbb{R}^{n \times n}$ satisfies $\mathbf{0} \leq B \leq A$, then $\rho(B) \leq \rho(A)$, with equality if and only if $B = A$.

(iv) Rank$[\rho(A)I - A] = n - 1$, so that the eigenvectors of A associated with the eigenvalue $\rho(A)$ are multiples of each other.

(v) $\rho(A)$ is a simple eigenvalue of A.

(vi) If λ is any other eigenvalue of A, then $|\lambda| < \rho(A)$.

Proof. Since $r(A; \cdot)$ is an upper semi-continuous function on $\mathbb{S}_n$ and $\mathbb{S}_n$ is a compact subset of $\mathbb{R}_+^n$, there exists a vector $\mathbf{v} \in \mathbb{S}_n$ such that $r(A; \mathbf{v}) = r(A)$. For brevity, let r stand for $r(A)$. Also, let l be an integer such that $A^l > \mathbf{0}$. Such an integer l exists because A is primitive.

(i). Since $r(A; \mathbf{v}) = r(A)$, it follows from the definition of $r(\cdot)$ that $\mathbf{v}A - r\mathbf{v} \geq \mathbf{0}$. (Note that $A_+ = A$ since A is a nonnegative matrix.) Let $\mathbf{z}$ denote $\mathbf{v}A - r\mathbf{v}$. If $\mathbf{z} = \mathbf{0}$ then $\mathbf{v}A = r\mathbf{v}$ and we are through, so suppose by way of contradiction that $\mathbf{z} \neq \mathbf{0}$. Then, since $\mathbf{z} \geq \mathbf{0}, \mathbf{z} \neq \mathbf{0}$, and $A^l > \mathbf{0}$, it follows that

$$\mathbf{0} < \mathbf{z}A^l = (\mathbf{v}A^l)A - r(\mathbf{v}A^l).$$

Now $\mathbf{v}A^l > \mathbf{0}$ since $\mathbf{v} \geq \mathbf{0}, \mathbf{v} \neq \mathbf{0}$ and $A^l > \mathbf{0}$. Hence $\mathbf{v}A^l$ can be scaled such that $y := (1/\mu)\mathbf{v}A^l$ belongs to $\mathbb{S}_n$. Since dividing by the constant μ does not affect the sign of anything, we have

$$\mathbf{0} < \mathbf{y}A - r\mathbf{y}.$$

Hence we can choose a number $\lambda > r$ such that $\mathbf{y}A - \lambda\mathbf{y} \geq \mathbf{0}$. So $r(\mathbf{y})$ is strictly larger than $r = r(A)$. But this contradicts the definition of $r(A)$ in (4.10). Hence it must be the case that $\mathbf{z} = \mathbf{0}$, that is, $\mathbf{v}A = r\mathbf{v}$. Therefore $r(A)$ is an eigenvalue of A. In turn this implies that $\rho(A) \geq r(A)$. But $\rho(A) \leq r(A)$ from Lemma 4.20. We conclude that $\rho(A) = r(A)$ and that $\rho(A)$ is an eigenvalue of A.

(ii). It has already been established that $\mathbf{v}A = r\mathbf{v}$ for some $\mathbf{v} \geq \mathbf{0}$, $\mathbf{v} \neq \mathbf{0}$. Hence $\mathbf{v}A^l = [\rho(A)]^l\mathbf{v}$. Moreover, $\mathbf{v} \geq \mathbf{0}, \mathbf{v} \neq \mathbf{0}$ and $A^l > \mathbf{0}$ together imply that $\mathbf{v}A^l > \mathbf{0}$. Finally $\mathbf{v} = (1/[\rho(A)]^l)\mathbf{v}A^l > \mathbf{0}$. Hence there exists a strictly positive row eigenvector of A corresponding to the eigenvalue $\rho(A)$.

(iii). Suppose $\mathbf{0} \leq B \leq A$. Let β be an eigenvalue of B and let $\mathbf{y} \in \mathbb{C}^n$ be an associated *column* eigenvector of B; that is, $B\mathbf{y} = \beta\mathbf{y}$. Since $B \leq A$, it follows that $\mathbf{x}B \leq \mathbf{x}A \; \forall x \in \mathbb{R}^n_+$. Also, $B\mathbf{y} = \beta\mathbf{y}$ implies that

$$|\beta|\mathbf{y}_+ \leq B\mathbf{y}_+ \leq A\mathbf{y}_+. \tag{4.11}$$

Multiplying both sides by $\mathbf{v}$ gives

$$|\beta|\mathbf{v}\mathbf{y}_+ \leq \mathbf{v}A\mathbf{y}_+ = \rho(A)\mathbf{v}\mathbf{y}_+.$$

Now $\mathbf{v}\mathbf{y}_+ > 0$ since $\mathbf{v} > \mathbf{0}$ and $\mathbf{y}_+ \geq \mathbf{0}, \mathbf{y}_+ \neq \mathbf{0}$. So we can divide both sides of the above inequality by $\mathbf{v}\mathbf{y}_+ > 0$, and conclude that $|\beta| \leq \rho(A)$. Since β is an arbitrary eigenvalue of A, it follows that $\rho(B) \leq \rho(A)$.

To prove the second part of the claim, suppose $|\beta| = \rho(A) = r$. Then (4.11) implies that

$$\mathbf{z} := A\mathbf{y}_+ - r\mathbf{y}_+ \geq \mathbf{0}.$$

As in the proof of Statement (i), if $\mathbf{z} \neq \mathbf{0}$, then we get a contradiction to the definition of $r(A)$. Hence $\mathbf{z} = \mathbf{0}$, or $A\mathbf{y}_+ = r\mathbf{y}_+$. Now we can multiply both sides by $A^l > \mathbf{0}$ to get $A(A^l)\mathbf{y}_+ = rA^l\mathbf{y}_+$ and in addition, $A^l\mathbf{y}_+ > \mathbf{0}$. Also, since $A\mathbf{y}_+ = r\mathbf{y}_+$, we have that $A^l\mathbf{y}_+ = r^l\mathbf{y}_+ > \mathbf{0}$, which means that $\mathbf{y}_+ > \mathbf{0}$. Now (4.11) also implies that

$$r\mathbf{y}_+ = B\mathbf{y}_+ = A\mathbf{y}_+, \text{ or } (B - A)\mathbf{y}_+ = \mathbf{0},$$

since the two extreme inequalities are in fact equalities. Since $B - A \leq \mathbf{0}$ and $\mathbf{y}_+ > \mathbf{0}$, the only way for $(B-A)\mathbf{y}_+$ to equal $\mathbf{0}$ is to have $B = A$. Hence Statement (iii) is proved.

(iv). Suppose $r\mathbf{y} = \mathbf{y}A$, so that $\mathbf{y}$ is also a row eigenvector of A corresponding to the eigenvalue $r = \rho(A)$. Then, as in the proof of Lemma 4.18 and Statement (i) above, it follows that $\mathbf{y}_+$ also satisfies $r\mathbf{y}_+ = \mathbf{y}_+A$. Also, by noting that $\mathbf{y}_+A^l = r^l\mathbf{y}_+ > \mathbf{0}$, and that $\mathbf{y}_+ \geq \mathbf{0}$ and $\mathbf{y}_+ \neq \mathbf{0}$, it follows that $\mathbf{y}_+A^l > \mathbf{0}$ and hence $\mathbf{y}_+ > \mathbf{0}$. Hence *every* row eigenvector of A corresponding to the eigenvalue $r = \rho(A)$ must have all nonzero components. Now let $\mathbf{v}$ denote the vector identified in the proof of Statement (i). Then $r(\mathbf{v}-c\mathbf{y}) = (\mathbf{v}-c\mathbf{y})A$ for all complex constants c. Suppose $\mathbf{y}$ is not a multiple of $\mathbf{v}$. Then it is possible to choose the constant c in such a way $\mathbf{v} - c\mathbf{y} \neq \mathbf{0}$, but at least one component of $\mathbf{v} - c\mathbf{y}$ equals zero. Since $\mathbf{v} - c\mathbf{y} \neq \mathbf{0}$, $\mathbf{v} - c\mathbf{y}$ is a row eigenvector of A corresponding to the eigenvalue $r = \rho(A)$, and it has at least one component equal to zero, which is a contradiction. Hence $\mathbf{y}$ is a multiple of $\mathbf{v}$. Thus it has been shown that the equation $\mathbf{y}(rI - A) = \mathbf{0}$ has only one independent solution, i.e., that $\text{Rank}(rI - A) = n - 1$.

(v). Let $\phi(\lambda) := \det(\lambda I - A)$ denote the characteristic polynomial of A, and let $\text{Adj}(\lambda I - A)$ denote the adjoint matrix of $\lambda I - A$. We have already seen that $\rho(A) = r$ is an eigenvalue of A; hence $\phi(r) = 0$. Now it is shown that $\phi'(r) \neq 0$, which is enough to show that r is a *simple* zero of the polynomial $\phi(\cdot)$ and hence a simple eigenvalue of A.

For brevity let $M(\lambda)$ denote $\text{Adj}(\lambda I - A)$. Then it is easy to see that each element of $M(\lambda)$ is a polynomial of degree $\leq n - 1$ in λ. For every value of λ, we have

$$(\lambda I_n - A)M(\lambda) = \phi(\lambda)I_n.$$

Differentiating both sides with respect to λ leads to

$$M(\lambda) + (\lambda I_n - A)M'(\lambda) = \phi'(\lambda).$$

Let $\mathbf{v} > \mathbf{0}$ denote the eigenvector found in the proof of Statement (i). Then $r\mathbf{v} = \mathbf{v}A$, or $\mathbf{v}(rI_n - A) = \mathbf{0}$. So if we right-multiply both sides of the above equation by $\mathbf{v}$ and substitute $\lambda = r$, we get

$$\phi'(r)\mathbf{v} = \mathbf{v}M(r) + \mathbf{v}(rI_n - A)M'(\lambda) = \mathbf{v}M(r). \tag{4.12}$$

Thus the proof is complete if it can be shown that $\mathbf{v}M(r) > \mathbf{0}$, because that is enough to show that $\phi'(r) > 0$.

We begin by establishing that no row of $M(r)$ is identically zero. Specifically, it is shown that the diagonal elements of $M(r)$ are all strictly positive. For each index $i \in \mathcal{N}$, the ii-th element of $M(r)$ is the principal minor given by

$$m_{ii}(r) = \det(rI_{n-1} - \bar{A}_i),$$

where $\bar{A}_i \in \mathbb{R}_+^{(n-1)\times(n-1)}$ is obtained from A by deleting the i-th row and i-th column. Now it is claimed that $\rho(\bar{A}_i) < \rho(A) = r$. To see this, suppose first (for notational convenience only) that $i = 1$. Then the two matrices

$$\bar{A}_1 = \begin{bmatrix} a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \dots & a_{nn} \end{bmatrix}, \text{ and } A_1^* = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

have exactly the same spectrum, except that A_1^* has an extra eigenvalue at zero. For other values of i, A_i^* equals A with the i-th row and i-th column set equal to zero, while $\bar{A}_i$ is obtained from A by deleting the i-th row and i-th column. Hence, for all i, the matrices A_i^* and $\bar{A}_i$ have the same spectrum, except that A_i^* has an extra eigenvalue at zero. Now $\mathbf{0} \leq A_i^* \leq A$. Moreover, $A_i^* \neq A$ since A is irreducible and thus cannot have an identically zero row or zero column. Hence by Statement (iii) above, $\rho(A_i^*) < \rho(A) = r$. In particular, r is not an eigenvalue of $\bar{A}_i$, and as a consequence,

$$m_{ii}(r) = \det(rI_{n-1} - \bar{A}_i) \neq 0 \; \forall i \in \mathcal{N}.$$

Actually, we can conclude that

$$m_{ii}(r) = \det(rI_{n-1} - \bar{A}_i) > 0 \; \forall i \in \mathcal{N}. \tag{4.13}$$

To see this, define the characteristic polynomial $\phi_i(\lambda) := \det(\lambda I - \bar{A}_i)$. Then $\phi_i(\lambda) > 0$ when l is positive and sufficiently large. So if $\phi_i(r) < 0$, then $\phi_i(\cdot)$ would have a real zero $> r$, which contradicts the fact that $\rho(A_i^*) < r$. Hence (4.13) is established.

We have thus far established that every row of $M(r)$ contains a positive element. Now

$$M(r)(rI_n - A) = \phi(r)I_n = \mathbf{0}.$$

Hence every nonzero row of $M(r)$ is a row eigenvector of A corresponding to the eigenvalue r. Since $m_{ii}(r) > 0$ for each i, it follows that every row of $M(r)$ is a positive multiple of $\mathbf{v}$. In other words, $M(r)$ has the form

$$M(r) = \mathbf{w}\mathbf{v}, \mathbf{w} \in \mathbb{R}_+^n.$$

Now let us return to (4.12). Substituting for $M(r)$ and "cancelling" $\mathbf{v}$ (which is permissible since $\mathbf{v} > \mathbf{0}$) shows that

$$\phi'(r) = \mathbf{w}\mathbf{v} > 0.$$

Hence r is a simple eigenvalue of A.

(vi). The claim is that if λ is any eigenvalue of A such that $|\lambda| = \rho(A) = r$, then of necessity $\lambda = \rho(A)$. Suppose $|\lambda| = \rho(A) = r$ with corresponding row eigenvector $\mathbf{z} \in \mathbb{C}^n$. Then

$$\lambda z_j = \sum_{i=1}^{n} z_i a_{ij}, \; \forall j \in \mathcal{N}, \text{ and } \lambda^l z_j = \sum_{i=1}^{n} z_i a_{ij}^{(l)}, \; \forall j \in \mathcal{N}.$$

As before, let $\mathbf{z}_+ \in \mathbb{R}_+^n$ denote the vector $[|z_1| \ldots |z_n|]$. Then, just as in the proof of Lemma 4.18, $\lambda\mathbf{z} = \mathbf{z}A$ implies that $|\lambda|\mathbf{z}_+ \leq \mathbf{z}_+A$. However, $|\lambda| = r$ by assumption. Hence, as in the proof of Statement (i), $r\mathbf{z}_+ \leq \mathbf{z}_+A$ implies that in fact $r\mathbf{z}_+$ *equals* $\mathbf{z}_+A$, and as a consequence $r^l\mathbf{z} = \mathbf{z}A^l$. In other words,

$$|\lambda^l z_j| = \left| \sum_{i=1}^{n} z_i a_{ij}^{(l)} \right| = \sum_{i=1}^{n} |z_i| a_{ij}^{(l)}, \; \forall j \in \mathcal{N}.$$

Since the z_i are complex numbers in general, and since $a_{ij}^{(l)} > 0$ for all i, j, the magnitude of the sum of $z_i a_{ij}^{(l)}$ equals the sum of the magnitudes of $z_i a_{ij}^{(l)}$ if and only if all these complex numbers are *aligned*, i.e., there is a *common* number θ such that

$$z_i a_{ij}^{(l)} = |z_i| a_{ij}^{(l)} \exp(\mathbf{i}\theta) \; \forall i, j \in \mathcal{N},$$

where $\mathbf{i} = \sqrt{-1}$. Dividing through by the positive number a_{ij} shows that

$$z_i = |z_i| \exp(\mathbf{i}\theta) \; \forall i \in \mathcal{N}.$$

Let us return to $r\mathbf{z}_+ = \mathbf{z}_+ A$. By Statement (iv), this implies that $\mathbf{z}_+ = c\mathbf{v}$ for some positive constant c, or

$$\mathbf{z} = c \exp(\mathbf{i}\theta)\mathbf{v}.$$

Therefore

$$\mathbf{z}A = c \exp(\mathbf{i}\theta)\mathbf{v}A = c \exp(\mathbf{i}\theta) r\mathbf{v} = r\mathbf{z}.$$

Since $\mathbf{z}A = \lambda z$, this shows that $r\mathbf{z} = \lambda\mathbf{z}$, and since at least one component of $\mathbf{z}$ is nonzero, this shows that $\lambda = r$. □

4.2.2 Perron-Frobenius Theorem for Irreducible Matrices

In this section, we study irreducible matrices that have a period $p \geq 2$. From Theorem 4.14, we know that an aperiodic irreducible matrix is primitive. So Theorem 4.21 applies to such matrices.

There are two distinct ways to approach the study of such matrices. First, we can examine the proof of Theorem 4.21 and see how much of it can be salvaged for periodic matrices. Second, we can use Theorem 4.19 in conjunction with Theorem 4.21. It turns out that each approach leads to its own distinctive set of insights. We begin with a preliminary result that is useful in its own right.

Lemma 4.22 *Suppose $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$. Then every nonzero eigenvalue of AB is also an eigenvalue of BA.*

Proof. Suppose $\lambda \neq 0$ is an eigenvalue of AB, and choose $\mathbf{x} \neq \mathbf{0}$ such that $AB\mathbf{x} = \lambda\mathbf{x}$. Clearly $B\mathbf{x} \neq \mathbf{0}$. Now

$$BAB\mathbf{x} = \lambda B\mathbf{x}.$$

Since $B\mathbf{x} \neq \mathbf{0}$, this shows that λ is also an eigenvalue of BA with corresponding eigenvector $B\mathbf{x}$. The converse follows by interchanging A and B throughout. □

Note that the above reasoning breaks down if $\lambda = 0$. Indeed, if $n \neq m$ (say $n > m$ to be precise), then the larger matrix (AB if $n > m$) must necessarily be rank deficient, and thus must have at least $n - m$ eigenvalues at zero. But if BA is nonsingular, then it will not have any eigenvalues at zero. Therefore Lemma 4.22 can be stated as follows: Given two matrices A, B of complementary dimensions (so that both AB and BA are well-defined), both AB and BA have the same spectrum, except for the possibility that one of them may have some extra eigenvalues at zero. In particular, $\rho(AB) = \rho(BA)$.

Theorem 4.23 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible and has period $p \geq 2$. Then*

(i) The eigenvalues of A exhibit cyclic symmetry with period p. Specifically, if λ is an eigenvalue of A, so are

$$\lambda \exp(\mathbf{i}2\pi k/p), k = 1, \ldots, p - 1.$$

(ii) In particular, $r = \rho(A)$ is an eigenvalue of A, and so are

$$r \exp(\mathbf{i}2\pi k/p), k = 1, \ldots, p-1.$$

Proof. **(i).** Suppose $\lambda \in \mathbb{C}$ is an eigenvalue of A, and note that if $\lambda = 0$, then $\lambda \exp(\mathbf{i}2\pi k/p) = 0$ for all k. So we need to study only the case where $\lambda \neq 0$. Without loss of generality, we can assume that A is in the canonical form (4.8), because permuting the rows and columns of a matrix does not change its spectrum. So we use the symbol A (instead of B) for the canonical form. Suppose $\lambda \in \mathbb{C}$ is an eigenvalue of A, and choose $\mathbf{x} \neq \mathbf{0}$ such that $\lambda\mathbf{x} = \mathbf{x}A$. Partition $\mathbf{x}$ commensurately with the canonical form. Then $\lambda\mathbf{x} = \mathbf{x}A$ can be also be partitioned, and implies that

$$\mathbf{x}_j A_{j,j+1} = \lambda \mathbf{x}_{j+1}, j = 0, \ldots, p-1, \tag{4.14}$$

where if $j = p-1$ we take $j+1 = 0$ since $p = 0 \mod p$. Now let $\alpha := \exp(\mathbf{i}2\pi/k)$ denote the p-th root of 1. Fix an integer k between 1 and $p-1$, and define the vector $\mathbf{y}^{(k)} \in \mathbb{C}^n$ by

$$\mathbf{y}_j^{(k)} = \alpha^{-jk}\mathbf{x}_j = \exp(-\mathbf{i}2\pi jk/p)\mathbf{x}_j, j = 0, \ldots, p-1.$$

Now it follows from (4.14) that

$$\begin{aligned} \mathbf{y}_j^{(k)} A_{j,j+1} &= \alpha^{-jk}\mathbf{x}_j A_{j,j+1} \\ &= \alpha^{-jk}\lambda\mathbf{x}_{j+1} = \lambda\alpha^k\alpha^{-(j+1)k}\mathbf{x}_{j+1} \\ &= \lambda\alpha^k\mathbf{y}_{j+1}. \end{aligned}$$

Hence $\mathbf{y}^{(k)}$ is an eigenvector of A corresponding to the eigenvalue $\lambda\alpha^k$. This argument can be repeated for each integer k between 1 and $p-1$. This establishes Statement (i).

(ii). Again, assume without loss of generality that A is in the canonical form (4.8), and define matrices $M_0, \ldots, M_{p-1}$ as in (4.9). Thus, as in the proof of Theorem 4.19, it follows that

$$A^p = \text{Block Diag}\{M_0, \ldots, M_{p-1}\}.$$

Hence the spectrum of A^p is the union of the spectra of $M_0, \ldots, M_{p-1}$. Moreover, it is clear from (4.9) that these p matrices are just cyclic products of A_{01} through $A_{p-1,0}$. Hence repeated application of Lemma 4.22 shows that each of these matrices has the same set of nonzero eigenvalues. (Since these matrices may have different dimensions, they will in general have different numbers of eigenvalues at zero.) As a consequence

$$\rho(M_0) = \rho(M_1) = \cdots = \rho(M_{p-1}) =: c, \text{ say.}$$

We also know from Theorem 4.19 that each of the p matrices $M_0, \ldots, M_{p-1}$ is primitive. Hence $\rho(M_i) = c$ is an eigenvalue of M_i for each i. Therefore c is a p-fold eigenvalue of A^p. Since the spectrum of A^p consists of just the p-th powers of the spectrum of A, we see that some p-th roots of c are also eigenvalues of A. However, it now follows from Statement (i) that in fact *every* p-th root of c is an eigenvalue of A. This is precisely Statement (ii). □

Theorem 4.24 *Suppose $A \in \mathbb{R}_+^{n\times n}$ is irreducible. Then*

(i) $\rho(A)$ is an eigenvalue of A.

(ii) There exists a row eigenvector $\mathbf{v} > \mathbf{0}$ of A corresponding to the eigenvalue $\rho(A)$.

(iii) If $B \in \mathbb{R}^{n\times n}$ satisfies $\mathbf{0} \leq B \leq A$, then $\rho(B) \leq \rho(A)$, with equality if and only if $B = A$.

(iv) Rank$[\rho(A)I - A] = n-1$, so that the eigenvectors of A associated with the eigenvalue $\rho(A)$ are multiples of each other.

(v) $\rho(A)$ is a simple eigenvalue of A.

(vi) If A is aperiodic, then every other eigenvalue λ of A satisfies $|\lambda| < \rho(A)$. Otherwise, if A has period $p \geq 2$, then each of the p numbers $\rho(A)\exp(\mathbf{i}2\pi k/p), k = 0, \ldots, p-1$ is an eigenvalue of A. All other eigenvalues λ of A satisfy $|\lambda| < \rho(A)$.

Remarks: Comparing Theorem 4.21 and 4.24, we see that only Statement (vi) is different in the two theorems.

Proof. The proof consists of mimicking the proof of Theorem 4.21 with minor changes. Hence we give the proof in a very sketchy form.

Since A is irreducible, it follows from Corollary 4.7 that the matrix

$$M := I + B = \sum_{i=0}^{n-1} A^i$$

is strictly positive. Moreover, if $\mathbf{z}\mathcal{A} = \lambda\mathbf{z}$, then

$$\mathbf{z}M = \left(\sum_{i=0}^{n-1} \lambda^i\right)\mathbf{z}.$$

(i). This is already established in Theorem 4.23, but we give an independent proof paralleling that of Theorem 4.21. Define $r(A)$ as in Lemma 4.20, and choose $\mathbf{v} \in \mathbb{S}_n$ such that $r(\mathbf{v}) = r(A)$. Define $\mathbf{z} = \mathbf{v}A - r\mathbf{v}$ where $r = r(A)$. Then $\mathbf{z} \geq \mathbf{0}$. If $\mathbf{z} = \mathbf{0}$ then we are through, so suppose by way of contradiction that $\mathbf{z} \neq \mathbf{0}$. Then $\mathbf{z}M > \mathbf{0}$ because $M > \mathbf{0}$. Thus $\mathbf{b} < \mathbf{z}M = \mathbf{v}AM - r\mathbf{v}M = \mathbf{v}MA - r\mathbf{v}M$ because A and M commute. Moreover $\mathbf{v}M > \mathbf{0}$ since $\mathbf{v} \in \mathbb{S}_n$ and $M > \mathbf{0}$. So if we define $\mathbf{y} = (1/\mu)\mathbf{v}M$, then $\mathbf{y} \in \mathbb{S}_n$ for a suitable scaling constant μ, and $\mathbf{y}A - r\mathbf{y} > \mathbf{0}$. This contradicts the definition of $r(A)$. Hence $\mathbf{z} = \mathbf{0}$ and r is an eigenvalue of A.

(ii). We know that $\mathbf{v}A = r\mathbf{v}$. So $\mathbf{v}AM = \mathbf{v}MA = r\mathbf{v}M$, and $\mathbf{v}M > \mathbf{0}$. So there exists a strictly positive eigenvector of A corresponding to r.

(iii). The proof is identical to the proof of Statement (iii) in Theorem 4.21, except that instead of multiplying by A^l we multiply by M.

(iv). Ditto.

(v). The proof is identical to that of Statement (v) in Theorem 4.21.

(vi). It has already been shown in Statement (ii) of Theorem 4.23 that each of the complex numbers $\rho(A) \exp \mathbf{i} 2\pi k/p, k = 0, \ldots, p-1$ is an eigenvalue of A. So it remains only to show that there are no other eigenvalues of A with magnitude $\rho(A)$. Suppose λ is an eigenvalue of A. Then λ^p is an eigenvalue of the matrix A^p. Since A has the canonical form (4.8), it follows that λ^p is an eigenvalue of one of the matrices M_i. Each of these matrices is primitive, so every eigenvalue other than $[\rho(A)]^p$ has magnitude strictly less than $[\rho(A)]^p$. Hence we either have $\lambda^p = [\rho(A)]^p$, or else $|\lambda|^p < [\rho(A)]^p$. This is precisely Statement (vi). □

Chapter Five

Markov Chains

In this chapter, we begin our study of Markov processes, which form the core topic of the book. We define the "Markov property," and show that all the relevant information about a Markov process assuming values in a finite set of cardinality n can be captured by a nonnegative $n \times n$ matrix called the "state transition matrix." Then we invoke the results of Chapter 4 on nonnegative matrices to analyze the temporal evolution of Markov processes. Then we proceed to a discussion of more advanced topics such as hitting times and ergodicity. Still more advanced topics such as mixing, and parameter estimation are discussed in the next chapter.

5.1 THE MARKOV PROPERTY AND THE STATE TRANSITION MATRIX

Throughout, $\mathbb{N}$ denotes a finite set $\{x_1, \ldots, x_n\}$.[1] For the purposes of the simplified setting studied here, we define a "stochastic process" over $\mathbb{N}$ to be a sequence of random variables $\{\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \ldots\}$ or $\{\mathcal{X}_t\}_{t=0}^{\infty}$ for short, where each $\mathcal{X}_t$ is a random variable assuming values in the set $\mathbb{N}$. Though the index t could in principle stand for just about anything, it is most common to think of t as representing "time." If we think of the parameter t as representing "time," then notions such as "past" and "future" make sense. Thus if $t < t'$, then $\mathcal{X}_t$ is a "past" variable for $\mathcal{X}_{t'}$, while $\mathcal{X}_{t'}$ is a "future" variable for $\mathcal{X}_t$. However, the index need not always represent time. For instance, when the stochastic process corresponds to the genome sequence of an organism, the set $\mathbb{N}$ is the four symbol nucleotide alphabet $\{A, C, G, T\}$, and the sequencing is spatial rather than temporal. Similarly when we think of the primary structure of a protein, the set $\mathbb{N}$ is the twenty symbol set of amino acids. In the case of both DNA and proteins, the sequences have a definite spatial direction and therefore cannot be "read backwards." This spatial orientation allows us to replace the "past" and "future" by "earlier" and "later."

Another point worth emphasizing is is the abstract nature of the set $\mathbb{N}$. We write $\mathbb{N} = \{x_1, \ldots, x_n\}$ as opposed to $\mathbb{N} = \{1, \ldots, n\}$ to dispel the notion that the elements of the set $\mathbb{N}$ can be identified with *integers*. (Having said this, we will sometimes write $\mathbb{N} = \{1, \ldots, n\}$ if that makes the notation

[1] Note that until now we had been using the symbols $\mathbb{A}$ and $\mathbb{B}$ to denote finite sets.

less cumbersome.) Going back to the DNA example, in this book we write the four nucleotides in the order $\{A, C, G, T\}$, thus arranging Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) in English alphabetical order. But some authors write them in the order A, T, C, G or some other permutation of the four symbols. Clearly, in order to be meaningful, our methods of analysis *cannot* depend on the order in which these symbols are arranged, and must possess an inherent "permutation invariance." To summarize the discussion, the ordering of the elements within the set $\mathbb{N}$ is not meaningful or permanent, but the ordering of the parameter t *is* indeed meaningful.

Let us return to the "stochastic process" $\{\mathcal{X}_t\}_{t=0}^{\infty}$. For each integer T, the random variables $(\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_T)$ have a *joint distribution*, which is a probability distribution over the *finite* set X^{T+1}. Since in this book we try to "avoid the infinite" to the extent possible, we will not speak about the "joint law" of all the infinitely many random variables taken together. However, it is essential to emphasize that our exposition is somewhat constricted due to this self-imposed restriction, and is definitely somewhat impoverished as a consequence. The ability to "cope with the infinite" in a mathematically meaningful and consistent manner is one of the substantial achievements of axiomatic probability theory.

Now we introduce a very fundamental property called the Markov property.

Definition 5.1 *The process* $\{\mathcal{X}_t\}_{t=0}^{\infty}$ *is said to* **possess the Markov property***, or to be a* **Markov process***, if for every* $t \geq 1$ *and every sequence* $u_0 \ldots u_{t-1} u_t \in \mathbb{N}^{t+1}$*, it is true that*

$$\Pr\{\mathcal{X}_t = u_t | \mathcal{X}_0 = u_0, \ldots \mathcal{X}_{t-1} = u_{t-1}\} = \Pr\{\mathcal{X}_t = u_t | \mathcal{X}_{t-1} = u_{t-1}\}. \quad (5.1)$$

Recall that a conditional probability of $\mathcal{X}_t$, irrespective of on what measurements it is conditioned, is a probability distribution on the set $\mathbb{N}$. The Markov property states therefore that the conditional probability distribution of the "current state" $\mathcal{X}_t$ depends only on the "immediate past" state $\mathcal{X}_{t-1}$, and not on any of the previous states. Thus adding some more measurements prior to time $t-1$ does not in any way alter the conditional probability distribution of $\mathcal{X}_t$.

For convenience, we introduce the notation $\mathcal{X}_j^k$ to denote the entity $(\mathcal{X}_i, j \leq i \leq k$. Alternatively, $\mathcal{X}_j^k = (\mathcal{X}_j, \mathcal{X}_{j+1}, \ldots, \mathcal{X}_{k-1}, \mathcal{X}_k)$. Clearly this notation makes sense only if $j \leq k$. With this notation, we can rephrase Definition 5.1 as follows: The stochastic process $\{\mathcal{X}_t\}$ is a Markov process if, for every $(u_0, \ldots, u_t) \in \mathbb{N}^{t+1}$, it is true that

$$\Pr\{\mathcal{X}_t = u_t | \mathcal{X}_0^{t-1} = u_0 \ldots u_{t-1}\} = \Pr\{\mathcal{X}_t = v | \mathcal{X}_{t-1} = u_{t-1}\}.$$

For *any* stochastic process $\{\mathcal{X}_t\}$ and *any* sequence $u_0 \ldots u_{t-1} u_t \in \mathbb{N}^{t+1}$, we can always write

$$\Pr\{\mathcal{X}_0^t = u_0 \ldots u_t\} = \Pr\{\mathcal{X}_0 = u_0\} \cdot \prod_{i=0}^{t-1} \Pr\{\mathcal{X}_{i+1} = u_{i+1} | \mathcal{X}_0^i = u_0 \ldots u_i\}.$$

This follows from repeated application of the definition of conditional probability. However, if the process under study is a Markov process, then the above formula can be simplified to

$$\Pr\{\mathcal{X}_0^t = u_0 \ldots u_t\} = \Pr\{\mathcal{X}_0 = u_0\} \cdot \prod_{i=0}^{t-1} \Pr\{\mathcal{X}_{i+1} = u_{i+1} | \mathcal{X}_i = u_i\}. \quad (5.2)$$

Thus, with a Markov process, the length of the "tail" on which the conditioning is carried out is always one.

In the probability literature, one often uses the name "Markov chain" to denote a Markov process $\{\mathcal{X}_t\}$ where the underlying parameter t assumes only discrete values (as opposed to taking values in a continuum, such as $\mathbb{R}_+$ for example). In the present book, attention is restricted only to the case where t assumes values in $\mathbb{Z}_+$, the set of nonnegative integers. Accordingly, we use the expressions "Markov process" and "Markov chain" interchangeably.

The formula (5.2) demonstrates the importance of the quantity

$$\Pr\{\mathcal{X}_{t+1} = u | \mathcal{X}_t = v\},$$

viewed as a function of three entities: The "current" state $v \in \mathbb{N}$, the "next state" $u \in \mathbb{N}$, and the "current time" $t \in \mathbb{Z}_+$. Recall that $\mathbb{N}$ is a finite set. Let us fix the time t for the time being, and define the quantity

$$a_{ij}(t) := \Pr\{\mathcal{X}_{t+1} = j | \mathcal{X}_t = i\}, i, j \in \mathbb{N} = \{1, \ldots, n\}, t \in \mathbb{Z}_+. \quad (5.3)$$

Thus $a_{ij}(t)$ is the probability of making a transition from the current state i to the next state j at time t.

Definition 5.2 *The $n \times n$ matrix $A(t) = [a_{ij}(t)]$ is called the* **state transition matrix** *of the Markov process at time t. The Markov chain is said to be* **homogenous** *if $A(t)$ is a constant matrix for all $t \in \mathbb{Z}_+$, and* **inhomogenous** *otherwise.*

Lemma 5.3 *Suppose $\{\mathcal{X}_t\}$ is a Markov process assuming values in a finite set $\mathbb{N}$ of cardinality n, and let $A(t)$ denote its state transition matrix at time t. Then $A(t)$ is a* **stochastic matrix** *for all t. That is:*

$$a_{ij}(t) \in [0, 1] \; \forall i, j \in \mathbb{N}, t \in \mathbb{Z}_+.$$

$$\sum_{j=1}^{n} a_{ij}(t) = 1 \; \forall i \in \mathbb{N}, t \in \mathbb{Z}_+.$$

Proof. Both properties are obvious from the definition. A conditional probability always lies between 0 and 1. Moreover, the sum of the conditional probabilities over all possible outcomes equals one. □

Lemma 5.4 *Suppose $\{\mathcal{X}_t\}$ is a Markov process assuming values in a finite set $\mathbb{N}$ of cardinality n, and let $A(t)$ denote its state transition matrix at time t. Suppose the initial state $\mathcal{X}_0$ is distributed according to $\mathbf{c}_0 \in \mathbb{S}_n$. That is,*

$$\Pr\{\mathcal{X}_0 = x_i\} = c_i \; \forall i \in \mathbb{N}.$$

Then for all $t \geq 0$, the state $\mathcal{X}_t$ is distributed according to

$$\mathbf{c}_t = \mathbf{c}_0 A(0)A(1)\ldots A(t-1). \tag{5.4}$$

Proof. Pick an arbitrary element $u_t \in \mathbb{N}$. Then it follows from (5.2) that

$$\begin{aligned}\Pr\{\mathcal{X}_t = u_t\} &= \sum_{u_0 u_1 \ldots u_t \in \mathbb{N}^{t+1}} \Pr\{\mathcal{X}_0 = u_0\} \cdot \prod_{i=0}^{t-1} \Pr\{\mathcal{X}_{i+1} = u_{i+1} | \mathcal{X}_i = u_i\} \\ &= \sum_{u_0 u_1 \ldots u_t \in \mathbb{N}^{t+1}} c_{u_0} a_{u_0,u_1}(0) \ldots a_{u_{t-1},u_t}(t-1).\end{aligned}$$

But the right side is just the u_t-th component of $\mathbf{c}_t = \mathbf{c}_0 A(0)A(1)\ldots A(t-1)$ written out in expanded form. □

Example 5.1 This example is a variation on the card game "blackjack," in which the objective is to keep drawing cards until the total value of the cards drawn equals or exceeds twenty one. In the present simplified version, the thirteen cards are replaced by a four-sided die. (It may be mentioned that in many ancient cultures, dice were made from animal bones, and had only four sides since they were oblong.) The four sides are labelled as 0, 1, 2 and 3 (as opposed to the more conventional 1, 2, 3 and 4), and are equally likely to appear on any one throw. A player rolls the die again and again, and $\mathcal{X}_t$ denotes the accumulated score after t throws. If the total *exactly* equals nine, the player wins; otherwise he loses. It is assumed that the outcome of each throw is independent of all the previous throws.

It is now shown that $\{\mathcal{X}_t\}$ is a Markov process assuming values in the set $\mathbb{N} := \{0, 1, \ldots, 8, W, L\}$ of cardinality eleven. Let $\mathcal{Y}_t$ denote the outcome of the roll of the die at time t. Then

$$\Pr\{\mathcal{Y}_t = 0\} = \Pr\{\mathcal{Y}_t = 1\} = \Pr\{\mathcal{Y}_t = 2\} = \Pr\{\mathcal{Y}_t = 3\} = 1/4.$$

Now let us examine the distribution of $\mathcal{X}_t$. We know that $\mathcal{X}_t = \mathcal{X}_{t-1} + \mathcal{Y}_t$, except that if $\mathcal{X}_{t-1} + \mathcal{Y}_t = 9$ we take $\mathcal{X}_t = W$ (win), and if $\mathcal{X}_{t-1} + \mathcal{Y}_t > 9$ we take $\mathcal{X}_t = L$ (loss). If $\mathcal{X}_{t-1} = W$ or L, then the game is effectively over, and we take $\mathcal{X}_t = \mathcal{X}_{t-1}$. These observations can be summarized in the following rules: If $\mathcal{X}_{t-1} \leq 5$, then

$$\begin{aligned}\Pr\{\mathcal{X}_t = \mathcal{X}_{t-1}\} = \Pr\{\mathcal{X}_t = \mathcal{X}_{t-1} + 1\} = \\ \Pr\{\mathcal{X}_t = \mathcal{X}_{t-1} + 2\} = \Pr\{\mathcal{X}_t = \mathcal{X}_{t-1} + 3\} = 1/4.\end{aligned}$$

If $\mathcal{X}_{t-1} = 6$, then

$$\Pr\{\mathcal{X}_t = 6\} = \Pr\{\mathcal{X}_t = 7\} = \Pr\{\mathcal{X}_t = 8\} = \Pr\{\mathcal{X}_t = W\} = 1/4.$$

If $\mathcal{X}_{t-1} = 7$, then

$$\Pr\{\mathcal{X}_t = 7\} = \Pr\{\mathcal{X}_t = 8\} = \Pr\{\mathcal{X}_t = W\} = \Pr\{\mathcal{X}_t = L\} = 1/4.$$

If $\mathcal{X}_{t-1} = 8$, then

$$\Pr\{\mathcal{X}_t = 8\} = \Pr\{\mathcal{X}_t = W\} = 1/4, \Pr\{\mathcal{X}_t = L\} = 2/4.$$

Finally, if $\mathcal{X}_{t-1} = W$ or L, then $\Pr\{\mathcal{X}_t = \mathcal{X}_{t-1}\} = 1$.

The process $\{\mathcal{X}_t\}$ is a Markov process because the probability distribution of $\mathcal{X}_t$ depends *only* on the value of $\mathcal{X}_{t-1}$, and does not at all depend on *how* the score happened to reach $\mathcal{X}_{t-1}$. In other words, when it comes to determing the probability distribution of $\mathcal{X}_t$, only the value of $\mathcal{X}_{t-1}$ is relevant, and all past values of $\mathcal{X}_i, i \leq t-2$ are irrelevant.

The state transition matrix of the Markov process is an 11×11 matrix given by

$$A = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 2/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Since the transition matrix does not explicitly depend on t, the Markov chain is homogeneous.

It is natural that the game begins with the initial score equal to zero. Thus the "random variable" $\mathcal{X}_0$ equals zero, or in other words, the initial distribution $\mathbf{c} \in \mathbb{R}^{1 \times 11}$ has a 1 in the first component and zeros elsewhere. Now repeated application of the formula (5.4) gives the distributions of the random variables $\mathcal{X}_1, \mathcal{X}_2$, etc. If $\mathbf{c}_t$ denotes the distribution of $\mathcal{X}_t$, then we have

$$\mathbf{c}_0 = \mathbf{c} = [1 \;\; 0 \ldots 0],$$

$$\mathbf{c}_1 = \mathbf{c}_0 A = [1/4 \;\; 1/4 \;\; 1/4 \;\; 1/4 \;\; 0 \ldots 0],$$

$$\mathbf{c}_2 = [1/16 \;\; 2/16 \;\; 3/16 \;\; 4/16 \;\; 3/16 \;\; 2/16 \;\; 1/16 \;\; 0 \;\; 0 \;\; 0 \;\; 0],$$

and so on. One noteworthy feature of this Markov process is that it is *nonstationary.* Note that each $\mathcal{X}_t$ has a different distribution. The precise definition of a stationary process is that the joint distribution (or law) of all the infinitely many random variables $\{\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \ldots\}$ is the same as the joint law of the variables $\{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \ldots\}$. A necessary, but not sufficient, condition for stationarity is that each individual random variable must have the same distribution. This condition does not hold in the present case. So it is important to note that a homogeneous Markov chain can still be nonstationary – it depends on what the initial distribution is.

Another noteworty point about this Markov chain is that if we examine the distribution $\mathbf{c}_t$, then $\Pr\{\mathcal{X}_t = 0 \text{ through } 8\}$ approaches zero as $t \to \infty$. In plain English, all games will "eventually" wind up in either a win or a loss. All other states are "transient." This idea is made precise in the next section.

5.2 DYNAMICS OF STATIONARY MARKOV CHAINS

In this section, we study the dynamics of Markov chains where the state transition matrix is constant with time. By applying the general results on nonnegative matrices from Chapter 4, we show that it is possible to partition the state space into "recurrent" and "transient" states. Then we analyze the dynamics in greater detail, and derive explicit formulas for the probability that a trajectory will hit a specified subset of the state space, as well as the average time needed to do so.

5.2.1 Recurrent and Transient States

In this section, we specialize the contents of Chapter 4 to stochastic matrices, which are a special kind of nonnegative matrix, and thereby draw some very useful conclusions about the temporal evolution of stationary Markov chains.

Definition 5.5 *Suppose A is a stochastic matrix of dimension $n \times n$; that is, $A \in [0,1]^{n \times n}$, and $\sum_{j=1}^{n} a_{ij} = 1$ for all $i \in \mathbb{N}$. Then a vector $\boldsymbol{\pi} \in \mathbb{S}_n$ is said to be a* **stationary distribution** *of A if $\boldsymbol{\pi} A = \boldsymbol{\pi}$.*

The significance of a stationary distribution is obvious. Suppose $\{\mathcal{X}_t\}$ is a homogeneous Markov chain with the state transition matrix A. We have seen (as in Example 5.1) that, depending on the initial distribution, the resulting process $\{\mathcal{X}_t\}$ may have different distributions at different times. However, suppose A has a stationary distribution $\boldsymbol{\pi}$ (and at the moment we don't know whether a given matrix *has* a stationary distribution). Then $\boldsymbol{\pi} A^t = \boldsymbol{\pi}$ for all t. So if $\mathcal{X}_0$ has the distribution $\boldsymbol{\pi}$, then it follows from (5.4) that $\mathcal{X}_t$ also has the same distribution $\boldsymbol{\pi}$ for all values of t. To put it in words, if a Markov chain is started off in a stationary distribution (assuming there is one), then all future states also have the same stationary distribution. It is therefore worthwhile to ascertain whether a given stochastic matrix A does indeed have a stationary distribution, and if so, to determine the set of *all* stationary distributions of A.

Theorem 5.6 *Suppose A is a stochastic matrix. Then*

1. *$\rho(A) = 1$ where $\rho(\cdot)$ is the spectral radius.*

2. *By a symmetric permutation of rows and columns, A can be put into the canonical form*

$$A = \begin{array}{c} \\ \mathcal{E} \\ \mathcal{I} \end{array} \begin{array}{c} \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \left[\begin{array}{cc} P & \mathbf{0} \\ R & Q \end{array} \right] \end{array}, \tag{5.5}$$

   where $\mathcal{E}$ is the set of essential states and $\mathcal{I}$ is the set of inessential states.

3. *If the set $\mathcal{I}$ is nonempty, then R contains at least one positive element.*

4. *By further row and column permutations, P can be put in the form*

$$P = \begin{array}{c} \\ \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_s \end{array} \begin{array}{c} \begin{array}{ccc} \mathcal{E}_1 & \dots & \mathcal{E}_s \end{array} \\ \left[\begin{array}{ccc} P_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & P_s \end{array} \right] \end{array}, \tag{5.6}$$

where each $\mathcal{E}_i$ is a communicating class and each P_i is irreducible.

5. *$\rho(P_i) = 1$ for $i = 1, \dots, s$, and each P_i has a unique invariant distribution $\mathbf{v}_i \in \mathbb{S}_{n_i}$, where $n_i = |\mathcal{E}_i|$.*

6. *If $\mathcal{I}$ is nonempty, then $\rho(Q) < 1$.*

7. *The matrix A has at least one stationary distribution. The set of all stationary distributions of A is given by*

$$\{[\lambda_1 \mathbf{v}_1 \dots \lambda_s \mathbf{v}_s \ \ \mathbf{0}], \lambda_i \geq 0 \ \forall i, \sum_{i=1}^{s} \lambda_i = 1\}. \tag{5.7}$$

8. *Let $\mathbf{c} \in \mathbb{S}_n$ be an arbitrary initial distribution. If $\mathcal{I}$ is nonempty, permute the components of $\mathbf{c}$ to be compatible with (5.5), and write $\mathbf{c} = [\mathbf{c}_{\mathcal{E}} \ \ \mathbf{c}_{\mathcal{I}}]$. Partition $\mathbf{c}_t = \mathbf{c}A^t$ as $\mathbf{c}_t = [\mathbf{c}_{\mathcal{E}}^{(t)} \ \ \mathbf{c}_{\mathcal{I}}^{(t)}]$. Then $\mathbf{c}_{\mathcal{I}}^{(t)} \to \mathbf{0}$ as $t \to \infty$, irrespective of $\mathbf{c}$.*

As a prelude to the proof, we present a lemma that may be of some interest in its own right.

Lemma 5.7 *Suppose $M \in \mathbb{R}_+^{n \times n}$, and define*

$$\mu(M) := \max_{i \in \mathbb{N}} \sum_{j=1}^{n} m_{ij}.$$

Then

$$\rho(M) = r(M) \leq \mu(M), \tag{5.8}$$

where $r(\cdot)$ is defined in Lemma 4.3 and (4.9).

Proof. The equality of $\rho(M)$ and $r(M)$ follows from Statement (i) of Theorem 4.24. Select an arbitrary vector $\mathbf{x} \in \mathbb{S}_n$. It is shown that $r(\mathbf{x}) \leq \mu(M)$; then (5.8) follows from (4.9), the definition of $r(M)$. So suppose $\mathbf{x} \in \mathbb{S}_n$ is arbitrary, and choose an index $j^* \in \mathbb{N}$ such that

$$x_{j^*} = \max_{j \in \mathbb{N}} x_j.$$

The index j^* need not be unique, but this does not matter. We can choose any one index j^* such that the above holds. Then

$$(\mathbf{x}M)_{j^*} = \sum_{i=1}^{n} x_i m_{ij^*} \leq \sum_{i=1}^{n} x_{j^*} m_{ij^*} \leq \mu(M) x_{j^*}.$$

So

$$r(\mathbf{x}) = \min_{j \in \mathbb{N}} \frac{(\mathbf{x}M)_j}{x_j} \leq \frac{(\mathbf{x}M)_{j^*}}{x_{j^*}} = \mu(M).$$

So the desired conclusion (5.8) follows. □

Proof. (of Theorem 5.6): Everything here follows as almost a routine consequence of the results that have alreadby been proved in Chapter 4.

(1) Let $\mathbf{e}_n$ denote the column vector consisting of all one's. Then the fact that A is stochastic can be expressed as

$$A\mathbf{e}_n = \mathbf{e}_n.$$

This shows that $\lambda = 1$ is an eigenvalue of A, with corresponding column eigenvector $\mathbf{e}_n$. So $\rho(A) \geq 1$. On the other hand, since every row of A sums to one, it follows from Lemma 5.7 that $\mu(A) = 1$, whence $\rho(A) \leq r(A) \leq \mu(A) = 1$. Combining these two observations shows that indeed $\rho(A) = 1$.

(2) The canonical form (5.5) is a ready consequence of Theorem 4.8. In fact (5.5) is the same as (4.4).

(3) Since A is stochastic, no row of A can be identically zero. So the statement follows from Theorem 4.8.

(4) This statement also follows from Theorem 4.8.

(5) Since A is stochastic, it is obvious from (5.5) and (5.6) that each P_i is also stochastic. Hence by Statement 1 it follows that $\rho(P_i) = 1$ for all i. But now we have the additional information that each P_i is irreducible. Hence we conclude from Statement (iv) of Theorem 4.24 that there is a positive eigenvector $\phi_i \in \mathbb{R}_+^{n_i}$ such that $\phi_i P_i = \phi_i$, and that all row eigenvectors of P_i corresponding to the eigenvalue one are multiples of this ϕ_i. Choose $\mathbf{v}_i \in \mathbb{S}_{n_i}$ to be a multiple of ϕ_i. Obviously $\mathbf{v}_i$ is unique.

(6) Suppose $\mathcal{I}$ is nonempty, and let $m = |\mathcal{I}|$ denote the cardinality of $\mathcal{I}$. Partition A^m as

$$A^m = \begin{array}{cc} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} P^m & \mathbf{0} \\ R^{(m)} & Q^m \end{array} \right] \end{array},$$

Then, from the second statement of Lemma 4.4, it follows that each row of $R^{(m)}$ contains a nonzero element. Thus each row sum of Q^m is strictly less than one. Now it follows from Lemma 5.7 that $\rho(Q^m) < 1$. Since $\rho(Q^m) = [\rho(Q)]^m$, it follows that $\rho(Q) < 1$.

(7) Put A in the form (5.5) and look at the equation $\boldsymbol{\pi} A = \boldsymbol{\pi}$. If $\mathcal{I}$ is nonempty, partition $\boldsymbol{\pi}$ as $[\boldsymbol{\pi}_{\mathcal{E}} \;\; \boldsymbol{\pi}_{\mathcal{I}}]$. Then $\boldsymbol{\pi} A = \boldsymbol{\pi}$ becomes

$$[\boldsymbol{\pi}_{\mathcal{E}} \;\; \boldsymbol{\pi}_{\mathcal{I}}] \left[\begin{array}{cc} P & \mathbf{0} \\ R & Q \end{array} \right] = [\boldsymbol{\pi}_{\mathcal{E}} \;\; \boldsymbol{\pi}_{\mathcal{I}}].$$

Expanding this leads to

$$\boldsymbol{\pi}_{\mathcal{E}} P + \boldsymbol{\pi}_{\mathcal{I}} R = \boldsymbol{\pi}_{\mathcal{E}}, \text{ and } \boldsymbol{\pi}_{\mathcal{I}} Q = \boldsymbol{\pi}_{\mathcal{I}}.$$

But since $\rho(Q) < 1$, the matrix $Q - I$ is nonsingular, and so $\boldsymbol{\pi}_{\mathcal{I}} = \mathbf{0}$. This shows that all stationary distributions have zeros in all components corresponding to $\mathcal{I}$. So we seek all solutions to $\boldsymbol{\pi}_{\mathcal{E}} P = \boldsymbol{\pi}_{\mathcal{E}}$. If we now put P in the form (5.6) and partition $\boldsymbol{\pi}$ as $[\boldsymbol{\pi}_1 \ldots \boldsymbol{\pi}_s]$, then *each* $\boldsymbol{\pi}_i$ must satisfy $\boldsymbol{\pi}_i P_i = \boldsymbol{\pi}_i$. Each of these equations has a unique solution $\mathbf{v}_i \in \mathbb{S}_{n_i}$, by Statement 5. Hence the set of all stationary distributions is given by (5.7).

(8) If A is put in the form (5.5) and $\mathcal{I}$ is nonempty, then A^t has the form

$$A^t = \begin{array}{c c} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} P^t & \mathbf{0} \\ R^{(t)} & Q^t \end{array} \right] \end{array} .$$

Now, since $\rho(Q) < 1$, it follows that $Q^t \to \mathbf{0}$ as $t \to \infty$. Hence $\mathbf{c}_t = \mathbf{c}A^t$ implies that

$$\mathbf{c}_{\mathcal{I}}^{(t)} = \mathbf{c}_{\mathcal{I}} Q^t \to \mathbf{0} \text{ as } t \to \infty,$$

irrespective of what $\mathbf{c}_{\mathcal{I}}$ is. This is the desired conclusion. □

In the Markov chain parlance, the states in $\mathcal{I}$ are referred to as **transient states**, and those in $\mathcal{E}$ are referred to as **recurrent states**. Statement 8 of Theorem 5.6 gives the rationale for this nomenclature. Irrespective of the initial distribution of the Markov chain, we have that

$$\Pr\{\mathcal{X}_t \in \mathcal{I}\} \to 0 \text{ as } t \to \infty,$$

while

$$\Pr\{\mathcal{X}_t \in \mathcal{E}\} \to 1 \text{ as } t \to \infty.$$

Within the set $\mathcal{E}$ of recurrent states, the disjoint equivalence classes $\mathcal{E}_1, \ldots, \mathcal{E}_s$ are referred to as the **communicating classes**.

Example 5.2 Let us return to the modified blackjack game of Example 5.1. From the state transition matrix, it is obvious that each of the states 0 through 8 leads to W, but W does not lead to any of these states. (The same statement is also true with W replaced by L.) Thus all of these states are inessential and therefore transient. In contrast, both W and L are essential states, because they do not lead to any other states, and thus vacuously satisfy the condition for being essential. Hence it follows from Theorem 5.6 that

$$\Pr\{\mathcal{X}_t \in \{W, L\}\} \to 1 \text{ as } t \to \infty.$$

Thus all games end with the player either winning or losing; all intermediate states are transient. Within the set of essential states, since W does not lead to L and vice versa, both $\{W\}$ and $\{L\}$ are communicating classes (consisting of singleton sets).

5.2.2 Hitting Probabilities and Mean Hitting Times

Until now we have introduced the notions of recurrent states and transient states, and have shown that eventually all trajectories enter the set of recurrent states. In this subsection we analyze the dynamics of a Markov chain in somewhat greater detail and study the probability that a trajectory will hit a specified subset of the state space, as well as the average time needed to do so.

To motivate these concepts, let us re-examine the modified blackjack game of Example 5.1. It is shown in Example 5.2 that all trajectories will hit either W (win) or L (loss) with probability one. That is a very coarse analysis of the trajectories. It would be desirable to know the *probability of winning (or losing)* from a given starting position. The hitting probability formalizes this notion. Further, it would be desirable to know *the expected number of moves* that would result in a win or loss. The mean hitting time formalizes this notion.

Suppose $\{\mathcal{X}_t\}_{t\geq 0}$ is a Markov process assuming values in a finite state space $\mathbb{N} = \{1, \ldots, n\}$. A subset $S \subseteq \mathbb{N}$ is said to be **absorbing** if $\mathcal{X}_t \in S \Rightarrow \mathcal{X}_{t+1} \in S$ (and by extension, that $\mathcal{X}_{t+k} \in S$ for all $k \geq 1$). Clearly S is absorbing if and only if $a_{ij} = 0$ for all $i \in S, j \notin S$.

Next, suppose $S \subseteq \mathbb{N}$, not necessarily an absorbing set. Define

$$h(S; i, t) := \Pr\{\mathcal{X}_t \in S | \mathcal{X}_0 = i\}.$$

Thus $h(S; i, t)$ is the probability that a trajectory of the Markov process starting at time 0 in state i "hits" the set S at time t. In the same spirit, define

$$\bar{h}(S; i, t) := \Pr\{\exists l, 0 \leq l \leq t, \text{ s.t. } \mathcal{X}_l \in S | \mathcal{X}_0 = i\}. \tag{5.9}$$

Thus $\bar{h}(S; i, t)$ is the probability that a trajectory of the Markov process starting at time 0 in state i "hits" the set S at or before time t. Note that if S is an absorbing set, then $h(S; i, t) = \bar{h}(S; i, t)$. However, if S is not an absorbing set, then the two quantities need not be the same. In particular, $\bar{h}(S; i, t)$ is a nondecreasing function of t, whether or not S is an absorbing set. The same cannot be said of $h(S; i, t)$. Now let us define

$$g(S; i, t) := \Pr\{\mathcal{X}_s \notin S \text{ for } 0 \leq s \leq t-1 \& \mathcal{X}_t \in S | \mathcal{X}_0 = i\}. \tag{5.10}$$

Then $g(S; i, t)$ is the probability that a trajectory of the Markov process starting at time 0 in state i "hits" the set S *for the first time* at time t. From this definition, it is easy to see that

$$\bar{h}(S; i, t) = \bar{h}(S; i, t-1) + g(S; i, t),$$

and as a result

$$\bar{h}(S; i, l) = \sum_{l=0}^{t} g(S; i, l).$$

Now, since the sequence $\{\bar{h}(S; i, t)\}$ is increasing and bounded above by 1 (since every $\bar{h}(S; i, t)$ is a probability), the sequence converges to some limit as $t \to \infty$. This limit is denoted by $\bar{h}(S; i)$ and called the **hitting probability** of the set S given the initial state i. Incidentally, the same argument also shows that $g(S; i, t) \to 0$ as $t \to \infty$. In words, the probability of hitting a set for the first time at time t approaches zero as t becomes larger.

Next, let us define an integer-valued random variable denoted by $\lambda(S;i)$, whereby $\lambda(S;i) = t$ if the trajectory hits S for the first time at t. We can think of $\lambda(S;i)$ as the (random) time to hit the set S starting at time 0 in the state i. From the above discussion, it is clear that $\lambda(S;i)$ equals t with probability $g(S;i,t)$. We must, however, explicitly permit $\lambda(S;i)$ equal infinity in case

$$\sum_{t=0}^{\infty} g(S;i,t) < 1.$$

The **mean hitting time** $\tau(S;i)$ is defined as the expected value of $\lambda(S;i)$; thus

$$\tau(S;i) := \sum_{t=0}^{\infty} t g(S;i,t) + \infty \cdot \left[1 - \sum_{t=0}^{\infty} g(S;i,t) \right]. \tag{5.11}$$

If the quantity inside the square brackets is positive, then the mean hitting time is taken as infinity, by convention. Even if this quantity is zero, the mean hitting time could still be infinite if the hitting time is a heavy-tailed random variable.

In spite of the apparent complexity of the above definitions, there is a very simple explicit characterizations of both the hitting probability and mean hitting time. The derivation of these characterizations is the objective of this subsection.

Theorem 5.8 *The vector* $\bar{\mathbf{h}}(S) := [\bar{h}(S;1) \ldots \bar{h}(S;n)]$ *is the minimal nonnegative solution of the set of equations*

$$v_i = 1 \textit{ if } i \in S, v_i = \sum_{j \in \mathbb{N}} a_{ij} v_j \textit{ if } i \notin S. \tag{5.12}$$

Proof. Here, by a "minimal nonnegative solution," we mean that (i) $\bar{\mathbf{h}}(S)$ satisfies (5.12), and (ii) if $\mathbf{v}$ is any other nonnegative vector that satisfies the same equations, then $\mathbf{v} \geq \bar{\mathbf{h}}(S)$. To prove (i), observe that if $i \in S$, then $h(S;i,0) = 1$, whence $\bar{h}(S;i,t) = \bar{h}(S;i,0) = 1$ for all t. Hence $\bar{h}(S;i)$ being the limit of this sequence also equals one. On the other hand, if $i \notin S$, then

$$h(S;i,t+1) = \sum_{j \in \mathbb{N}} a_{ij} h(S;j,t).$$

This equation states that the probability of hitting S at time $t+1$ starting from the initial state i at time 0 equals the probability of hitting S at time t starting from the initial state j at time 0, weighted by the probability of making the transition from i to j. In writing this equation, we have used both the Markovian nature of the process as well as its stationarity. Now the same reasoning shows that

$$\bar{h}(S;i,t+1) = \sum_{j \in \mathbb{N}} a_{ij} \bar{h}(S;j,t).$$

Letting $t \to \infty$ in both sides of the above equation shows that

$$\bar{h}(S;i) = \sum_{j \in \mathbb{N}} a_{ij} \bar{h}(S;j).$$

Thus the vector of hitting probabilities satisfies (5.12).

To establish the second statement, suppose $\mathbf{v} \in \mathbb{R}^n_+$ is some solution of (5.12). Then $v_i = 1$ for all $i \in S$. For $i \notin S$, we get from (5.12),

$$\begin{aligned} v_i &= \sum_{j \in \mathbb{N}} a_{ij} v_j = \sum_{j \in S} a_{ij} v_j + \sum_{j \notin S} a_{ij} v_j \\ &= \sum_{j \in S} a_{ij} + \sum_{j \notin S} a_{ij} v_j. \end{aligned}$$

Now we can substitute a second time for v_j to get

$$\begin{aligned} v_i &= \sum_{j \in S} a_{ij} + \sum_{j \notin S} a_{ij} v_j \\ &= \sum_{j \in S} a_{ij} + \sum_{j \notin S} \sum_{k \in S} a_{ij} a_{jk} \\ &+ \sum_{j \notin S} \sum_{k \notin S} a_{ij} a_{jk} v_k. \end{aligned}$$

This process can be repeated. If we do this l times, we get

$$\begin{aligned} v_i &= \sum_{j_1 \in S} a_{ij_1} \\ &+ \sum_{j_1 \notin S} \sum_{j_2 \in S} a_{ij_1} a_{j_1 j_2} + \ldots \\ &+ \sum_{j_1 \notin S, \ldots, j_{l-1} \notin S} \sum_{j_l \in S} a_{ij_1} a_{j_1 j_2} \ldots a_{j_{l-1} j_l} \\ &+ \sum_{j_1 \notin S, \ldots, j_{l-1} \notin S} \sum_{j_l \notin S} a_{ij_1} a_{j_1 j_2} \ldots a_{j_{l-1} j_l} v_{j_l}. \end{aligned} \tag{5.13}$$

Next, observe that for each index l, we have

$$\sum_{j_1 \notin S, \ldots, j_{l-1} \notin S} \sum_{j_l \in S} a_{ij_1} a_{j_1 j_2} \ldots a_{j_{l-1} j_l} = g(S;i,l),$$

because the left side is precisely the probability that, starting in state i at time 0, the trajectory hits S for the first time at time l. Moreover, since $i \notin S$, it is clear that $g(S;i,0) = 0$. Finally, from (5.13), it is clear that the last term in the summation is nonnegative, because $\mathbf{v}$ is a nonnegative vector. Thus it follows from (5.13) that

$$v_i \geq \sum_{s=1}^{l} g(S;i,l) = \sum_{s=0}^{l} g(S;i,s) = \bar{h}(S;i,l).$$

Since this is true for each l, letting $l \to \infty$ shows that $v_i \geq \bar{h}_i(S;i)$, as desired. □

In applying (5.12), it is often convenient to rewrite it as

$$\bar{h}(S;i) = 1 \text{ if } i \in S, \bar{h}(S;i) = \frac{1}{1-a_{ii}} \sum_{j \neq i} a_{ij} \bar{h}(S;j) \text{ if } i \not\in S. \tag{5.14}$$

If $a_{ii} < 1$, then the fraction above makes sense. If $a_{ii} = 1$, then a trajectory starting in state i just stays there, so clearly $\bar{h}(S;i) = 0$. Note that the right side of (5.14) is a convex combination of the quantities $\bar{h}(S;j)$ for $j \neq i$, because $\sum_{j \neq i} a_{ij} = 1 - a_{ii}$.

Example 5.3 Let us return to the simplified "blackjack" example of Examples 5.1 and 5.2. We have already seen that every trajectory approaches W or L with probability 1. Now let us compute the probability of winning or losing from a given initial state. To simplify notation, let us write $\bar{h}(W;i)$ to denote $\bar{h}(\{W\};i)$, and define $\bar{h}(L;i)$ in an analogous fashion.

From (5.14), it is obvious that $\bar{h}(W;W) = 1$ and $\bar{h}(L;L) = 1$, because both are absorbing states. For the same reason, we have $\bar{h}(W;L) = 0$ and $\bar{h}(L;W) = 0$. Working backwards, we have from (5.12) that

$$\bar{h}(W;8) = \frac{1}{3} a_{8W} \bar{h}(W;W) + \frac{2}{3} a_{8L} \bar{h}(W;L) = \frac{1}{3} a_{WW} = \frac{1}{3}.$$

Similarly

$$\bar{h}(L;8) = \frac{1}{3} a_{8W} \bar{h}(L;W) + \frac{2}{3} a_{8L} \bar{h}(L;L) = \frac{2}{3} a_{LL} = \frac{2}{3}.$$

It is hardly surprising that $\bar{h}(W;8) + \bar{h}(L;8) = 1$, because these are the only two absorbing states. In fact it is true that $\bar{h}(W;i) + \bar{h}(L;i) = 1$ for every initial state i. Next,

$$\bar{h}(W;7) = \frac{1}{3} \bar{h}(W;8) + \frac{1}{3} \bar{h}(W;W) + \frac{1}{3} \bar{h}(W;L) = \frac{4}{9}.$$

Proceeding in this manner, we get the table below.

i	$\bar{h}(W;i)$	$\bar{h}(L;i)$
8	1/3	2/3
7	4/9	5/9
6	16/27	11/27
5	37/81	44/81
4	121/243	122/243
3	376/729	353/729
2	1072/2187	1118/2187
1	3289/6561	3272/6561
0	9889/19683	9794/19683

Thus we see that the probability of winning from a particular starting state goes up and down like a yo-yo. The states 4 and 1 are closest to being "neutral" in that the odds of winning and losing are roughly equal, while the state 6 offers the best prospects for winning. This is not surprising, because from the initial state 6 one cannot possibly lose in one time step, but winning in one time step is possible.

Thus far we have analyzed the hitting probability $\bar{h}(S;i)$. Next we analyze the mean hitting $\tau(S;i)$ defined in (5.11). It turns out that $\tau(S;i)$ also satisfies a simple linear recursion analogous to (5.12).

Theorem 5.9 *Suppose that the hitting time $\tau(S;i)$ is finite for all $i \in \mathbb{N}$. Then the vector $\tau(S)$ of mean hitting times $[\tau(S;1)\ldots\tau(S;n)]$ is the minimal nonnegative solution of the equations*

$$\alpha_i = 0 \textit{ if } i \in S, \alpha_i = 1 + \sum_{j \notin S} a_{ij}\alpha_j \textit{ if } i \notin S. \tag{5.15}$$

Proof. As in Theorem 5.8, we need to establish two statements: First, the vector of mean hitting times satisfies (5.15). Second, if α is any other solution of (5.15), then $\alpha_i \geq \tau(S;i)$ for all i.

To prove the first statement, suppose first that $i \in S$. Then clearly $\tau(S;i) = 0$. Next, suppose $i \notin S$, and consider all possible next states j; the transition from i to j occurs with probability a_{ij}. With $\mathcal{X}_0 = i$ and $\mathcal{X}_1 = j$, we have

$$\begin{aligned}
&\Pr\{\mathcal{X}_l \notin S \text{ for } 0 \leq l \leq t-1 \text{ and } \mathcal{X}_t \in S|\mathcal{X}_0 = i \& \mathcal{X}_1 = j\} \\
&\qquad = \Pr\{\mathcal{X}_l \notin S \text{ for } 1 \leq l \leq t-1 \text{ and } \mathcal{X}_t \in S|\mathcal{X}_1 = j\} \\
&\qquad = \Pr\{\mathcal{X}_l \notin S \text{ for } 0 \leq l \leq t-2 \text{ and } \mathcal{X}_t \in S|\mathcal{X}_0 = j\}.
\end{aligned}$$

Here the first equation follows from the Markovian nature of the process $\{\mathcal{X}_t\}$, because the behavior of $\mathcal{X}_l$ for $l \geq 2$ depends only on $\mathcal{X}_1$ and is independent of $\mathcal{X}_0$. The second equation follows from the stationarity of the process. Hence, if $\mathcal{X}_0 = i \notin S$, we can distinguish between two possibilities: If $\mathcal{X}_1 = j \in S$, then the mean hitting time from then onwards is zero, and the mean hitting time from the start is 1. If $\mathcal{X}_1 = j \notin S$, then the mean hitting time from then onwards is $\tau(S;j)$, and the mean hitting time from the beginning is $1 + \tau(S;j)$. We can average over all of these events to get

$$\begin{aligned}
\tau(S;i) &= \sum_{j \in S} a_{ij} + \sum_{j \notin S} a_{ij}[1 + \tau(S;j)] \\
&= 1 + \sum_{j \notin S} a_{ij}\tau(S;j),
\end{aligned}$$

because

$$\sum_{j \in S} a_{ij} + \sum_{j \notin S} a_{ij} = \sum_{j \in \mathbb{N}} a_{ij} = 1.$$

Therefore the vector $\tau(S)$ of mean hitting times satisfies (5.15).

To prove the second statement, suppose α is any nonnegative solution of

(5.15). Then $\alpha_i = 0$ for $i \in S$. For $i \notin S$, it follows from (5.15) that

$$\begin{aligned} \alpha_i &= 1 + \sum_{j \notin S} a_{ij} \alpha_j \\ &= 1 + \sum_{j \notin S} a_{ij} \left[1 + \sum_{k \notin S} a_{jk} \alpha_k \right] \\ &= 1 + \sum_{j \notin S} a_{ij} + \sum_{j \notin S} \sum_{k \notin S} a_{ij} a_{jk} \alpha_k. \end{aligned} \tag{5.16}$$

This process can be repeated by substituting for α_k. If we do this l times, we get

$$\begin{aligned} \alpha_i &= 1 + \sum_{j_1 \notin S} a_{ij_1} + \sum_{j_1, j_2 \notin S} a_{ij_1} a_{j_1 j_2} + \dots \\ &+ \sum_{j_1, \dots, j_{l-1} \notin S} a_{ij_1} a_{j_1 j_2} \cdots a_{j_{l-2} j_{l-1}} \\ &+ \sum_{j_1, \dots, j_l \notin S} a_{ij_1} a_{j_1 j_2} \cdots a_{j_{l-1} j_l} \alpha_{j_l}. \end{aligned} \tag{5.17}$$

Now the last term is nonnegative since $\alpha \geq \mathbf{0}$. As for the remaining terms, since $i \notin S$ it is clear that the hitting time $\lambda(S; i) \geq 1$ with probability one. Thus

$$\Pr\{\lambda(S; i) \geq 1\} = 1.$$

More generally, it is easy to see that

$$\begin{aligned} \sum_{j_1, \dots, j_{l-1} \notin S} a_{ij_1} a_{j_1 j_2} \cdots a_{j_{l-2} j_{l-1}} &= \Pr\{\mathcal{X}_s \notin S \text{ for } 0 \leq s \leq l-1 | \mathcal{X}_0 = i\} \\ &= \Pr\{\lambda(S; i) \geq l\}, \end{aligned}$$

because the left side is the probability that the trajectory starting at $i \notin S$ at time 0 does not hit the set S during the first $l-1$ transitions. Hence, after neglecting the last term on the right side of (5.17) because it is nonnegative, we get

$$\alpha_i \geq \sum_{k=1}^{l} \Pr\{\lambda(S; i) \geq k\}. \tag{5.18}$$

To conclude the proof, we observe that, as a consequence of (5.10), we have

$$\Pr\{\lambda(S; i) \geq k\} = \sum_{t=k}^{\infty} \Pr\{\lambda(S; i) = t\} = \sum_{t=k}^{\infty} g(S; i, t). \tag{5.19}$$

We need not worry about the convergence of the infinite summation because $\Pr\{\lambda(S; i) \geq k\} \leq 1$ and so the sum converges. Substituting from (5.19) into (5.18) shows that

$$\alpha_i \geq \sum_{k=1}^{l} \sum_{t=k}^{\infty} g(S; i, t) \geq \sum_{k=1}^{l} \sum_{t=k}^{l} g(S; i, t).$$

Now a simple counting argument shows that

$$\sum_{k=1}^{l} \sum_{t=k}^{l} g(S; i, t) = \sum_{t=1}^{l} t g(S; i, t). \tag{5.20}$$

An easy way to see this is to consider the triangular matrix below, where g_t is shorthand for $g(S; i, t)$.

$$\begin{bmatrix} g_1 & g_2 & \dots & g_{l-1} & g_l \\ 0 & g_2 & \dots & g_{l-1} & g_l \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & g_{l-1} & g_l \\ 0 & 0 & \dots & 0 & g_l \end{bmatrix}$$

The left side of (5.20) is obtained by first summing each row and then adding the row sums; the right side of (5.20) is obtained by first summing each column and then adding the column sums. Clearly both procedures give the same answer, which is what (5.20) says. Let us substitute from (5.19) and (5.20) into (5.18). This gives

$$\alpha_i \geq \sum_{k=1}^{l} t g(S; i, t), \ \forall l.$$

Now letting $l \to \infty$ shows that $\alpha_i \geq \tau(S; i) \ \forall i$, which is the desired statement. □

Example 5.4 Let us again return to the simplified blackjack game and compute the expected duration of the game starting from each initial state. Let us define $E = \{W, L\}$, so that $\mathcal{X}_t \in E$ suggests that the game has ended. We can apply (5.15) to compute the mean hitting times. It is again convenient to rewrite (5.15) as

$$\tau(S; i) = \frac{1}{1 - a_{ii}} \left[1 + \sum_{j \notin S, j \neq i} a_{ij} \tau(S; j) \right].$$

It is clear that $\tau(E; W) = \tau(E; L) = 0$, as both W and L belong to E. If $\mathcal{X}_0 = 8$, then

$$\tau(E; 8) = \frac{4}{3},$$

since the summation on the right side is empty and is thus taken as zero. Next,

$$\tau(E; 7) = \frac{4}{3}(1 + a_{78}\tau(E; 8)) = \frac{4}{3}(1 + 1/3) = \frac{16}{9}.$$

Proceeding in this manner, we get the table below.

i	$\tau(E;i)$	$\approx$
8	4/3	1.333
7	16/9	1.778
6	64/27	2.370
5	256/81	3.160
4	916/243	3.700
3	3232/729	4.433
2	11200/2187	5.121
1	37888/6561	5.775
0	126820/19683	6.443

5.3 ERGODICITY OF MARKOV CHAINS

5.3.1 Motivation

Suppose $\{\mathcal{X}_t\}_{t=0}^{\infty}$ is a stationary stochastic process assuming values in a finite set $X = \{x_1, \ldots, x_n\}$. We have not given a precise definition of stationarity, but we do know one important consequence of stationarity: Every one of the random variables $\mathcal{X}_t$ has exactly the same distribution. Let $\boldsymbol{\pi}$ denote this distribution, so that

$$p_i = \Pr\{\mathcal{X}_t = x_i\}, i = 1, \ldots, n.$$

Thus $\boldsymbol{\pi}$ is the same for all values of t because of stationarity. Now suppose f is a function mapping the set X into the real numbers $\mathbb{R}$. Then we can think of $f(\mathcal{X})$ itself as a random variable, as discussed in Section 2.1. Moreover, since $f(\mathcal{X})$ is a *real-valued* random variable, it is possible to define the expected value of $f(\mathcal{X})$ as

$$E[f(\mathcal{X}), P_\pi] = \sum_{i=1}^{n} f(x_i)\pi_i. \tag{5.21}$$

As discussed in Section 2.1, the above formula is valid even if the numbers $f(x_1), \ldots, f(x_n)$ are not necessarily distinct.

Computing the expected value of $f(\mathcal{X})$ using (5.21) is straight-forward *provided* one knows the values π_i, that is, provided one knows the probability distribution of the random variable $\mathcal{X}_t$. However, there are many situations in which the statistics of the process under study are not known. For example, suppose someone gives us a coin, and tells us that there is a payout of 1 (in some currency) whenever the coin turns up 'heads' (H) and a payback of -1 whenever the coin turns up 'tails' (T). Before entering into the game, we would like to know the expected payout

$$G = f(H)\pi_H + f(T)\pi_T = \pi_H - \pi_T$$

in this case. However, since we are seeing the coin for the first time, we have no way of knowing the values p_H and p_T. So we toss the coin a number of times, and generate a sequence of outcomes, in the form $HTHHTHTTHT \cdots$. This sequence of outcomes is called a 'sample path' or a 'realization' of the stochastic process $\{\mathcal{X}_t\}$, where $\mathcal{X}_t$ is the outcome of the coin toss at time t. We can compute the fraction of heads in the sample path and call it $\hat{\pi}_H$. The hat over π is to remind ourselves that $\hat{\pi}_H$ is only an *approximation* to the true but unknown value π_H. The number $\hat{\pi}_H$ is called the **empirical probability** of heads. It is itself a random variable, because if we repeat the experiment, we will in general get a different value for $\hat{\pi}_H$. The number $\hat{\pi}_T$ is defined analogously. Then we can *compute*

$$\hat{G} = f(H)\hat{\pi}_H + f(T)\hat{\pi}_T$$

based on the sample path. But how close is $\hat{G}$ to G?

More generally, suppose $\{X_t\}$ is a stationary stochastic process that assumes values in a set $\mathbb{N} = \{x_1, \ldots, x_n\}$. Then we can generate a sample path $u_0 u_1 \ldots u_{T-1}$ of the process up to time $T-1$, and compute the empirical probabilities.

$$\hat{\pi}_i = \frac{1}{T} \sum_{t=0}^{T-1} I_i(u_t),$$

where $I_i(\cdot)$ is the 'indicator function'

$$I_i(u) = \begin{cases} 1, & \text{if} \quad u = x_i, \\ 0, & \text{if} \quad u \neq x_i. \end{cases}$$

So $\hat{\pi}_i$ is just the fraction of times that the symbol x_i occurs in the sample path. Then we can compute the quantity

$$\hat{E}(f) := \sum_{i=1}^{n} f(x_i)\hat{\pi}_i = \frac{1}{T} \sum_{t=0}^{T-1} f(u_t). \tag{5.22}$$

Again, we can ask: How close is $\hat{E}(f)$ to $E(f)$? In the sequel, we refer to $E(f)$ as the **true mean** and to $\hat{E}(f)$ as the **empirical mean**.

The simplest situation to visualize is the case where $\{\mathcal{X}_t\}$ is an i.i.d. (independent, identically distributed) process where each $\mathcal{X}_t$ has the distribution $\boldsymbol{\pi}$. In this case, each $f(\mathcal{X}_t)$ is a random variable that is independent of each $f(\mathcal{X}_\tau)$ whenever $t \neq \tau$. Therefore the samples $\{f(\mathcal{X}_0), f(\mathcal{X}_1), \ldots, f(\mathcal{X}_{T-1})\}$ can be thought of as an 'ensemble' of identical replicas of the single random variable $f(\mathcal{X})$, and of $\hat{E}(f)$ as an 'ensemble average.' In this situation, it is natural to expect that the empirical mean $\hat{E}(f)$ 'converges' in some appropriate sense to the true mean $E(f)$. A formal statement and proof of this result are given in Chapter 6.

A variation of this result is the case where the $\mathcal{X}_t$'s are 'nearly independent,' and where a suitably defined 'index of dependence' between $\mathcal{X}_t$ and $\mathcal{X}_\tau$ approaches zero as the difference $|t - \tau|$ approaches infinity. The notion of 'mixing,' discussed in Chapter 6, is a formalization of this idea. It can be shown that even when the stochastic process $\{\mathcal{X}_t\}$ is merely 'mixing' instead of being i.i.d., the empirical mean $\hat{E}(f)$ still converges to the true mean

$E(f)$, though possibly at a slower rate than when the process $\{\mathcal{X}_t\}$ is i.i.d. This is also shown in Chapter 6.

The topic of the present section is 'ergodicity,' which is a still weaker property than mixing. Thanks to seminal work in probability theory in the 1930's, it is known that ergodicity is sufficient for the empirical mean $\hat{E}(f)$ to converge to the true mean $E(f)$; mixing is not required. Accordingly in this section we study the ergodicity properties of homogeneous Markov chains.

One last comment before we conclude the discussion of the motivation for studying ergodicity. If the state space of a Markov process is uncountably infinite, then the probability *distribution* $\boldsymbol{\pi}$ gets replaced by an appropriate probability *measure*. In computing the expected value of a function, the *summation* in (5.21) gets replaced by an *integral* with respect to the associated probability measure. In such a general situation, *even when the underlying probability measure is known*, it may be messy or intractable to compute the expected value using this integral formulation. In contrast, even when the state space of a Markov process is an uncountably infinite set, the formula (5.22) for the empirical mean $\hat{E}(f)$ still involves only averaging a finite number of real-valued measurements $f(u_t)$. So the message is that the formula (5.22) for generating an approximation to $E(f)$ can be valuable even when the statistical properties of the process $\{\mathcal{X}_t\}$ are known precisely. Indeed, well-known methods such as Monte Carlo simulation are based on precisely this principle.

Finally, the reader is cautioned that the words 'ergodic' and 'ergodicity' are used by different authors to mean different properties. This problem seems to arise only in the Markov chain literature. Both mixing and ergodicity are properties that can be defined for *arbitrary* stationary stochastic processes, not just homegeneous Markov chains, and not just stochastic processes assuming values in a finite set. In this very general setting, there is only one definition. In the 'pure' stochastic process literature, a process is defined to be ergodic if, in some precisely defined sense, it explores all parts of the state space. Ergodicity is actually a very weak property. Yet it is sufficient for the empirical mean $\hat{E}(f)$ to converge to the true mean $E(f)$. The proof of the ergodic theorem in the 1930's was a remarkable achievement. The definition of ergodicity used in the present book is consistent with the more general definition; that is, a Markov process over a finite state space is ergodic in the sense defined here if and only if it is ergodic in the general sense used in the stochastic process literature. Unfortunately in the Markov process literature, the adjective 'ergodic' is often used to describe a much stronger property. Thus the reader should verify the definitions in a particular source before comparing across sources.

5.3.2 Ergodicity of a Stationary Markov Process

In this subsection, the focus is on the following question: Suppose $A \in \mathbb{R}_+^{n \times n}$ is a stochastic matrix, and we define

$$\bar{A}(T) := \frac{1}{T} \sum_{t=0}^{T-1} A^t.$$

We refer to $\bar{A}(T)$ as the 'ergodic average' of A over T time instants. The question of interest here is: Does $\bar{A}(T)$ have a limit as $T \to \infty$, and if so, what does the limit look like? It is shown in the sequel that the limit *does* exist for *every* stochastic matrix, and an explicit formula for the limit is derived.

The motivation for studying the question is fairly obvious. Suppose $\{\mathcal{X}_t\}$ is a homogeneous Markov chain with the state space $\mathbb{N}$ and transition matrix A. Suppose $f : \mathbb{N} \to \mathbb{R}$ is a given function, and that $\boldsymbol{\pi}$ is a given stationary distribution of A. (Recall that there could be more than one stationary distribution of A.) We wish to compute the expected value of the random variable $f(\mathcal{X})$ with respect to the stationary distribution $\boldsymbol{\pi}$. But actually we may not know either A or $\boldsymbol{\pi}$. So for this purpose we start off the Markov chain with some initial distribution $\mathbf{c}_0 = \mathbf{c}$. The underlying presumption is that we have a way of generating a sample path $\{u_0, u_1, \ldots\}$ where the initial state is distributed according to $\mathbf{c}$ (which we choose), and subsequent state transitions are according to the possibly unknown matrix A. Then $\mathcal{X}_t$ is distributed according to $\mathbf{c}_t = \mathbf{c}A^t$, and the corresponding random variable $f(\mathcal{X})$ also has the same distribution. Thus, if we form the ergodic average $\hat{E}(f)$ according to (5.22), then we can think of $\hat{E}(f)$ as a random variable on the set $\mathbb{N}$, whose underlying distribution is

$$\bar{\mathbf{c}}_T := \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{c}_t = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{c}A^t.$$

Hence, if the limit

$$\bar{A} := \lim_{T \to \infty} \bar{A}_T = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} A^t \tag{5.23}$$

exists, then the limit distribution of the random variable $\hat{E}(f)$ will equal $\mathbf{c}\bar{A}$.

Finally then, suppose it is the case that $\mathbf{c}\bar{A}$ equals the stationary distribution $\boldsymbol{\pi}$ for *every* initial distribution $\mathbf{c}$. If this were to be so, then the ergodic average $\hat{E}(f)$ would converge (in some sense not made very precise) to the 'correct' or true mean $E[f(\mathcal{X})]$ where the expectation is taken with respect to the distribution $\boldsymbol{\pi}$. This then is the basis of a *mathematical treatment* of ergodic theory. In order for this approach to work, the desired property is that $\mathbf{c}\bar{A} = \boldsymbol{\pi}$ for *every* initial distribution $\mathbf{c} \in \mathbb{S}_n$. In general this property will not hold, so we make a slight modification. We assume that, while the stationary distribution $\boldsymbol{\pi}$ may not be known, at least we know which components of $\boldsymbol{\pi}$ are zero and which are not. To put it another way, we know at least the index set $S = \{i : \pi_i = 0\}$. Then we ensure that the initial distribution $\mathbf{c}$ has zeros in the indices belonging to S; in other words, we ensure that $\pi_i = 0 \Rightarrow c_i = 0$, or in the language of Chapter 3, that $\mathbf{c} \ll \boldsymbol{\pi}$ ($\mathbf{c}$ is dominated by $\boldsymbol{\pi}$). So now we can ask a slightly restricted question: Is it the case that $\mathbf{c}\bar{A} = \boldsymbol{\pi}$ *for every* $\mathbf{c} \in \mathbb{S}_n$ *such that* $\mathbf{c} \ll \boldsymbol{\pi}$? This question is quite tractable, and we give a complete answer in this subsection.

Now we state the main result of this subsection.

Theorem 5.10 *Suppose $A \in \mathbb{R}_+^{n\times n}$ is a stochastic matrix, and assume without loss of generality that A is in the canonical form (5.5). Define the eigenvectors $\mathbf{v}_i \in \mathbb{S}_{n_i}$ as in Statement 5 of Theorem 5.6, and let $\mathbf{e}_{n_i}$ denote the column vector consisting of n_i one's. Then the ergodic limit $\bar{A}$ defined in (5.23) exists and is given by*

$$\bar{A} = \begin{array}{c c} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} \bar{P} & \mathbf{0} \\ \bar{R} & \mathbf{0} \end{array} \right] \end{array}, \tag{5.24}$$

where

$$\bar{P} = \begin{array}{c c} & \begin{array}{ccc} \mathcal{E}_1 & \ldots & \mathcal{E}_s \end{array} \\ \begin{array}{c} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_s \end{array} & \left[\begin{array}{ccc} \mathbf{e}_{n_1}\mathbf{v}_1 & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{e}_{n_s}\mathbf{v}_s \end{array} \right] \end{array}, \tag{5.25}$$

$$\bar{R} = (I - Q)^{-1} R \bar{P}. \tag{5.26}$$

The proof of the theorem proceeds via a series of lemmas. But first we give a few explanatory notes.

1. The matrix $\bar{P}$ in (5.25) is block-diagonal, and the number of blocks equals the number of communicating classes. Each of the diagonal block is a rank one matrix (namely $\mathbf{e}_{n_i}\mathbf{v}_{n_i}$).

2. It is easy to construct examples where A^t by itself does not converge to anything. For instance, let

$$A = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right].$$

Then $A^2 = I$, and as a result $A^{2k} = I$ and $A^{2k+1} = A$. So A^t does not have a limit as $t \to \infty$. On the other hand,

$$\bar{A}_{2k} = \frac{1}{2}(A + I), \bar{A}_{2k+1} = \frac{k}{2k+1}(A + I) + \frac{1}{2k+1}A.$$

So the ergodic limit exists and equals

$$\bar{A} = \frac{1}{2}(A + I) = \left[\begin{array}{cc} 0.5 & 0.5 \\ 0.5 & 0.5 \end{array} \right] = \left[\begin{array}{c} 1 \\ 1 \end{array} \right] [0.5 \ \ 0.5].$$

This is consistent with (5.25), since A is irreducible ($s = 1$) and its unique stationary distribution is $[0.5 \ \ 0.5]$.

Lemma 5.11 *Suppose $A \in \mathbb{R}_+^{n \times n}$ is irreducible and stochastic, and let $\mathbf{v} \in \mathbb{S}_n$ denote its stationary distribution. Then the ergodic limit $\bar{A}$ exists and equals $\mathbf{e}_n \mathbf{v}$.*

Proof. Let p denote the period of the matrix A. Then we know from Theorem 4.24 that A has exactly p eigenvalues on the unit circle, namely

$$\lambda_0 = 1, \lambda_k = \exp(\mathbf{i} w \pi k/p), k = 1, \ldots, p-1,$$

and the remaining $n-p$ eigenvalues of A all have magnitude strictly less than one. Let $\mathbf{v}_0 = \mathbf{v}, \mathbf{v}_1, \ldots, \mathbf{v}_p$ denote row eigenvectors of A corresponding to the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_p$, and choose other row vectors $\mathbf{v}_{p+1}, \ldots, \mathbf{v}_n$ in such a way that

$$VAV^{-1} = \begin{array}{cc} & \begin{array}{cc} p & n-p \end{array} \\ \begin{array}{c} p \\ n-p \end{array} & \left[\begin{array}{cc} \Lambda & \mathbf{0} \\ \mathbf{0} & S \end{array} \right] \end{array} , \tag{5.27}$$

where $\Lambda = \text{Diag}\{\lambda_0, \ldots, \lambda_{p-1}\}$, and $\rho(S) < 1$. Now note that $\lambda_1, \ldots, \lambda_p$ are all roots of one, and therefore $\lambda_k^p - 1 = 0$ for $k = 1, \ldots, p-1$. Moreover, since

$$\lambda^p - 1 = (\lambda - 1) \left(\sum_{i=0}^{p-1} \lambda^i \right),$$

it follows that

$$\sum_{i=0}^{p-1} \lambda_k^i = 0, k = 1, \ldots, p-1.$$

So for every integer l, we have that

$$\sum_{t=lp}^{(l+1)p-1} \Lambda^t = \begin{array}{cc} & \begin{array}{cc} 1 & p-1 \end{array} \\ \begin{array}{c} 1 \\ p-1 \end{array} & \left[\begin{array}{cc} l & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right] \end{array} .$$

Moreover, since $\rho(S) < 1$, it follows that $S^t \to \mathbf{0}$ as $t \to \infty$. So

$$\bar{A}_T = \frac{1}{T} \sum_{t=0}^{T-1} A^t = V^{-1} \bar{Z}_T V + O(1/T),$$

where

$$\bar{Z}_T = \begin{array}{cc} & \begin{array}{ccc} 1 & p-1 & n-p \end{array} \\ \begin{array}{c} 1 \\ p-1 \\ n-p \end{array} & \left[\begin{array}{ccc} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array} \right] \end{array} .$$

So the ergodic limit exists and equals $\mathbf{w}_1 \mathbf{v}$, where $\mathbf{w}_1$ is the first column of V^{-1}. Hence the proof is complete once it is shown that $\mathbf{w}_1 = \mathbf{e}_n$. For this purpose, note that since A is stochastic, its rows all add up to one; that is, $A\mathbf{e}_n = \mathbf{e}_n$. Now if we let U denote the matrix on the right side of (5.27), define $W = V^{-1}$, and rewrite (5.27) as $AW = WU$, then the first column of $AW = WU$ says that $A\mathbf{w}_1 = \mathbf{w}_1$. Since $\lambda_0 = 1$ is a simple eigenvalue of A, it follows that $\mathbf{w}_1 = \alpha\mathbf{e}_n$ for some proportionality constant α. To determine this constant, observe that $VW = I_n$ implies that $\mathbf{v}\mathbf{w}_1 = 1$. But since $\mathbf{v}\mathbf{e}_n = 1$ (because $\mathbf{v} \in \mathbb{S}_n$), the constant α equals one, and $\mathbf{w}_1 = eb_1$. □

Corollary 5.12 *Suppose $A \in \mathbb{R}_+^{n\times n}$ is primitive and stochastic, and let $\mathbf{v} \in \mathbb{S}_n$ denote its stationary distribution. Then*

$$A^t \to \mathbf{e}_n\mathbf{v} \text{ as } t \to \infty. \tag{5.28}$$

Proof. Since A is primitive, it is irreducible and also aperiodic. So from Theorem 4.21, we know that all eigenvalues of A have magnitude less than one except for $\lambda_0 = 1$. So we can choose

$$V = \left[\begin{array}{c} \mathbf{v} \\ V_2 \end{array}\right], W = V^{-1} = [\mathbf{e}_n | W_2]$$

such that

$$VAV^{-1} = \begin{array}{cc} & \begin{array}{cc} 1 & n-1 \end{array} \\ \begin{array}{c} 1 \\ n-1 \end{array} & \left[\begin{array}{cc} 1 & \mathbf{0} \\ \mathbf{0} & S \end{array}\right] \end{array} =: U, \text{ say,}$$

where $\rho(S) < 1$. So

$$A^t = V^{-1}U^tV = [\mathbf{e}_n|W_2]\left[\begin{array}{cc} 1 & \mathbf{0} \\ \mathbf{0} & S^t \end{array}\right]\left[\begin{array}{c} \mathbf{v} \\ V_2 \end{array}\right] \to \mathbf{e}_n\mathbf{v} \text{ as } t \to \infty,$$

after noting that $S^t \to 0$ as $t \to \infty$ because $\rho(S) < 1$. □

Proof. (of Theorem 5.10): Suppose A is in the canonical form (5.5) and (5.6). Then

$$A^t = \begin{array}{cc} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} P^t & \mathbf{0} \\ R^{(t)} & Q^t \end{array}\right] \end{array},$$

where $R^{(t)}$ can be computed recursively from partitioning $A^t = AA^{t-1}$, as follows:

$$R(1) = R, R^{(t)} = RP^{t-1} + QR^{(t-1)}, \text{ for } t \geq 2. \tag{5.29}$$

So the ergodic average $\bar{A}_T$ is given by

$$\bar{A}_T = \frac{1}{T}\sum_{t=0}^{T-1} A^t = \begin{array}{cc} & \begin{array}{cc} \mathcal{E} & \mathcal{I} \end{array} \\ \begin{array}{c} \mathcal{E} \\ \mathcal{I} \end{array} & \left[\begin{array}{cc} \bar{P}_T & \mathbf{0} \\ \bar{R}_T & \bar{Q}_T \end{array}\right] \end{array},$$

where $\bar{P}_T, \bar{R}_T, \bar{Q}_T$ are defined in the obvious fashion. The ergodic limit of each matrix is now analyzed separately.

First, since we have from (5.6) that

$$P = \text{Block Diag } \{P_1, \ldots, P_s\},$$

it is clear that

$$\bar{P}_T = \text{Block Diag } \{(P_1)_T, \ldots, (P_s)_T\}.$$

Next, observe that each P_i is irreducible. So it follows from Lemma 5.7 that each P_i has an ergodic limit, which equals $\mathbf{e}_{n_i}\mathbf{v}_i$. So the ergodic limit of P is given by (5.25).

Next, note that $\rho(Q) < 1$ from Statement 6 of Theorem 5.6. So $Q^t \to \mathbf{0}$ as $t \to \infty$, whence $\bar{Q} = \mathbf{0}$.

Now it remains only to compute $\bar{R}$. Since we take $A^0 = I_n$, and the lower triangular block of I_n is $\mathbf{0}$, we can take $R^{(0)} = \mathbf{0}$. With this convention, the formula (5.29) is consistent even for $t = 1$. So

$$\begin{aligned}
\bar{R}_T &= \frac{1}{T}\sum_{t=0}^{T-1} R(t) \\
&= \frac{1}{T}\sum_{t=1}^{T-1} R(t) \text{ since } R^{(0)} = \mathbf{0} \\
&= \frac{1}{T}\sum_{t=1}^{T-1} \{RP^{t-1} + QR^{(t-1)}\} \\
&= \frac{1}{T} R \sum_{t=0}^{T-2} P^t + \frac{1}{T} Q \sum_{t=0}^{T-2} R(t) \\
&= \frac{T-1}{T} R\bar{P}_{T-1} + \frac{T-1}{T} Q\bar{R}_{T-1}.
\end{aligned}$$

So $\bar{R}_T$ satisfies this time-varying recursive equation. However, its limit behaviour is easy to analyze. As $T \to \infty$, the constant $(T-1)/T \to 1$, the matrix $\bar{P}_{T-1}$ approaches the ergodic limit $\bar{P}$, and both $\bar{R}_T$ and $\bar{R}_{T-1}$ approach a constant matrix $\bar{R}$. Thus taking the limit as $T \to \infty$ in the preceding equation shows that $\bar{R}$ satisfies

$$\bar{R} = R\bar{P} + Q\bar{R}.$$

Note that $\rho(Q) < 1$ so that $I - Q$ is nonsingular. So the solution of the above equation for $\bar{R}$ is given by (5.26). □

Suppose A is a stochastic matrix and $\boldsymbol{\pi}$ is a stationary distribution of A. We refer to $(\boldsymbol{\pi}, A)$ as a **Markovian pair**.

Definition 5.13 *Suppose $(\boldsymbol{\pi}, A)$ is a Markovian pair. Then $(\boldsymbol{\pi}, A)$ is said to be* **ergodic** *if*

$$\mathbf{c}\bar{A} = \boldsymbol{\pi} \text{ whenever } \mathbf{c} \in \mathbb{S}_n \text{ and } \mathbf{c} \ll \boldsymbol{\pi}. \tag{5.30}$$

The motivation for the above definition has already been given at the beginning of the preceding subsection. Note that (5.30) can be expressed equivalently as

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{c}A^t = \boldsymbol{\pi} \ \forall \mathbf{c} \in \mathbb{S}_n \text{ such that } \mathbf{c} \ll \boldsymbol{\pi}. \tag{5.31}$$

Theorem 5.14 *Suppose $(\boldsymbol{\pi}, A)$ is a Markovian pair, and let $\mathcal{C}_1, \ldots, \mathcal{C}_s$ denote the communicating classes of the matrix A. Then $(\boldsymbol{\pi}, A)$ is ergodic if and only if $\boldsymbol{\pi}$ is concentrated on exactly one of the communicating classes.*

Proof. From Theorem 5.6 and (5.7) we know that $\boldsymbol{\pi}$ must have the form

$$\boldsymbol{\pi} = [\lambda_1 \mathbf{v}_1 \ldots \lambda_s \mathbf{v}_s \ \ \mathbf{0}], \text{ where } [\lambda_1 \ldots \lambda_s] \in \mathbb{S}_s.$$

The theorem says that the pair $(\boldsymbol{\pi}, A)$ is ergodic if and only if all the λ_i's are zero except for one.

"If": Suppose only one of the λ_i's is nonzero and the rest are zero. Renumber the communicating classes such that $\lambda_1 = 1$ and $\lambda_i = 0$ for $i \geq 2$. Thus

$$\boldsymbol{\pi} = [\mathbf{v}_1 \ \ \mathbf{0} \ldots \mathbf{0} \ \ \mathbf{0}] = [\mathbf{v}_1 \ \ \mathbf{0}],$$

where we have aggregated all the zero vectors into one. Suppose now $\mathbf{c} \in \mathbb{S}_n$ is arbitrary except that $\mathbf{c} \ll \boldsymbol{\pi}$. Then $\mathbf{c}$ is also concentrated only on the class $\mathcal{C}_1$. Thus $\mathbf{c}$ has the form

$$\mathbf{c} = [\mathbf{c}_1 \ \ \mathbf{0}],$$

where $\mathbf{c}_1 \in \mathbb{S}_{n_1}$ and n_1 is the number of states in the communicating class $\mathcal{C}_1$. Now, from (5.24) and (5.25), we get

$$\mathbf{c}\bar{A} = [\mathbf{c}_1 \mathbf{e}_{n_1} \mathbf{v}_1 \ \ \mathbf{0}] = [\mathbf{v}_1 \ \ \mathbf{0}] = \boldsymbol{\pi},$$

since $\mathbf{c}_1 \mathbf{e}_{n_1} = 1$ by virtue of the fact that $\mathbf{c}_1 \in \mathbb{S}_{n_1}$.

"Only If": Suppose $\boldsymbol{\pi}$ has the form (5.7) and that at least two of the λ_i's are nonzero. Renumber the communcating classes such that $\lambda_1 \neq 0$. Then $\boldsymbol{\pi}$ has the form

$$\boldsymbol{\pi} = [\lambda_1 \mathbf{v}_1 \ \ \boldsymbol{\pi}_2 \ \ \mathbf{0}],$$

where $\boldsymbol{\pi}_2 \neq \mathbf{0}$ because it contains the nonzero subvector $\lambda_2 \mathbf{v}_2$. Now choose

$$\mathbf{c} = [\mathbf{v}_1 \ \ \mathbf{0} \ \ \mathbf{0}].$$

Then $\mathbf{c} \ll \boldsymbol{\pi}$. However

$$\mathbf{c}\bar{A} = [\mathbf{v}_1 \ \ \mathbf{0} \ \ \mathbf{0}] \neq \boldsymbol{\pi}.$$

Hence it is not true that $\mathbf{c}\bar{A} = \boldsymbol{\pi}$ whenever $\mathbf{c} \ll \boldsymbol{\pi}$. □

5.3.3 A General Convergence Result

In this section we state and prove a convergence result that is in complete contrast to Theorem 5.14 in that the present result requires no special hypotheses at all. We first state and prove the theorem, and then demonstrate that the theorem contains "less than meets the eye."

Theorem 5.15 *Suppose $A \in [0,1]^{n \times n}$ is stochastic and that $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{S}_n$. Then*

$$H(\boldsymbol{\phi}A \| \boldsymbol{\psi}A) \leq H(\boldsymbol{\phi} \| \boldsymbol{\psi}), \tag{5.32}$$

provided $\boldsymbol{\phi} \ll \boldsymbol{\psi}$ so that $H(\boldsymbol{\phi} \| \boldsymbol{\psi})$ is finite.

Theorem 5.16 *Suppose $A \in [0,1]^{n\times n}$ is stochastic and let $\boldsymbol{\pi} \in \mathbb{S}_n$ be a stationary distribution of A. Let $\boldsymbol{\phi} \in \mathbb{S}_n$ be arbitrary. If $\boldsymbol{\phi} \ll \boldsymbol{\pi}$, then the sequence $\{H(\boldsymbol{\phi}A^t\|\boldsymbol{\pi})\}_{t\geq 0}$ is nonincreasing and converges as $t \to \infty$.*

Proof. (of Theorem 5.15) We apply the log sum inequality from Lemma 3.25. Fix an index $j \in \{1,\ldots,n\}$, and apply (3.35) with

$$a_i \leftarrow \phi_i a_{ij}, b_i \leftarrow \psi_i a_{ij}.$$

If $\boldsymbol{\psi} \gg \boldsymbol{\phi}$, then $b_i = 0 \Rightarrow a_i = 0$, so we can apply (3.35). This leads to

$$\sum_{i=1}^n \phi_i a_{ij} \log \frac{\phi_i}{\psi_i} \geq \sum_{i=1}^n \phi_i a_{ij} \log \left(\frac{\sum_{i=1}^n \phi_i a_{ij}}{\sum_{i=1}^n \psi_i a_{ij}} \right), \ \forall j.$$

Hence we can sum both sides with respect to j, and the inequality still holds. Now the left side becomes

$$\begin{aligned}\sum_{j=1}^n \sum_{i=1}^n \phi_i a_{ij} \log \frac{\phi_i}{\psi_i} &= \sum_{i=1}^n \phi_i \left(\sum_{j=1}^n a_{ij} \right) \log \frac{\phi_i}{\psi_i} \\ &= \sum_{i=1}^n \phi_i \log \frac{\phi_i}{\psi_i} = H(\boldsymbol{\phi}\|\boldsymbol{\psi}),\end{aligned}$$

since $\sum_{j=1}^n a_{ij} = 1$ for all i. The right side becomes

$$\begin{aligned}&\sum_{j=1}^n \sum_{i=1}^n \phi_i a_{ij} \log \left(\frac{\sum_{i=1}^n \phi_i a_{ij}}{\sum_{i=1}^n \psi_i a_{ij}} \right) \\ &= \sum_{j=1}^n (\boldsymbol{\phi}A)_j \log \frac{(\boldsymbol{\phi}A)_j}{(\boldsymbol{\psi}A)_j} \\ &= H(\boldsymbol{\phi}A\|\boldsymbol{\psi}A).\end{aligned}$$

This establishes the desired inequality. □

Proof. (of Theorem 5.16) Apply (5.32) with $\boldsymbol{\psi} = \boldsymbol{\pi}$ and note that $\boldsymbol{\pi}A = \boldsymbol{\pi}$. So in this case (5.32) implies that

$$H(\boldsymbol{\phi}\|\boldsymbol{\pi}) \geq H(\boldsymbol{\phi}A\|\boldsymbol{\pi}).$$

We can apply this inequality recursively to see that

$$H(\boldsymbol{\phi}A^t\|\boldsymbol{\pi}) \geq H(\boldsymbol{\phi}A^{t+1}\|\boldsymbol{\pi}), \ \forall t.$$

Since the quantity $H(\boldsymbol{\phi}A^t\|\boldsymbol{\pi})$ is bounded below by zero, the sequence $\{H(\boldsymbol{\phi}A^t\|\boldsymbol{\pi})\}$ converges to some limit as $t \to \infty$. □

At first glance Theorem 5.16 appears to be counter-intuitive. It states that, with no assumptions whatsoever, the t step distribution $\boldsymbol{\phi}A^t$ gets ever closer to *every* stationary distribution $\boldsymbol{\pi}$ of A. (Note that there is no assumption in Theorem 5.16 that the stationary distribution is unique.) How can this conclusion be reconciled with Theorem 5.14?

The explanation lies in what Theorem 5.16 does *not* say. While Theorem 5.16 assures us that the divergence $H(\boldsymbol{\phi}A^t\|\boldsymbol{\pi})$ converges monotonically to some limit, *there is no guarantee that the limit is zero.* This is the extra property that is guaranteed by Theorem 5.14. This is illustrated through an example.

Example 5.5 Suppose

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\pi}_1 = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.5 \end{bmatrix}^t, \boldsymbol{\pi}_2 = \begin{bmatrix} 0.4 \\ 0.4 \\ 0.2 \end{bmatrix}^t, \boldsymbol{\phi} = \begin{bmatrix} 0.3 \\ 0.5 \\ 0.2 \end{bmatrix}^t .$$

Then both $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are stationary distributions of A, and moreover $\boldsymbol{\phi} \ll \boldsymbol{\pi}_i$ for $i = 1, 2$. Because of the nature of A, we have

$$\boldsymbol{\phi} A^t = \boldsymbol{\phi} \text{ if } t \text{ is even}, \boldsymbol{\phi} A^t = [0.5 \;\; 0.3 \;\; 0.2] \text{ if } t \text{ is odd}.$$

As a result

$$H(\boldsymbol{\phi} A^t \| \boldsymbol{\pi}_1) = 0.3 \log 1.2 + 0.5 \log 2 + 0.2 \log 0.4 \; \forall t,$$

$$H(\boldsymbol{\phi} A^t \| \boldsymbol{\pi}_2) = 0.3 \log 0.75 + 0.5 \log 1.25 \; \forall t.$$

Thus, as t varies, the divergence $H(\boldsymbol{\phi} A^t \| \boldsymbol{\pi}_i)$ remains constant at a nonzero value for *each* value of i.

Chapter Six

Markov Processes: Mixing and Estimation

In this chapter, we study the problem of estimating various quantities on the basis of observations. In Section 6.2 we state a fundamental bound, known as the Hoeffding inequality, that gives a precise quantitative estimate of the *rate* at which estimates based on observations (known as "empirical estimates") converge to their true values. Hoeffding's inequality in its original form applies to the case where successive observations of a random variable are independent. Clearly, if we observe the successive states of a Markov chain, the observations will *not* be independent. Thus we need to be able to adjust the Hoeffding and other such inequalities to the case of dependent observations. This is achieved by introducing the notions of α-mixing and β-mixing, which quantify the extent to which two random variables are dependent. Using these ideas, we can address questions such as the following: Suppose we observe a sample path of a Markov chain whose state transition matrix is unknown. Can we *estimate* the unknown state transition matrix in terms of the observed sequence of states? If so, at what "rate" does the estimated state transition matrix converge to the true but unknown state transition matrix? We conclude the chapter by defining the "divergence rate" between two Markov chains evolving over a common state space, and then giving an explicit formula for this divergence rate.

6.1 MIXING COEFFICIENTS AND ASSOCIATED INEQUALITIES

6.1.1 Mixing Coefficients Between Random Variables

"Mixing" is a way of quantifying the idea that two random variables are "nearly independent." Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in finite sets $\mathbb{A} := \{1, \ldots, n\}$ and $\mathbb{B} := \{1, \ldots, m\}$ respectively.[1] Let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution, and let P_ϕ denote their joint probability measure. Recall from Section 2.2 that the "marginal" distributions of $\mathcal{X}$

[1] In reality, the two random variables can assume values in any finite sets $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_m\}$ respectively. In other words, the range of the two random variables need not be numbers, but can be any arbitrary labels. However, identifying a_i with i and b_j with j simplifies the notation considerably.

and $\mathcal{Y}$ respectively corresponding to $\boldsymbol{\phi}$ are defined by

$$(\boldsymbol{\phi}_{\mathcal{X}})_i := \sum_{j=1}^{m} \phi_{ij}, (\boldsymbol{\phi}_{\mathcal{Y}})_j = \sum_{i=1}^{n} \phi_{ij}.$$

Thus $\boldsymbol{\phi}_{\mathcal{X}} \in \mathbb{S}_n, \boldsymbol{\phi}_{\mathcal{Y}} \in \mathbb{S}_m$. Recall from Chapter 2 that the random variables $\mathcal{X}$ and $\mathcal{Y}$ are said to be **independent** if

$$\phi_{ij} = (\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}_{\mathcal{Y}})_j, \ \forall i \in \mathbb{A}, j \in \mathbb{B}. \tag{6.1}$$

The main difficulty with the above definition of independence is that it is "binary" – either two random variables are independent or they are not. There is nothing in between. And yet it would be desirable to have some way of saying that that two random variables $\mathcal{X}$ and $\mathcal{Y}$ are "nearly independent," and making that statement precise by *quantifying* just far (or near) $\mathcal{X}$ and $\mathcal{Y}$ are to being independent. In this section, we introduce two distinct coefficients, called the α-mixing and the β-mixing coefficients respectively, that serve this purpose. There are several other types of mixing coefficients that are used in advanced theories of stochastic processes. But these two notions are good enough for the elementary discussions in this book.

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in $\mathbb{A} = \{1, \ldots, n\}$ and $\mathbb{B} = \{1, \ldots, m\}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote their joint distribution. Suppose $A \subseteq \mathbb{A}, B \subseteq \mathbb{B}$. Then their cartesian product $A \times B$ is defined in the familiar fashion as

$$A \times B := \{(i, j) : i \in A, j \in B\}.$$

Thus $A \times B$ is a subset of $\mathbb{A} \times \mathbb{B}$. Now suppose $\mathcal{X}$ and $\mathcal{Y}$ were independent random variables. Then their joint distribution $\boldsymbol{\phi}$ factors nicely as in (6.1). As a consequence, we have that

$$\begin{aligned} P_{\phi}(A \times B) &= \sum_{(i,j) \in (A \times B)} \phi_{ij} \\ &= \sum_{i \in A} \sum_{j \in B} \phi_{ij} \\ &= \sum_{i \in A} \sum_{j \in B} (\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}_{\mathcal{Y}})_j \\ &= \left(\sum_{i \in A} (\boldsymbol{\phi}_{\mathcal{X}})_i \right) \cdot \left(\sum_{j \in B} (\boldsymbol{\phi}_{\mathcal{Y}})_j \right) \\ &= P_{\phi_{\mathcal{X}}}(A) \cdot P_{\phi_{\mathcal{Y}}}(B). \end{aligned}$$

Therefore, if $\mathcal{X}$ and $\mathcal{Y}$ are not independent, then the difference between $P(A \times B)$ and $P_{\phi_{\mathcal{X}}}(A) \cdot P_{\phi_{\mathcal{Y}}}(B)$ provides a quantitative estimate of the "nonindependence" of $\mathcal{X}$ and $\mathcal{Y}$. This is the motivation for the definitions of the α-mixing and β-mixing coefficients.

Definition 6.1 *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are random variables assuming values in $\mathbb{A} = \{1, \ldots, n\}$ and $\mathbb{B} = \{1, \ldots, m\}$ respectively, and let $\boldsymbol{\phi} \in \mathbb{S}_{nm}$ denote* *their joint distribution. Then the* α**-mixing coefficient** *between* $\mathcal{X}$ *and* $\mathcal{Y}$ *is denoted by* $\alpha(\mathcal{X}, \mathcal{Y})$ *and is defined as*

$$\alpha(\mathcal{X}, \mathcal{Y}) := \max_{A \subseteq \mathbb{A}, B \subseteq \mathbb{B}} |P_\phi(A \times B) - P_{\phi_\mathcal{X}}(A) \cdot P_{\phi_\mathcal{Y}}(B)|. \tag{6.2}$$

Definition 6.2 *Suppose* $\mathcal{X}$ *and* $\mathcal{Y}$ *are random variables assuming values in* $\mathbb{A} = \{1, \ldots, n\}$ *and* $\mathbb{B} = \{1, \ldots, m\}$ *respectively. Let* $\boldsymbol{\phi}$ *denote their joint distribution, and let* $\boldsymbol{\phi}_\mathcal{X}, \boldsymbol{\phi}_{\mathbf{y}}$ *denote the marginal distributions on* $\mathbb{A}, \mathbb{B}$ *respectively. Then the* β*-mixing coefficient between* $\mathcal{X}$ *and* $\mathcal{Y}$ *is denoted by* $\beta(\mathcal{X}, \mathcal{Y})$ *and is defined by*

$$\beta(\mathcal{X}, \mathcal{Y}) := \rho(\boldsymbol{\phi}, \boldsymbol{\phi}_\mathcal{X} \times \boldsymbol{\phi}_\mathcal{Y}).$$

The β-mixing coefficient has a very intuitive interpretation. Suppose two random variables $\mathcal{X}, \mathcal{Y}$ have the joint probability measure $P_\mathcal{X} \times P_\mathcal{Y}$. Then they would have the same marginal distributions as under P, but would be independent. Thus the total variation metric $\rho(P, P_\mathcal{X} \times P_\mathcal{Y})$ quantifies the extent to which the two random variables *fail* to be independent.

To study further the properties of the two mixing coefficients, let us define the quantity

$$\delta_{ij} := \phi_{ij} - (\boldsymbol{\phi}_\mathcal{X})_i \cdot (\boldsymbol{\phi}_\mathcal{Y})_j, \ \forall i, j. \tag{6.3}$$

Since $\boldsymbol{\phi}_\mathcal{X}, \boldsymbol{\phi}_{\mathbf{y}}$ are the marginal distributions of $\boldsymbol{\phi}$, it is clear that

$$\sum_{i=1}^n \delta_{ij} = 0 \ \forall j, \sum_{j=1}^m \delta_{ij} = 0 \ \forall i. \tag{6.4}$$

Let us extend the definition of δ_{ij} in (6.3) to sets by defining

$$\delta(A, B) := \sum_{i \in A} \sum_{j \in B} \delta_{ij}, \ \forall A \subseteq \mathbb{A}, B \subseteq \mathbb{B}. \tag{6.5}$$

In particular, if $A = \{i\}, B = \{j\}$, then $\delta(A, B) = \delta_{ij}$. So the notation is consistent with (6.3). With this definition, it is clear that

$$P_\phi(A \times B) - P_{\phi_\mathcal{X}}(A) \cdot P_{\phi_\mathcal{Y}}(B) = \delta(A, B).$$

Therefore it follows directly from Definition 6.1 that

$$\alpha(\mathcal{X}, \mathcal{Y}) = \max_{A \subseteq \mathbb{A}, B \subseteq \mathbb{B}} \delta(A, B). \tag{6.6}$$

On the other hand, it readily follows from Theorem 2.8 that

$$\beta(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m |\delta_{ij}| \tag{6.7}$$

$$= \sum_{i=1}^n \sum_{j=1}^m \max\{\delta_{ij}, 0\}, \tag{6.8}$$

Thus it is quite straight-forward to compute $\beta(\mathcal{X}, \mathcal{Y})$. In contrast, in order to compute $\alpha(\mathcal{X}, \mathcal{Y})$, in general we have no option but to examine all 2^n

subsets of $\mathbb{A}$ and all 2^m subsets of $\mathbb{B}$.[2] As shown in the next example, it is not possible in general to simplify the expression for the α-mixing coefficient.

Example 6.1 The purpose of this example is to show that it is *not* possible to replace the right side of (6.2) by the simpler expression

$$\max_{i \in \mathbb{A}, j \in \mathbb{B}} |\phi_{ij} - (\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}_{\mathcal{Y}})_j|.$$

This is shown through a numerical example. Let $n = m = 4$, and suppose the joint distribution of $\mathcal{X}, \mathcal{Y}$ is as shown in the table below.

$\mathcal{X} \setminus \mathcal{Y}$	1	2	3	4	$\mathbf{p}_{\mathcal{X}}$
1	0.055	0.110	0.060	0.025	0.250
2	0.055	0.050	0.060	0.035	0.200
3	0.035	0.100	0.160	0.105	0.400
4	0.005	0.040	0.070	0.035	0.150
$\mathbf{p}_{\mathcal{Y}}$	0.150	0.300	0.350	0.200	1.000

Joint Distribution of $\mathcal{X}$ and $\mathcal{Y}$

Now the distribution of the product $\boldsymbol{\phi}_{\mathcal{X}} \times \boldsymbol{\phi}_{\mathcal{Y}}$ is shown below.

$(\boldsymbol{\phi}_{\mathcal{X}})_i \cdot (\boldsymbol{\phi}\mathcal{Y})_j$	1	2	3	4
1	0.0375	0.0750	0.0875	0.0500
2	0.0300	0.0600	0.0700	0.0400
3	0.0600	0.1200	0.1400	0.0800
4	0.0225	0.0450	0.0525	0.0300

Distribution of the $\boldsymbol{\phi}_{\mathcal{X}} \times \boldsymbol{\phi}_{\mathcal{Y}}$

Next, we display the values of the quantity δ_{ij} for all i, j.

δ_{ij}	1	2	3	4
1	0.0175	0.0350	-0.0275	-0.0250
2	0.0250	-0.0100	-0.0100	-0.0050
3	-0.0250	-0.0200	0.0200	0.0250
4	-0.0175	-0.0050	0.0175	0.0050

Values of δ_{ij}

The largest entry by absolute value in the δ matrix is 0.0350. And yet, if we choose $A = B = \{3, 4\}$, then

$$\delta(A, B) = \sum_{i \in A} \sum_{j \in B} \delta_{ij} = 0.0675 > \max_{i \in \mathbb{A}, j \in \mathbb{B}} \delta_{ij}.$$

Now we have defined two distinct mixing coefficients. Hence it is natural to ask how they are related.

Theorem 6.3 *Let $\mathcal{X}, \mathcal{Y}, \boldsymbol{\phi}$ be as before. Then*

$$\beta(\mathcal{X}, \mathcal{Y}) \geq 2\alpha(\mathcal{X}, \mathcal{Y}), \text{ or } \alpha(\mathcal{X}, \mathcal{Y}) \leq 0.5\beta(\mathcal{X}, \mathcal{Y}). \tag{6.9}$$

[2] Actually, one can exclude the cases $A = \emptyset, A = \mathbb{A}, B = \emptyset, B = \mathbb{B}$. Also, if we examine A, B, we need not examine A^c, B^c where A^c denotes the complement of A. However, the point still remains valid: It is necessary to examine $O(2^{n+m})$ choices of A and B.

Proof. Recall the definition of the quantity $\delta(A, B)$ from (6.5). If A^c denotes the complement of A, then it is easy to see that

$$\delta(A, B) = -\delta(A^c, B) = -\delta(A, B^c) = \delta(A^c, B^c).$$

By definition, $\alpha(\mathcal{X}, \mathcal{Y})$ is the maximum value of $|\delta(A, B)|$ as A varies over all subsets of $\mathbb{A}$ and B varies over all subsets of $\mathbb{B}$. In view of (6.4), there exist sets $A \subseteq \mathbb{A}, B \subseteq \mathbb{B}$ such that $\delta(A, B) = \delta(A^c, B^c) = \alpha(\mathcal{X}, \mathcal{Y})$, that is,

$$\sum_{i \in A} \sum_{j \in B} \delta_{ij} = \sum_{i \in A^c} \sum_{j \in B^c} \delta_{ij} = \alpha(\mathcal{X}, \mathcal{Y}).$$

Since δ_{ij} can be either positive or negative, it follows that

$$\sum_{i \in A} \sum_{j \in B} \max\{\delta_{ij}, 0\} \geq \sum_{i \in A} \sum_{j \in B} \delta_{ij} = \alpha(\mathcal{X}, \mathcal{Y}).$$

Similarly,

$$\sum_{i \in A^c} \sum_{j \in B^c} \max\{\delta_{ij}, 0\} \geq \sum_{i \in A^c} \sum_{j \in B^c} \delta_{ij} = \alpha(\mathcal{X}, \mathcal{Y}).$$

Finally

$$\begin{aligned} \beta(\mathcal{X}, \mathcal{Y}) &= \rho(P, P_{\mathcal{X}} \times P_{\mathcal{Y}}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \max\{\delta_{ij}, 0\} \\ &\geq \sum_{i \in A} \sum_{j \in B} \max\{\delta_{ij}, 0\} + \sum_{i \in A^c} \sum_{j \in B^c} \max\{\delta_{ij}, 0\} \\ &= 2\alpha(\mathcal{X}, \mathcal{Y}). \end{aligned}$$

This is the desired result. □

Example 6.2 Consider again the probability distribution $\boldsymbol{\phi}$ of Example 6.1. Then it can be verified through enumerating all possible subsets $A \subseteq \mathbb{A}, B \subseteq \mathbb{B}$ that $\alpha(\mathcal{X}, \mathcal{Y}) = 0.0675$. Now, from (6.8) it follows that

$$\beta(\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{4} \sum_{j=1}^{4} \max\{\delta_{ij}, 0\} = 0.14.$$

Hence $0.5\beta(\mathcal{X}, \mathcal{Y}) = 0.07$ is a very tight upper bound for $\alpha(\mathcal{X}, \mathcal{Y})$. The advantage of $\beta(\mathcal{X}, \mathcal{Y})$ over $\alpha(\mathcal{X}, \mathcal{Y})$ is that $\beta(\mathcal{X}, \mathcal{Y})$ can be computed in $O(nm)$ operations, whereas in general computing $\alpha(\mathcal{X}, \mathcal{Y})$ exactly requires $O(2^{n+m})$ operations.

6.1.2 Inequalities Associated with Mixing Coefficients

Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in the sets $\mathbb{A} = \{1, \ldots, n\}$, $\mathbb{B} = \{1, \ldots, m\}$ respectively. Let $\boldsymbol{\phi}$ denote their joint distribution and let P_{ϕ} denote their joint probability measure. Suppose f is a function of the random variable $\mathcal{X}$ and g is a function of the random variable $\mathcal{Y}$. Then the product $f(\mathcal{X})g(\mathcal{Y})$ depends on both $\mathcal{X}$ and $\mathcal{Y}$. Suppose we wish to compute the expected value $E[fg, P_\phi]$. If $\mathcal{X}$ and $\mathcal{Y}$ were independent, then we know from Theorem 2.12 that the expected value of fg would be just the product of the expected values of f and g respectively. Thus

$$E[fg, P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}] = E[f, P_{\phi_\mathcal{X}}] \cdot E[g, P_{\phi_\mathcal{Y}}].$$

Now suppose that $\mathcal{X}$ and $\mathcal{Y}$ are "nearly independent." Then we may hope that the expected value $E[fg, P_\phi]$ would be "close" to $E[f, P_{\phi_\mathcal{X}}] \cdot E[g, P_{\phi_\mathcal{Y}}]$. Theorem 6.4 below gives an upper bound of the approximation error, that is, the difference between the two quantities $E[fg, P_\phi]$ and $E[f, P_{\phi_\mathcal{X}}] \cdot E[g, P_{\phi_\mathcal{Y}}]$, in terms of the α-mixing coefficient $\alpha(\mathcal{X}, \mathcal{Y})$. Next, suppose we have a function $h(\mathcal{X}, \mathcal{Y})$ that may not be of the product form $f(\mathcal{X})g(\mathcal{Y})$. Theorem 6.6 below estimates the difference between the expected values $E[h, P_\phi]$ and $E[h, P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}]$ in terms of the β-mixing coefficient $\beta(\mathcal{X}, \mathcal{Y})$.

Theorem 6.4 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in $\mathbb{A} = \{1, \ldots, n\}$ and $\mathbb{B} = \{1, \ldots, m\}$ respectively, and let P_ϕ denote their joint probability measure. Let $P_{\phi_\mathcal{X}}, P_{\phi_\mathcal{Y}}$ denote the marginal measures of P_ϕ. Then*

$$|E[fg, P_\phi] - E[f, P_{\phi_\mathcal{X}}] \cdot E[g, P_{\phi_\mathcal{Y}}]| \leq 4\bar{f}\bar{g}\alpha(\mathcal{X}, \mathcal{Y}), \tag{6.10}$$

where

$$\bar{f} = \max_i |f_i|, \bar{g} = \max_j |g_j|.$$

The proof of the theorem is based on a preliminary lemma.

Lemma 6.5 *Let $\mathcal{X}, \mathcal{Y}, P_\phi$ be as in Theorem 6.4, and suppose $\eta : \mathbb{A} \to \{-1, 1\}, \xi : \mathbb{B} \to \{-1, 1\}$. Then*

$$|E[\eta\xi, P_\phi] - E[\eta, P_{\phi_\mathcal{X}}] \cdot E[\xi, P_{\phi_\mathcal{Y}}]| \leq 4\alpha(\mathcal{X}, \mathcal{Y}). \tag{6.11}$$

Remark: It is obvious that (6.11) is a special case of (6.10) because $\bar{\eta} = \bar{\xi} = 1$.

Proof. of Lemma 6.5: Define sets $A_+, A_- \subseteq \mathbb{A}$ and $B_+, B_- \subseteq \mathbb{B}$ as follows:

$$A_+ = \{i : \eta_i = 1\}, A_- = \{i : \eta_i = -1\},$$

$$B_+ = \{j : \xi_j = 1\}, B_- = \{j : \xi_j = -1\}.$$

Then A_+, A_- partition $\mathbb{A}$ while B_+, B_- partition $\mathbb{B}$. As a result, the four product sets $A_+ \times B_+, A_+ \times B_-, A_- \times B_+, A_- \times B_-$ partition $\mathbb{A} \times \mathbb{B}$. Moreover, $\eta_i\xi_j = 1$ whenever $ij \in A_+ \times B_+$ or $ij \in A_- \times B_-$, while $\eta_i\xi_j = -1$ whenever $ij \in A_+ \times B_-$ or $ij \in A_- \times B_+$. Hence

$$E[\eta\xi, P_\phi] = P_\phi(A_+ \times B_+) + P_\phi(A_- \times B_-) - P_\phi(A_+ \times B_+) - P_\phi(A_- \times B_+).$$

In the same way, we have

$$E[\eta, P_{\phi_\mathcal{X}}] = P_{\phi_\mathcal{X}}(A_+) - P_{\phi_\mathcal{X}}(A_-), E[\xi, P_{\phi_\mathcal{Y}}] = P_{\phi_\mathcal{Y}}(B_+) - P_{\phi_\mathcal{Y}}(B_-).$$

Hence

$$\begin{aligned} E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[\xi, P_{\phi_{\mathcal{Y}}}] = {} & P_{\phi_{\mathcal{X}}}(A_+) \cdot P_{\phi_{\mathcal{Y}}}(B_+) + P_{\phi_{\mathcal{X}}}(A_-) \cdot P_{\phi_{\mathcal{Y}}}(B_-) \\ & - P_{\phi_{\mathcal{X}}}(A_+) \cdot P_{\phi_{\mathcal{Y}}}(B_-) - P_{\phi_{\mathcal{X}}}(A_-) \cdot P_{\phi_{\mathcal{Y}}}(B_+). \end{aligned}$$

So, if we define

$$\epsilon = E[\eta\xi, P_{\phi}] - E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[\xi, P_{\phi_{\mathcal{Y}}}],$$

then

$$\begin{aligned} \epsilon = {} & [P(A_+ \times B_+) - P_{\phi_{\mathcal{X}}}(A_+) \cdot P_{\phi_{\mathcal{Y}}}(B_+)] \\ & + [P(A_- \times B_-) - P_{\phi_{\mathcal{X}}}(A_-) \cdot P_{\phi_{\mathcal{Y}}}(B_-) \\ & + [P_{\phi_{\mathcal{X}}}(A_+) \cdot P_{\phi_{\mathcal{Y}}}(B_-) - P(A_+ \times B_+] \\ & + [P_{\phi_{\mathcal{X}}}(A_-) \cdot P_{\phi_{\mathcal{Y}}}(B_+) - P(A_- \times B_+)]. \end{aligned}$$

Now note that each of the four quantities on the right side is bounded above $\alpha(\mathcal{X}, \mathcal{Y})$ and below by $-\alpha(\mathcal{X}, \mathcal{Y})$. Hence

$$-4\alpha(\mathcal{X}, \mathcal{Y}) \leq \epsilon \leq 4\alpha(\mathcal{X}, \mathcal{Y}),$$

or

$$|\epsilon| \leq 4\alpha(\mathcal{X}, \mathcal{Y}).$$

This is the desired inequality. □

Proof. of Theorem 6.4: Define a function $g_{\mathcal{X}} : \mathbb{A} \to \mathbb{R}$ as follows:

$$(g_{\mathcal{X}})_i := \frac{\sum_{j=1}^{m} g_j \phi_{ij}}{(\boldsymbol{\phi}_{\mathcal{X}})_i} = \frac{\sum_{j=1}^{m} g_j \phi_{ij}}{\sum_{j=1}^{m} \phi_{ij}}. \tag{6.12}$$

We recognize from Definition 2.18 that $g_{\mathcal{X}}$ is just the conditional expectation of $g(\mathcal{Y})$, viewed as a function of both $\mathcal{X}, \mathcal{Y}$, with respect to $\mathcal{X}$; compare with (2.33). Then it follows from Theorem 2.20 that

$$E[fg, P_{\phi}] = E[fg_{\mathcal{X}}, P_{\phi_{\mathcal{X}}}]. \tag{6.13}$$

Since $E[g, P_{\phi_{\mathcal{Y}}}]$ is just a constant, it follows from the above that

$$\begin{aligned} E[fg, P_{\phi}] - E[f, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}] & = E[fg_{\mathcal{X}}, P_{\phi_{\mathcal{X}}}] - E[f, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}] \\ & = E[f(g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]), P_{\phi_{\mathcal{X}}}]. \end{aligned}$$

Hence, if we define

$$\gamma := E[fg, P_{\phi}] - E[f, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}],$$

then it follows that

$$\gamma = E[f(g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]), P_{\phi_{\mathcal{X}}}] \leq \bar{f} E[|g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]|, P_{\phi_{\mathcal{X}}}]. \tag{6.14}$$

Now define a function $\eta : \mathbb{A} \to \{-1, 1\}$ by

$$\eta_i := \text{sign}\{(g_{\mathcal{X}})_i - E[g, P_{\phi_{\mathcal{Y}}}])\}, i = 1, \ldots, n, \tag{6.15}$$

where

$$\text{sign}(x) := \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{if } x < 0. \end{cases}$$

Then it is easy to see that

$$|g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]| = \eta \cdot (g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]),$$

whence (6.14) becomes

$$\gamma \leq \bar{f} E[\eta(g_{\mathcal{X}} - E[g, P_{\phi_{\mathcal{Y}}}]), P_{\phi_{\mathcal{X}}}]. \tag{6.16}$$

The right side of this equation can be manipulated further using Theorem 2.20. This theorem implies that

$$E[\eta g_{\mathcal{X}}, P_{\phi_{\mathcal{X}}}] = E[\eta g, P_{\phi}].$$

Substituting this into (6.16) shows that

$$\gamma \leq \bar{f}\{E[\eta g, P_{\phi}] - E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}]\} \leq \bar{f}\zeta, \tag{6.17}$$

where

$$\zeta := E[\eta g, P_{\phi}] - E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}].$$

Now ζ looks exactly like γ, except that the real-valued function f has been replaced by the bipolar-valued function η of $\mathcal{X}$.

To proceed further, we repeat the process, this time taking conditional expectations with respect to $\mathcal{Y}$. This leads to

$$\begin{aligned}
\zeta &= E[\eta g, P_{\phi}] - E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}] \\
&= E[\eta_{\mathcal{Y}} g, P_{\phi_{\mathcal{Y}}}] - E[\eta, P_{\phi_{\mathcal{X}}}] \cdot E[g, P_{\phi_{\mathcal{Y}}}] \\
&= E[g \cdot (\eta_{\mathcal{Y}} - E[\eta, P_{\phi_{\mathcal{X}}}]), P_{\phi_{\mathcal{Y}}}].
\end{aligned}$$

Now in analogy with (6.15), let us define the function $\xi : \mathbb{B} \to \{-1, 1\}$ by

$$\xi_j := \operatorname{sign}\{(\eta_{\mathcal{Y}})_j - E[\eta, P_{\phi_{\mathcal{X}}}]\}, \ \forall j,$$

and observe that

$$|(\eta_{\mathcal{Y}})_j - E[\eta, P_{\phi_{\mathcal{X}}}]| = \xi_j \cdot \{(\eta_{\mathcal{Y}})_j - E[\eta, P_{\phi_{\mathcal{X}}}]\} \ \forall j.$$

Then

$$\begin{aligned}
\zeta &= E[g \cdot (\eta_{\mathcal{Y}} - E[\eta, P_{\phi_{\mathcal{X}}}]), P_{\phi_{\mathcal{Y}}}] \\
&\leq \bar{g} E[|\eta_{\mathcal{Y}} - E[\eta, P_{\phi_{\mathcal{X}}}]|, P_{\phi_{\mathcal{Y}}}] \\
&= \bar{g} E[\xi(\eta_{\mathcal{Y}} - E[\eta, P_{\phi_{\mathcal{X}}}]), P_{\phi_{\mathcal{Y}}}] \\
&= \bar{g}\{E[\xi \eta_{\mathcal{Y}}, P_{\phi_{\mathcal{Y}}}] - E[\xi, P_{\phi_{\mathcal{Y}}}] \cdot E[\eta, P_{\phi_{\mathcal{X}}}]\} \\
&= \bar{g}\{E[\xi \eta, P_{\phi}] - E[\xi, P_{\phi_{\mathcal{Y}}}] \cdot E[\eta, P_{\phi_{\mathcal{X}}}]\}.
\end{aligned}$$

Substituting this bound into (6.17) shows that

$$\gamma \leq \bar{f}\bar{g}\{E[\xi\eta, P_{\phi}] - E[\xi, P_{\phi_{\mathcal{Y}}}] \cdot E[\eta, P_{\phi_{\mathcal{X}}}]\},$$

where η, ξ are now *bipolar-valued* functions. Now we can invoke Lemma 6.5, specifically (6.11), to conclude that

$$\gamma \leq 4\bar{f}\bar{g}\alpha(\mathcal{X}, \mathcal{Y}).$$

Similar reasoning shows that

$$\gamma \geq -4\bar{f}\bar{g}\alpha(\mathcal{X}, \mathcal{Y}).$$

Finally these two inequalities together show that

$$|\gamma| \leq 4\bar{f}\bar{g}\alpha(\mathcal{X}, \mathcal{Y}).$$

This is the desired conclusion. □

Theorem 6.6 *Suppose $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in $\mathbb{A} = \{1, \ldots, n\}$ and $\mathbb{B} = \{1, \ldots, m\}$ respectively, and let P_ϕ denote their joint probability measure. Let $P_{\phi_\mathcal{X}}, P_{\phi_\mathcal{Y}}$ denote the marginal measures of P_ϕ. Suppose $h : \mathbb{A} \times \mathbb{B} \to [a, b]$. Then*

$$|E[h, P_\phi] - E[h, P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}]| \leq (b - a)\beta(\mathcal{X}, \mathcal{Y}). \tag{6.18}$$

Proof. Let us scale the function h so that it assumes values in the unit interval, by defining

$$h'_{ij} = \frac{h_{ij} - a}{b - a}, \ \forall i, j.$$

Then a routine calculation shows that

$$E[h', P_\phi] - E[h', P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}] = \frac{E[h, P_\phi] - E[h, P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}]}{b - a}.$$

Now, since $h' : \mathbb{A} \times \mathbb{B} \to [0, 1]$, it follows from Lemma 2.10 that

$$|E[h', P_\phi] - E[h', P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}]| \leq \rho(P, P_{\phi_\mathcal{X}} \times P_{\phi_\mathcal{Y}}) = \beta(\mathcal{X}, \mathcal{Y}),$$

which in turn implies (6.18). □

6.2 MIXING COEFFICIENTS OF A MARKOV PROCESS

6.3 ESTIMATING MEANS FROM SAMPLE PATHS: HOEFFDING'S INEQUALITY

Suppose we are given a coin with two sides which we denote by heads (H) and tails (T). We wish to know what the probability is of the coin turning up heads. If we knew the detailed mass distribution of the coin, perhaps we might be able to compute this probability starting from first principles. But it would appear to be more natural to toss a number of times and observe how many times heads appears. Suppose we toss the coin 100 times and heads appears 62 times. Then the ratio $62/100 = 0.62$ is called the **empirical probability** of heads, and should not be confused with the true but unknown probability of heads, call it $p(H)$. Let $\hat{p}_m(H)$ denote the empirical probability of heads after m coin tosses, where the hat above the p serves to remind us that $\hat{p}_m(H)$ is only an approximation to $p(H)$. Now $\hat{p}_m(H)$ is itself a *random variable* assuming values in the interval $[0, 1]$. If we were to toss the coin another 100 times, there is no reason to suppose that we would once again get 62 heads. An old theorem in probability theory tells us that $\hat{p}_m(H)$ approaches $p(H)$ 'almost surely' as $m \to \infty$. In a badly written book on probability, the probability of heads $p(H)$ is even 'defined' as the limit of $\hat{p}_m(H)$ as $m \to \infty$. However what concerns us at present is not the *asymptotic* behavior of $\hat{p}_m(H)$, but rather its *finite time* behavior. Hoeffding's inequality is a famous theorem that allows us to deduce how close $\hat{p}_m(H)$ is to $p(H)$ after a finite number m of trials.

Theorem 6.7 (Hoeffding's Inequality) *Suppose $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ are independent random variables, where $\mathcal{Y}_i$ assumes values in the bounded interval $[a_1, b_i]$. Then for each real number ϵ, we have*

$$\Pr\{\sum_{i=1}^{m}[\mathcal{Y}_i - E(\mathcal{Y}_i)] \geq \epsilon\} \leq \exp\left[-2\epsilon^2 / \sum_{i=1}^{m}(b_i - a_i)^2\right], \tag{6.19}$$

where $E(\mathcal{Y}_i)$ denotes the expected value of $\mathcal{Y}_i$.

The proof of Hoeffding's inequality uses the following auxiliary lemma.

Lemma 6.8 *Suppose $\mathcal{X}$ is a zero-mean random variable assuming values in the interval $[a, b]$. Then for any $s > 0$, we have*

$$E[\exp(s\mathcal{X})] \leq \exp(s^2(b-a)^2/8).$$

Proof. **(of Lemma 6.8)**: Since the exponential is a convex function, the value of e^{sx} is bounded by the corresponding convex combination of its extreme values; that is,

$$\exp(sx) \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}, \ \forall x \in [a, b].$$

Now take the expectation of both sides, and use the fact that $E(\mathcal{X}) = 0$. This gives

$$\begin{aligned} E[\exp(s\mathcal{X})] &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} \\ &=: \exp(\phi(u)), \end{aligned}$$

where $p := -a/(b-a), u := s(b-a)$, and $\phi(u) := -pu + \ln(1 - p + pe^u)$. Clearly $\phi(u) = 0$. Moreover, a routine calculation shows that

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

whence $\phi'(u) = 0$ as well. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq 0.25.$$

Hence by Taylor's theorem, there exists a $\theta \in [0, u]$ such that

$$\phi(u) = \frac{\phi''(\theta)u^2}{2} \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

This completes the proof. □

Proof. **(of Theorem 6.7)**: For any nonnegative random variable, we have from Corollary 2.26 that

$$\Pr\{\mathcal{X} \geq \epsilon\} \leq e^{-s\epsilon}E[\exp(s\mathcal{X})].$$

Now apply this inequality to the random variable

$$\mathcal{Z}_m := \sum_{i=1}^{m} (\mathcal{Y}_i - E(\mathcal{Y}_i)),$$

which has zero mean since the $\mathcal{Y}_i$'s are independent. Then

$$\begin{aligned}
\Pr\{\mathcal{Z}_m \geq \epsilon\} &\leq e^{-s\epsilon}\, E\left[\exp\left(s \sum_{i=1}^{m} (\mathcal{Y}_i - E(\mathcal{Y}_i))\right)\right] \\
&= e^{-s\epsilon} \prod_{i=1}^{m} E[\exp(s(\mathcal{Y}_i - E(\mathcal{Y}_i)))] \text{ by independence} \\
&\leq e^{-s\epsilon} \prod_{i=1}^{m} \exp[s^2 (b_i - a_i)^2/8] \text{ by Lemma 6.8} \\
&= \exp\left[-s\epsilon + s^2 \sum_{i=1}^{m} \frac{(b_i - a_i)^2}{8}\right] \\
&= \exp\left[\frac{-2\epsilon^2}{\sum_{i=1}^{m} (b_i - a_i)^2}\right], \qquad (6.20)
\end{aligned}$$

where the last step follows by choosing

$$s = \frac{4\epsilon}{\sum_{i=1}^{m} (b_i - a_i)^2}.$$

This completes the proof. □

A useful (and widely used) 'corollary' of Hoeffding's inequality is obtained when we take repeated and independent measurements of *the same* random variable. Because of its importance, we state the 'corollary' as a theorem.

Theorem 6.9 (Hoeffding's Inequality for i.i.d. processes) *Suppose* $\mathcal{Y}$ *is a random variable assuming values in a bounded interval* $[a, b]$*, and that* $y_1, \ldots, y_m$ *are independent realizations of* $\mathcal{Y}$*. Then for each* $\epsilon > 0$*, we have*

$$\Pr\left\{\frac{1}{m} \sum_{i=1}^{m} y_i - E(\mathcal{Y}) \geq \epsilon\right\} \leq \exp[-2m\epsilon^2/(b-a)^2], \qquad (6.21)$$

$$\Pr\left\{\frac{1}{m} \sum_{i=1}^{m} y_i - E(\mathcal{Y}) \leq -\epsilon\right\} \leq \exp[-2m\epsilon^2/(b-a)^2], \qquad (6.22)$$

$$\Pr\left\{\left|\frac{1}{m} \sum_{i=1}^{m} y_i - E(\mathcal{Y})\right| \geq \epsilon\right\} \leq 2\exp[-2m\epsilon^2/(b-a)^2], \qquad (6.23)$$

Proof. To prove (6.21), apply (6.19) with ϵ replaced by $m\epsilon$, and $a_i = a, b_i = b$ for all i. To prove (6.22), apply (6.21) with $\mathcal{Y}$ replaced by $-\mathcal{Y}$. Finally (6.23) is a direct consequence of (6.21) and (6.22). □

Theorem 6.9 has a natural interpretation in terms of estimating the mean or expected value of a random variable on the basis of successive independent measurements. Suppose $\mathcal{Y}$ is the random variable whose mean we wish to estimate, and for this purpose we have generated m independent samples $y_1, \ldots, y_m$ of $\mathcal{Y}$. The quantity

$$\hat{E}_m(\mathcal{Y}) := \frac{1}{m} \sum_{i=1}^{m} y_i$$

is called the **empirical mean** of $\mathcal{Y}$, as it is just the average of the m observations of $\mathcal{Y}$. Now, inequalities (6.21) through (6.23) quantify the *rate* at which the empirical mean converges to the true mean $E(\mathcal{Y})$, as the number of samples m approaches infinity. The bound (6.23) states that, after we have drawn m independent samples and computed the empirical mean $\hat{E}_m(\mathcal{Y})$ as above, we can say with confidence $1 - 2\exp[-2m\epsilon^2/(b-a)^2]$ that the empirical mean is within ϵ of the true mean $E(\mathcal{Y})$. The inequalities (6.2) and (6.22) give 'one-sided' bounds on the likelihood that $\hat{E}_m(\mathcal{Y}) \geq E(\mathcal{Y}) + \epsilon$ and $\hat{E}_m(\mathcal{Y}) \leq E(\mathcal{Y}) - \epsilon$ respectively. It is noteworthy that the right sides of all three inequalities *approach* zero as $m \to \infty$, but will never *exactly equal* zero.

Hoeffding's inequality was proved in 1963; see [56]. Since then various researchers have attempted to improve the bound, but could not succeed in doing so. And it is no wonder. In 1990, Massart [81] proved that Hoeffding's inequality is, in a very precise sense, the 'best possible' inequality.

Note that Hoeffding's inequality is stated here for real-valued random variables. So how can it be applied to random variables that assume values in some discrete set that has no obvious interpretation as a subset of the real numbers (e.g. the set of nucleotides)? The trick is to associate a *binary-valued* random variable, assuming the (real) values 0 and 1, with the random variable assuming values in an abstract set. To illustrate, let us return to the problem of estimating the probability of a coin turning up heads. Let us define a random variable $\mathcal{Y}$ such that $\mathcal{Y} = 1$ if the coin toss turns up heads, and $\mathcal{Y} = 0$ if the coin toss turns up tails. Then it is clear that $\Pr\{\mathcal{Y} = 1\} = p_H$, the probability of heads. Moreover, $E(\mathcal{Y})$ also equals $p(H)$. Hence the fraction of heads that turn up during a coin toss experiment, which we have called $\hat{p}_m(H)$ earlier, is the empirical mean of $\mathcal{Y}$. Therefore we can apply Hoeffding's inequality with $b = 1, a = 0$ and assert that

$$\Pr\{|\hat{p}_m(H) - P(H)| \geq \epsilon\} \leq 2\exp(-2m\epsilon^2).$$

Suppose that a random variable $\mathcal{X}$ assumes not just two but some finite number n of values, which need not be real numbers. Specifically, suppose $\mathcal{X}$ assumes values in $\mathbb{A} = \{a_1, \ldots, a_n\}$. Suppose we generate m independent realizations of $\mathcal{X}$, denoted by $x_1, \ldots, x_m$. Then, *for each index* i, we can define an associated binary-valued random variable $\mathcal{Y}_i$ as follows: $\mathcal{Y}_i = 1$ if $\mathcal{X} = a_i$, and $\mathcal{Y}_i = 0$ otherwise. With this association, it is clear that the expected value of $\mathcal{Y}_i$ is precisely $\Pr\{\mathcal{X} = a_i\} =: p_i$. Now, using the realizations $x_1, \ldots, x_m$, let us define n different empirical probabilities as follows:

$$\hat{p}_m(i) := \frac{1}{m} \sum_{j=1}^{m} I_{\{x_j = a_i\}},$$

where I denotes the indicator function. Thus $\hat{p}_m(i)$ is precisely the fraction of times that the outcome a_i appears amongst the m trials. Now we can apply Hoeffding's inequality to each of the n binary-valued random variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ and assert that

$$\Pr\{|\hat{p}_m(i) - p_i| \geq \epsilon\} \leq 2 \exp(-2m\epsilon^2), i = 1, \ldots, n.$$

These n separate bounds can be combined into the single, and very useful, bound

$$\Pr\{|\hat{p}_m(i) - p_i| \leq \epsilon \; \forall i\} \geq 1 - 2n \exp(-2m\epsilon^2). \tag{6.24}$$

Since n, the cardinality of the set $\mathbb{A}$, appears explicitly on the right side of the above equation, this approach is not useful for infinite sets, and alternate approaches need to be devised. However, for random variables $\mathcal{X}$ assuming values in a finite set of cardinality n, (6.24) states that, after m independent trials, we can state with confidence $1 - 2n \exp(-2m\epsilon^2)$ that *every one of the n estimates $\hat{p}_m(i)$* is within ϵ of its true value.

6.4 ESTIMATING THE STATE TRANSITION MATRIX

Chapter Seven

Introduction to Large Deviation Theory

7.1 PROBLEM FORMULATION

In this chapter, we take some baby steps in a very important part of probability theory, known as large deviation theory.[1] We begin by describing briefly the motivation for the problem under study. Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set. Let $\mathcal{M}(\mathbb{A})$ denote the set of all probability distributions on the set $\mathbb{A}$. Clearly one can identify $\mathcal{M}(\mathbb{A})$ with the n-simplex $\mathbb{S}_n$. Suppose $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A})$ is a fixed but possibly unknown probability distribution, and $\mathcal{X}$ is a random variable assuming values in $\mathbb{A}$ with the distribution $\boldsymbol{\mu}$. In order to estimate $\boldsymbol{\mu}$, we generate independent samples $x_1, \ldots, x_l, \ldots$, where each x_i belongs to $\mathbb{A}$, is distributed according to $\boldsymbol{\mu}$, and is independent of x_j for $j \neq i$. The symbol $\mathbf{x}_1^l := x_1 \ldots x_l \in \mathbb{A}^l$ denotes the multisample that represents the outcome of the first l experiments. Based on this multisample, we can construct an **empirical distribution** $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ as follows:

$$(\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l))_i := \frac{1}{l} \sum_{j=1}^{l} I_{\{x_j = a_i\}}, \tag{7.1}$$

where I denotes the indicator function. Thus

$$I_{\{x_j = a_i\}} = \begin{cases} 1 & \text{if} \quad x_j = a_i, \\ 0 & \text{if} \quad x_j \neq a_i, \end{cases}$$

In words, (7.1) simply states that $\hat{\boldsymbol{\mu}}_i(\mathbf{x})$ equals the fraction of the samples $x_1, \ldots, x_l$ that equal the symbol a_i. Since every sample x_i has to equal one of the a_i's, it is easy to see that $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ is also a probability distribution on $\mathbb{A}$. Moreover $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ is a 'random' element of $\mathcal{M}(\mathbb{A})$ since it is based on the random multisample $\mathbf{x}$. Thus we can think of $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$ as a stochastic process that assumes values in $\mathcal{M}(\mathbb{A})$ and ask: As $l \to \infty$, does this process converge to the true but possibly unknown measure $\boldsymbol{\mu}$ that is generating the samples, and if so, at what rate?

To address this question, the first thing we do is to convert the question from one of studying a stochastic process into one of studying a sequence of real numbers. Suppose $\Gamma \subseteq \mathcal{M}(\mathbb{A})$ is some set of probability distributions. Then $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}_{l \geq 1}$ is a sequence of real numbers. So it makes sense to study the behavior of this sequence as $l \to \infty$. What is the interpretation of 'Pr' in this context? Clearly the empirical distribution $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ depends only on the first l samples $\mathbf{x}_1^l$. So $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} = P_\mu^l\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}$.

[1] I firmly resisted the temptation to say 'some small steps in large deviation theory'.

Suppose now that $\boldsymbol{\mu} \not\in \bar{\Gamma}$; thus the true measure $\boldsymbol{\mu}$ that is generating the random samples does not belong to the closure of the set Γ. Since Γ is a subset of $\mathbb{R}^n$, a finite-dimensional space, all topologies are equivalent; however, to be definite, we can use the total variation metric to define what is meant by the closure of a subset of $\mathcal{M}(\mathbb{A})$. Now, we believe that as we draw more and more samples, the empirical measure $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ will converge to the true measure $\boldsymbol{\mu}$ (in some vague sense not yet made precise). Hence, we believe that if $\boldsymbol{\mu} \not\in \bar{\Gamma}$, then the sequence of real numbers $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}$ will converge to zero. Large deviation theory is concerned with the *rate* at which this sequence converges to zero, and how the rate depends on the set Γ and the true distribution $\boldsymbol{\mu}$.

Specifically, suppose this sequence converges to zero at an exponential rate; that is

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \sim c_1 \exp(-lc_2).$$

Then the constant c_2 is the rate of convergence (which will in general depend on both Γ and $\boldsymbol{\mu}$). How can we 'get at' this rate? We can compute the quantity

$$\frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \sim \frac{\log c_1}{l} - c_2,$$

and observe that as $l \to \infty$, the negative of this quantity approaches c_2. Motivated by this observation, we define something called the 'rate function'. Since we will modify the definition almost at once, let us call this a 'trial definition'.

Let us call a function $I : \mathcal{M}(\mathbb{A}) \to \mathbb{R}_+$ a 'rate function' if it has the following properties: (i) Whenever $\Gamma \subseteq \mathcal{M}(\mathbb{A})$ is an *open set*, we have

$$- \inf_{\boldsymbol{\nu} \in \Gamma} I(\boldsymbol{\nu}) \leq \liminf_{l \to \infty} \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}. \tag{7.2}$$

(ii) Whenever $\Gamma \subseteq \mathcal{M}(\mathbb{A})$ is a *closed set*, we have

$$\limsup_{l \to \infty} \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \leq - \inf_{\boldsymbol{\nu} \in \Gamma} I(\boldsymbol{\nu}). \tag{7.3}$$

However, the above definition leaves a few issues unresolved. First, there are no specifications about the nature of the function I, so it could be quite erratic. Second, there is no requirement that the rate function be unique. Third, there are two separate conditions, one about what happens when Γ is an open set and another about what happens when Γ is a closed set, but nothing about what happens for 'in-between' sets Γ.

To overcome these issues, we introduce the notion of a lower semi-continuous function, and then that of a lower semi-continuous relaxation. Given a number $\epsilon > 0$, let us define $\mathcal{B}(\boldsymbol{\nu}, \epsilon)$ to be the open ball of radius centered at $\boldsymbol{\nu}$. Thus

$$\mathcal{B}(\boldsymbol{\nu}, \epsilon) := \{\boldsymbol{\phi} \in \mathbb{S}_n : \rho(\boldsymbol{\nu}, \boldsymbol{\phi}) < \epsilon\}.$$

Similarly, we define

$$\bar{\mathcal{B}}(\boldsymbol{\nu}, \epsilon) := \{\boldsymbol{\phi} \in \mathbb{S}_n : \rho(\boldsymbol{\nu}, \boldsymbol{\phi}) \leq \epsilon\}.$$

A function $f : \mathcal{M}(\mathbb{A}) \to \mathbb{R}$ is said to be **lower semi-continuous** if

$$\boldsymbol{\nu}_i \to \boldsymbol{\nu}^* \;\Rightarrow\; f(\boldsymbol{\nu}^*) \leq \liminf_i f(\boldsymbol{\nu}_i).$$

Contrast this with the definition of continuity, which is

$$\boldsymbol{\nu}_i \to \boldsymbol{\nu}^* \;\Rightarrow\; f(\boldsymbol{\nu}^*) = \lim_i f(\boldsymbol{\nu}_i).$$

Now, given any function $I : \mathcal{M}(\mathbb{A}) \to \mathbb{R}_+$, we define its **lower semi-continuous relaxation** I_* by

$$I_*(\boldsymbol{\nu}) := \lim_{\epsilon \to 0} \inf_{\boldsymbol{\phi} \in \mathcal{B}(\boldsymbol{\nu}, \epsilon)} I(\boldsymbol{\phi}).$$

It is left as a problem to verify that $I_*(\cdot)$ is indeed lower semi-continuous.

Now suppose that a rate function $I(\cdot)$ satisfies (7.2) and (7.3). Then so does its lower semi-continuous relaxation $I_*(\cdot)$. Moreover, for every $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{A})$, we can write

$$\begin{aligned} -I_*(\boldsymbol{\nu}) &= \lim_{\epsilon \to 0} \liminf_{l \to \infty} \frac{1}{l} \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \mathcal{B}(\boldsymbol{\nu}, \epsilon)\} \\ &= \lim_{\epsilon \to 0} \limsup_{l \to \infty} \frac{1}{l} \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \bar{\mathcal{B}}(\boldsymbol{\nu}, \epsilon)\}. \end{aligned}$$

This shows that if the rate function is lower semi-continuous, then it is unique. Putting it another way, it is conceivable that two distinct functions can satisfy (7.2) and (7.3), but if so, they will both have exactly the same lower semi-continuous relaxation. Finally, if the rate function $I(\cdot)$ is lower semi-continuous (we drop the subscript for convenience), then the two equations (7.2) and (7.3) can be combined into the single equation

$$\begin{aligned} -\inf_{\boldsymbol{\nu} \in \Gamma^o} I(\boldsymbol{\nu}) &\leq \liminf_{l \to \infty} \frac{1}{l} \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \\ &\leq \limsup_{l \to \infty} \frac{1}{l} \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \leq -\inf_{\boldsymbol{\nu} \in \bar{\Gamma}} I(\boldsymbol{\nu}), \end{aligned} \tag{7.4}$$

where Γ^o denotes the interior of the set Γ. All of these statements are not easy for a beginner to see, so the problems at the end of the section give a step by step guide to proving these assertions.

Now we are ready to state the definition of the rate function as it is normally given in the literature on large deviation theory.

Definition 7.1 *Let the symbols $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ be as above. Then a function $I : \mathcal{M}(\mathbb{A}) \to \mathbb{R}_+$ is said to be a (the)* **rate function** *of the stochastic process $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$ if*

1. *I is lower semi-continuous.*

2. *For every set $\Gamma \subseteq \mathcal{M}(\mathbb{A})$, the relationships in (7.4) hold.*

In such a case the process $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$ is said to satisfy the **large deviation property (LDP)** *with rate function $I(\cdot)$.*

In the above definition, the stochastic process under study is $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$, which assumes values in the *compact set* $\mathbb{S}_n$. We can think of a more general situation (not encountered in this book) where we study a stochastic process $\{\mathcal{X}_l\}$ that assumes values in $\mathbb{R}^n$. See for example [35, 48] for statements of the large deviation property in the general case. In such a case the rate function $I(\cdot)$ would have to be defined over all of $\mathbb{R}^n$, wherever the random variables $\mathcal{X}_l$ have their range. The rate function $I(\cdot)$ is said to be a **good rate function** if the so-called 'level sets'

$$L_\alpha := \{\mathbf{x} \in \mathbb{R}^n : I(\mathbf{x}) \leq \alpha\}$$

are compact for all α. However, we need not worry about 'good' rate functions in the present context since the domain of the rate function is anyhow a compact set; so the above condition is automatically satisfied.

Note that the assumption that I is lower semi-continuous makes the rate function unique if it exists. Suppose that the set Γ does not have any isolated points, i.e. that $\Gamma \subseteq \overline{\Gamma^o}$. Suppose also that the function $I(\cdot)$ is continuous, and not merely lower semi-continuous (and observe in passing that the function $\boldsymbol{\nu} \mapsto H(\boldsymbol{\nu} \| \boldsymbol{\mu})$ is indeed continuous). Then the two extreme infima in (7.4) coincide. As a result the liminf and limsup are equal, which means that both equal the limit of the sequence. This means that

$$\lim_{l \to \infty} \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} = -\inf_{\boldsymbol{\nu} \in \Gamma} I(\boldsymbol{\nu}). \tag{7.5}$$

One of the attractions of the large deviation property is this precise estimate of the rate at which $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}$ approaches zero. Let us define α to be the infimum on the right side of (7.5). Then (7.5) can be rewritten as

$$\frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \to -\alpha \text{ as } l \to \infty.$$

This means that, very roughly speaking,

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\} \sim c_1(l) \exp(c_2(l)),$$

where $c_1(l)$ is subexponential in the sense that $(\log c_1(l))/l$ approaches zero as $l \to \infty$, while $c_2 \to -\alpha$ as $l \to \infty$.

In the next section, it is shown that for the specific problem discussed here (namely, i.i.d. processes assuming values in a finite set $\mathbb{A}$ with distribution $\boldsymbol{\mu}$), the rate function is actually $I(\boldsymbol{\nu}) = H(\boldsymbol{\nu} \| \boldsymbol{\mu})$. But in anticipation of that, let us discuss the implications of the definition. Equation (7.5) means that the rate at which $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}$ approaches zero depends only on the *infimum value* of the rate function $I(\cdot)$ over Γ. To understand the implications of this, consider the diagram below, which shows a 'large' set Γ_1, and a much smaller set Γ_2 shown in the figure. Suppose the rate function $I(\cdot)$ is continuous and assumes its minimum over Γ_1 at exactly one choice of $\boldsymbol{\nu}_*$, indicated by the red dot in the figure. Moreover, assume that $I(\boldsymbol{\nu}_*) > 0$; this is the case if the true probability distribution $\boldsymbol{\mu}$ does not belong to $\bar{\Gamma}$, because in this case $H(\boldsymbol{\nu}_* \| \boldsymbol{\mu}) > 0$. Now, since $\boldsymbol{\nu}_* \in \Gamma_2$, it is clear that

$$\inf_{\boldsymbol{\nu} \in \Gamma_1} I(\boldsymbol{\nu}) = \inf_{\boldsymbol{\nu} \in \Gamma_2} I(\boldsymbol{\nu}).$$

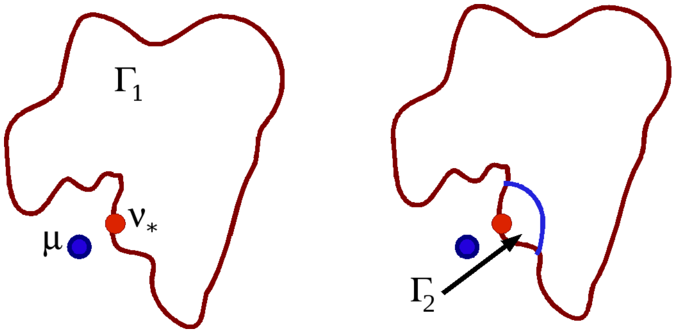


Figure 7.1 Implications of (7.1)

Moreover, if we let $\Gamma_1 \setminus \Gamma_2$ denote the complement of Γ_2 in Γ_1, that is,

$$\Gamma_1 \setminus \Gamma_2 = \{\boldsymbol{\nu} : \boldsymbol{\nu} \in \Gamma_1, \boldsymbol{\nu} \not\in \Gamma_2\},$$

then the continuity of the rate function implies that

$$\inf_{\boldsymbol{\nu} \in \Gamma_1 \setminus \Gamma_2} > \inf_{\boldsymbol{\nu} \in \Gamma_1} I(\boldsymbol{\nu}). \tag{7.6}$$

Now it is clear that, for every l, we have

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1\} = \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_2\} + \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1 \setminus \Gamma_2\}.$$

Therefore (7.6) means that the second quantity on the right side approaches zero at a faster exponential rate compared to the first quantity. So we can use this fact to deduce what precisely happens as $l \to \infty$. Since $\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1\}$ approaches zero at an exponential rate, we can think of $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1$ as a 'rare' event, which becomes more and more rare as $l \to \infty$. However, *if we now condition on this rare event occuring*, and ask where precisely within the large set Γ_1 the empirical distribution $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ is likely to lie, we see from the above discussion that

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_2 | \hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1\} = \frac{\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_2\}}{\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1\}} \to 1 \text{ as } l \to \infty.$$

Since the above argument can be repeated for *every* set Γ_2 such that (7.6) is true, we see that the conditional distribution of $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$, conditioned on the event that $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma_1$, becomes more and more peaked around $\boldsymbol{\nu}_*$ as $l \to \infty$. The above argument does not really depend on the fact that the rate function $I(\cdot)$ assumes its minimum value at exactly one point in the set Γ_1. If the minimum of the rate function occurs at finitely many points in Γ_1, then as $l \to \infty$, the conditional distribution above becomes more and more peaked around these minima.

7.2 LARGE DEVIATION PROPERTY FOR I.I.D. SAMPLES: SANOV'S THEOREM

In this section we derive the exact rate function for the process of empirical measures $\{\hat{\boldsymbol{\mu}}(x_1^l)\}$ defined in the preceding section, in the important case where the samples $x_1, x_2, \ldots$ are generated independently using a common distribution $\boldsymbol{\mu}$. Specifically, it is shown that the rate function is $I(\boldsymbol{\nu}) = H(\boldsymbol{\nu}\|\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the actual distribution generating the samples. This is known as Sanov's theorem (but it must be pointed out that the original Sanov's theorem is *not* limited to the case of finite alphabets). Along the way, we also introduce a very important *approach* known as 'the method of types'. This method makes the proofs very transparent and simple, and can also be extended to the case of non-i.i.d. (or dependent) processes.

Let us begin by restating the problem under study. Suppose $\mathbb{A} = \{a_1, \ldots, a_n\}$ is a finite set. Suppose $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A})$ is a fixed but possibly unknown probability distribution, and $\mathcal{X}$ is a random variable assuming values in $\mathbb{A}$ with the distribution $\boldsymbol{\mu}$. In order to estimate $\boldsymbol{\mu}$, we define an i.i.d. sequence $\{\mathcal{X}_t\}$ where each $\mathcal{X}_t$ has the distribution $\boldsymbol{\mu}$, and let $\mathbf{x} = x_1, \mathbf{x}_2, \ldots$ denote a sample path (or realization) of this i.i.d. process. Based on the first l samples of this realization, we can construct an **empirical distribution** $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ as follows:

$$(\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l))_i := \frac{1}{l} \sum_{j=1}^{l} I_{\{x_j = a_i\}}. \tag{7.7}$$

The objective is to make precise estimates the probability $P_\mu^l\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) \in \Gamma\}$ where $\Gamma \subseteq \mathcal{M}(\mathbb{A})$.

For this purpose, we introduce the 'method of types'. For the moment let us fix the integer l denoting the length of the multisample. Then it is easy to see that the empirical distribution $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ can take only a finite set of values. For one thing, it is clear from (7.7) that every element of $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ is a rational number with denominator equal to l. Let $\mathcal{E}(l, n)$ denote the set of all possible empirical distributions that can result from a multisample of length l.[2] We will denote elements of $\mathcal{E}(l, n)$ by symbols such as $\boldsymbol{\nu}, \boldsymbol{\phi}$ etc. Suppose $\mathbf{x}_1^l, \mathbf{y}_1^l \in \mathbb{A}^l$ are two multisamples. We define these two multisamples to be **equivalent** if they generate the same empirical estimate, i.e., if $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) = \hat{\boldsymbol{\mu}}(\mathbf{y}_1^l)$. It is easy to see that this is indeed an equivalence relation. For each distribution $\boldsymbol{\nu} \in \mathcal{E}(l, n)$, the set of all multisamples $\mathbf{x}_1^l \in \mathbb{A}$ such that $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) = \boldsymbol{\nu}$ is called the **type class** of $\boldsymbol{\nu}$, and is denoted by $T(\boldsymbol{\nu}, l)$.

Example 7.1 Suppose the alphabet $\mathbb{A}$ has cardinality 2. For simplicity we can write $\mathbb{A} = \{1, 2\}$. Suppose $l = 5$. Then there are only six possible empirical distributions, namely

$$\mathcal{E}(5, 2) = \{[0/5 \ \ 5/5], [1/5 \ \ 4/5], [2/5 \ \ 3/5], [3/5 \ \ 2/5], [4/5 \ \ 1/5], [5/5 \ \ 0/5]\}.$$

[2] It is clear that the possible empirical distributions depend only on n, the cardinality of the alphabet $\mathbb{A}$, and not on the nature of the elements of $\mathbb{A}$.

Next, suppose $\boldsymbol{\nu} = [2/5 \;\; 3/5]$. Then the type class $T(\boldsymbol{\nu}, 5)$ consists of all elements of $\mathbb{A}^5$ that contain precisely two 1's and three 2's. These can be written out explicitly, as

$$T([2/5 \;\; 3/5], 5) = \begin{array}{ll} \{ & [11222], [12122], [12212], [12221], [21122], \\ & [21212], [21221], [22112], [22121], [22211] \quad \}. \end{array}$$

Note that it is necessary to identify the type class not only with the empirical distribution $\boldsymbol{\nu}$ but also with the length l of the multisample. For instance, if we keep the same $\boldsymbol{\nu}$ but change l to 10, then $T(\boldsymbol{\nu}, 10)$ would consist of all elements of $\mathbb{A}^{10}$ that contain precisely four 1's and six 2's.

The 'method of types' consists of addressing (in no particular order) the following questions:

- What is the cardinality of $\mathcal{E}(l, n)$? In other words, how many distinct empirical distributions can be generated as $\mathbf{x}_1^l$ varies over $\mathbb{A}^l$? As we shall see shortly, an upper bound will do.
- What is the (log) likelihood of each multisample in $T(\boldsymbol{\nu}, l)$, and how is it related to $\boldsymbol{\nu}$?
- For each empirical distribution $\boldsymbol{\nu} \in \mathcal{E}(l, n)$, what is the cardinality of the associated type class $T(\boldsymbol{\nu}, l)$? We require both upper as well as lower bounds on this cardinality.

Let us address these questions in this order (though, as stated above, we could address them in any order we choose to).

Lemma 7.2 *We have that*

$$|\mathcal{E}(l, n)| = \left(\begin{array}{c} l + n - 1 \\ n - 1 \end{array} \right) = \frac{(l + n - 1)!}{l!(n - 1)!}. \tag{7.8}$$

Proof. The proof is by induction on n. If $n = 2$, then it is easy to see that $\mathcal{E}(l, 2)$ consists of all distributions of the form $[k/l \;\; (l-k)/l]$ as k varies from 0 to l. Thus $|\mathcal{E}(l, 2)| = l + 1$ for all l, and (7.8) holds for $n = 2$, for all l.

To proceed by induction, suppose (7.8) holds up to $n - 1$, for all l, and suppose $|\mathbb{A}| = n$. Suppose $\boldsymbol{\nu} \in \mathcal{E}(l, n)$. Thus each component of $l\boldsymbol{\nu}$ is an integer, and together they must add up to l. Let k denote the first component of $l\boldsymbol{\nu}$, and note that k can have the values $0, 1, \ldots, l$. If $k = l$ then the next $n - 1$ components of $l\boldsymbol{\nu}$ must all equal zero. Hence there is only one vector $\boldsymbol{\nu} \in \mathcal{E}(l, n)$ with the first component equal to 1. For each $k = 0, \ldots, l - 1$, the next $n - 1$ components of $l\boldsymbol{\nu}$ must all be integers and add up to $l - k$. Thus the next $n - 1$ components of $l\boldsymbol{\nu}$ can have $|\mathcal{E}(l - k, n - 1)|$ possible values. Thus we have established the following recursive relationship:

$$|\mathcal{E}(l, n)| = 1 + \sum_{k=0}^{l-1} |\mathcal{E}(l - k, n - 1)| = 1 + \sum_{k=1}^{l} |\mathcal{E}(k, n - 1)|,$$

after changing the dummy variable of summation. Therefore we need to solve the above recursion with the starting condition $|\mathcal{E}(l, 2)| = l + 1$.

Now let us recall the following property of binomial coefficients:

$$\binom{m}{n} = \binom{m-1}{n} + \binom{m-1}{n-1}.$$

This property can be verified by multiplying both sides by $n!(m-n)!$ and collecting terms. The above equation can be rewritten as

$$\binom{m-1}{n-1} = \binom{m}{n} - \binom{m-1}{n}.$$

Substituting this relationship into the recursive formula for $|\mathcal{E}(l,n)|$, and using the inductive hypothesis gives

$$\begin{aligned} |\mathcal{E}(l,n)| &= 1 + \sum_{k=1}^{l} \binom{k+n-2}{n-2} \\ &= 1 + \sum_{k=1}^{l} \left[\binom{k+n-1}{n-1} - \binom{k+n-2}{n-1} \right]. \end{aligned}$$

Note that when $k = 1$, we have that

$$\binom{k+n-2}{n-1} = 1.$$

which cancels the very first term of 1. So this is a telescoping sum, whereby the negative terms and positive terms cancel out, leaving only the very last positive term; that is,

$$|\mathcal{E}(l,n)| = \binom{l+n-1}{n-1}.$$

This completes the proof by induction. □

Lemma 7.3 *Suppose $\boldsymbol{\nu} \in \mathcal{E}(l,n)$, and suppose that the multisample $\mathbf{x}_1^l \in \mathbb{A}^l$ belongs to the corresponding type class $T(\boldsymbol{\nu}, l)$. Then*

$$\log P_\mu^l\{\mathbf{x}\} = lJ(\boldsymbol{\nu}, \boldsymbol{\mu}), \tag{7.9}$$

where $J(\cdot,\cdot)$ is the loss function defined in (3.30).

Proof. Since $\boldsymbol{\nu} \in \mathcal{E}(l,n)$, every component of $l\boldsymbol{\nu}$ is an integer. Let us define $l_i = l\nu_i$ for each i. Then the multisample $\mathbf{x}_1^l$ contains precisely l_i occurrences of the symbol a_i, for each i. Since the samples are independent of each other, the *order* in which the various symbols occur in $\mathbf{x}_1^l$ does not affect its likelihood. So it follows that

$$P_\mu^l\{\mathbf{x}\} = \prod_{i=1}^{l} \mu_i^{l_i},$$

whence

$$\log P_\mu^l\{\mathbf{x}\} = \sum_{i=1}^{n} l_i \log \mu_i = l \sum_{i=1}^{n} \nu_i \log \mu_i = lJ(\boldsymbol{\nu}, \boldsymbol{\mu}).$$

□

Now we come to the last of the questions raised above, namely: What is the cardinality of the type class $T(\boldsymbol{\nu}, l)$? Using the same symbols $l_i = l\nu_i$, we see that $|T(\boldsymbol{\nu}, l)|$ is the number of ways of choosing l symbols from $\mathbb{A}$ in such a way that the symbol a_i occurs precisely l_i times. This number equals

$$|T(\boldsymbol{\nu}, l)| = \frac{l!}{\prod_{i=1}^n l_1!}.$$

So we derive upper and lower bounds for $|T(\boldsymbol{\nu}, l)|$. As a part of doing this, we also derive another result that may be of independent interest.

Lemma 7.4 *Suppose* $\boldsymbol{\nu}, \boldsymbol{\phi} \in \mathcal{E}(l, n)$. *Then*

$$P_\nu^l(T(\boldsymbol{\phi}, l)) \leq P_\nu^l(T(\boldsymbol{\nu}, l)). \tag{7.10}$$

Remark: The lemma asserts that if we examine all the type classes, then the measure of each type class is maximum under the corresponding distribution.

Proof. Since $\boldsymbol{\nu}, \boldsymbol{\phi} \in \mathcal{E}(l, n)$, let us define $l_i = l\nu_i$ as above, and $k_i = l\phi_i$ for all i. Then it follows that

$$P_\nu^l(T(\boldsymbol{\phi}, l)) = |T(\boldsymbol{\phi}, l)| \prod_{i=1}^n \nu_i^{k_i} = \frac{l!}{\prod_{i=1}^n k_i!} \prod_{i=1}^n \nu_i^{k_i} = \frac{l!}{l^l} \prod_{i=1}^n \frac{l_i^{k_i}}{k_i!},$$

where we take advantage of the fact that the k_i's add up to l. Now the term $l!/l^l$ is just some integer constant that is independent of both $\boldsymbol{\nu}, \boldsymbol{\phi}$ and can thus be ignored in future calculations. Thus

$$\log P_\nu^l(T(\boldsymbol{\phi}, l)) = \text{const.} + \sum_{i=1}^n [k_i \log l_i - \log k_i!].$$

If we note that $\log(k_i!) = \sum_{j=1}^{k_i} \log s$, we can rewrite the above as

$$\log P_\nu^l(T(\boldsymbol{\phi}, l)) = \text{const.} + \sum_{i=1}^n \sum_{j=1}^{k_i} [\log l_i - \log j] = \text{const.} + \sum_{i=1}^n \sum_{j=1}^{k_i} \log \frac{l_i}{j}.$$

Let us define

$$c(k_1, \ldots, k_n) := \sum_{i=1}^n \sum_{j=1}^{k_i} \log \frac{l_i}{j}.$$

The claim is that above quantity is maximized when $k_i = l_i$ for all i.

To prove this, let us note first that

$$c(l_1, \ldots, l_n) = \sum_{i=1}^n \sum_{j=1}^{l_i} \log \frac{l_i}{j}.$$

To show that this is the maximum value for the function c, suppose we choose $k_i = l_i + 1$ and $k_{1'} = l_{i'} - 1$ for some indices i, i'. To make the notation less messy, let us rearrange the indices such that $k_1 = l_1 + 1, k_2 = l_2 - 1$ and $k_i = l_i$ for $i \geq 3$. Then it is easy to see that

$$c(l_1 + 1, l_2 - 1, l_3, \ldots, l_n) = c(l_1, \ldots, l_n) + \log \frac{l_1}{l_1 + 1} < c(l_1, \ldots, l_n).$$

Thus increasing any one of the indices by 1 from l_i, and decreasing another one to ensure that the k_i's still add up to l, can only decrease the function c. □

Lemma 7.5 *Let $H(\cdot)$ denote the entropy of a distribution, as before. Then*

$$[|\mathcal{E}(l,n)|]^{-1} \exp[lH(\boldsymbol{\nu})] \leq |T(\boldsymbol{\nu}, l)| \leq \exp[lH(\boldsymbol{\nu})], \ \forall \boldsymbol{\nu} \in \mathcal{E}(l,n). \tag{7.11}$$

Proof. Fix $\boldsymbol{\nu} \in \mathcal{E}(l,n)$, and let $\mathbf{x}_1^l \in T(\boldsymbol{\nu}, l)$. Then $\mathbf{x}_1^l$ contains precisely $l_i = l\nu_i$ occurrences of the symbol a_i. Therefore

$$P_\nu^l(\{\mathbf{x}\}) = \prod_{i=1}^n \nu_i^{l_i}, \ \forall \mathbf{x} \in T(\boldsymbol{\nu}, l),$$

$$\log P_\nu^l(\{\mathbf{x}\})) = \sum_{i=1}^n l_i \log \nu_i = l \sum_{i=1}^n \nu_i \log \nu_i = -lH(\boldsymbol{\nu}), \ \forall \mathbf{x} \in T(\boldsymbol{\nu}, l),$$

$$P_\nu^l(\{\mathbf{x}\})) = \exp[-lH(\boldsymbol{\nu})], \ \forall \mathbf{x} \in T(\boldsymbol{\nu}, l),$$

$$P_\nu^l(T(\boldsymbol{\nu}, l)) = |T(\boldsymbol{\nu}, l)| \exp[-lH(\boldsymbol{\nu})]. \tag{7.12}$$

Since P_ν^l is a probability measure on $\mathbb{A}^l$ and $T(\boldsymbol{\nu}, l) \subseteq \mathbb{A}^l$, it follows that

$$1 \geq P_\nu^l(T(\boldsymbol{\nu}, l)) \geq |T(\boldsymbol{\nu}, l)| \exp[-lH(\boldsymbol{\nu})],$$

which is the right inequality in (7.11).

To prove the left inequality in (7.11), observe that the various type classes $T(\boldsymbol{\phi}, l), \boldsymbol{\phi} \in E(l,n)$ partition the sample space $\mathbb{A}^l$. Therefore it follows from Lemma 7.5 that

$$\begin{aligned} 1 = P_\nu^l(\mathbb{A}^l) &= \sum_{\boldsymbol{\phi} \in \mathcal{E}(l,n)} P_\nu^l(T(\boldsymbol{\phi}, l)) \\ &\leq |\mathcal{E}(l,n)| P_\nu^l(T(\boldsymbol{\nu}, l)) \\ &= |\mathcal{E}(l,n)| |T(\boldsymbol{\nu}, l)| \exp[-lH(\boldsymbol{\nu})], \end{aligned}$$

where in the last step we make use of (7.12). This leads to the left inequality in (7.11). □

Theorem 7.6 (Sanov's Theorem for a Finite Alphabet) *The stochastic process $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$ defined in (7.7) satisfies the large deviation property with the rate function*

$$I(\boldsymbol{\nu}) = H(\boldsymbol{\nu} \| \boldsymbol{\mu}). \tag{7.13}$$

Proof. To show that the function $I(\cdot)$ defined above is indeed the rate function, we apply Definition 7.1. It is clear that the function $I(\cdot)$ is not only lower semi-continuous but is in fact continuous. So it remains only to verify the two inequalities in (7.4).

Suppose $\boldsymbol{\nu} \in \mathcal{E}(l,n)$, and let us ask: What is the probability that the empirical distribution $\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)$ equals $\boldsymbol{\nu}$? Clearly the answer is

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) = \boldsymbol{\nu}\} = P_\mu^l(T(\boldsymbol{\nu},l)) = |T(\boldsymbol{\nu},l)| P_\mu^l(\{\mathbf{x}_1^l\}) \text{ for any one } \mathbf{x}_1^l \in T(\boldsymbol{\nu},l).$$

Now if we use (7.9) and the right inequality in (7.11), we can conclude that

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) = \boldsymbol{\nu}\} \leq \exp[-lH(\boldsymbol{\nu}) + J(\boldsymbol{\nu}\|\boldsymbol{\mu})] = \exp[-lH(\boldsymbol{\nu}\|\boldsymbol{\mu})]. \tag{7.14}$$

Similarly, by using (7.9) and the left inequality in (7.11), we can conclude that

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l) = \boldsymbol{\nu}\} \geq [|\mathcal{E}(l,n)|]^{-1} \exp[-lH(\boldsymbol{\nu}\|\boldsymbol{\mu})]. \tag{7.15}$$

Now let $\Gamma \subseteq \mathcal{M}(\mathbb{A})$ be any set of probability distributions on $\mathbb{A}$. Then

$$\begin{aligned}\Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} &= \sum_{\boldsymbol{\nu} \in \mathcal{E}(l,n) \cap \Gamma} \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\} \\ &\leq |\mathcal{E}(l,n) \cap \Gamma| \sup_{\boldsymbol{\nu} \in \mathcal{E}(l,n) \cap \Gamma} \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\} \\ &\leq |\mathcal{E}(l,n)| \sup_{\boldsymbol{\nu} \in \mathcal{E}(l,n) \cap \Gamma} \exp[-lH(\boldsymbol{\nu}\|\boldsymbol{\mu})].\end{aligned}$$

Hence

$$\frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} \leq \frac{1}{l} \log |\mathcal{E}(l,n)| + \sup_{\boldsymbol{\nu} \in \Gamma} -H(\boldsymbol{\nu}\|\boldsymbol{\mu}). \tag{7.16}$$

Now we can make a crude estimate of $|\mathcal{E}(l,n)|$ using (7.8). It is clear that, if $l \geq n$, then $l + n - 1 \leq 2l$, so that

$$|\mathcal{E}(l,n)| \leq \frac{2^{n-1}}{(n-1)!} l^{n-1}, \ \forall l \geq n.$$

This is not a particularly 'clever' estimate, but as we shall see below, the constants in front will not matter anyhow when we take the limit as $l \to \infty$. Since $|\mathcal{E}(l,n)|$ is polynomial in l, the first term on the right side of (7.16) approaches zero as $l \to \infty$. Therefore

$$\limsup_{l \to \infty} \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} \leq \sup_{\boldsymbol{\nu} \in \Gamma} -H(\boldsymbol{\nu}\|\boldsymbol{\mu}) = -\inf_{\boldsymbol{\nu} \in \Gamma} H(\boldsymbol{\nu}\|\boldsymbol{\mu}).$$

This establishes the right inequality in (7.4).

To establish the left inequality in (7.4), we begin by showing that $\cup_l \mathcal{E}(l,n)$ is dense in $\mathcal{M}(\mathbb{A})$; that is, for every $\boldsymbol{\phi} \in \mathcal{M}(\mathbb{A})$ and every $\epsilon > 0$, there exists an integer l and a distribution $\boldsymbol{\nu} \in \mathcal{E}(l,n)$ such that $\rho(\boldsymbol{\nu},\boldsymbol{\phi}) \leq \epsilon$. In words, this says that every distribution in $\mathcal{M}(\mathbb{A})$ can be approximated arbitrarily closely by an empirical distribution in $\mathcal{E}(l,n)$ for sufficiently large l. To establish this fact, given ϵ and $\boldsymbol{\phi}$, choose an integer l such that $n/l \leq \epsilon$, or equivalently $l \geq n/\epsilon$. Then for every index $i = 1, \ldots, n$, there exists an integer q_i such that $|\phi_i - q_i/l| \leq 1/l$. Since $\sum_i \phi_i = 1$, we can also ensure that $\sum_i q_i = l$, if necessary by increasing or decreasing some of the q_i by one as needed. This ensures that there exists a distribution $\boldsymbol{\nu} = [q_1 \ldots q_n]/l \in \mathcal{E}(l, n)$ such that $|\phi_i - \nu_i| \leq 2/l$ for all i. For this choice of $\boldsymbol{\nu}$, it follows that

$$\rho(\boldsymbol{\nu}, \boldsymbol{\phi}) = \frac{1}{2} \sum_{i=1}^{n} |\nu_i - \phi_i| \leq n/l \leq \epsilon.$$

Now suppose $\boldsymbol{\nu}$ is an interior point of Γ. Then there is an open ball $\mathcal{B}(\boldsymbol{\nu})$ in $\mathcal{M}(\mathbb{A})$ that contains $\boldsymbol{\nu}$. Since $\cup_l \mathcal{E}(l, n)$ is dense in $\mathcal{M}(\mathbb{A})$, there exist a sequence of integers l_k and corresponding elements $\boldsymbol{\nu}_{l_k} \in \mathcal{E}(l_k, n) \cap \Gamma$ such that $l_k \to \infty$ and $\boldsymbol{\nu}_{l_k} \to \boldsymbol{\nu}$. Hence

$$\Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^{l_k}) \in \Gamma\} \geq \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^{l_k}) = \boldsymbol{\nu}_{l_k}\} \geq [|\mathcal{E}(l_k, n)|]^{-1} \exp[-l_k H(\boldsymbol{\nu}_{l_k} \| \boldsymbol{\mu})],$$

where the last step follows from (7.15). Therefore

$$\begin{aligned} \frac{1}{l_k} \log \Pr\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^{l_k}) \in \Gamma\} &\geq -\frac{1}{l_k} \log |\mathcal{E}(l_k, n)| - H(\boldsymbol{\nu}_{l_k} \| \boldsymbol{\mu}) \\ &\geq -H(\boldsymbol{\nu}_{l_k} \| \boldsymbol{\mu}) + o(1/l_k) \\ &\to -H(\boldsymbol{\nu} \| \boldsymbol{\mu}) \text{ as } k \to \infty. \end{aligned}$$

Hence it follows that

$$\liminf_{l \to \infty} \frac{1}{l} \log \Pr\{\boldsymbol{\nu}(\mathbf{x}_1^l) \in \Gamma\} \geq -H(\boldsymbol{\nu} \| \boldsymbol{\mu}), \ \forall \boldsymbol{\nu} \in \Gamma^o.$$

Since the above inequality holds for every $\boldsymbol{\nu} \in \Gamma^o$, we can conclude that

$$\liminf_{l \to \infty} \frac{1}{l} \log \Pr\{\boldsymbol{\nu}(\mathbf{x}_1^l) \in \Gamma\} \geq \sup_{\boldsymbol{\nu} \in \Gamma^o} -H(\boldsymbol{\nu} \| \boldsymbol{\mu}) = -\inf_{\boldsymbol{\nu} \in \Gamma^o} H(\boldsymbol{\nu} \| \boldsymbol{\mu}).$$

This is the left inequality in (7.4). Since both of the relationships in (7.4) hold with $I(\boldsymbol{\nu}) = H(\boldsymbol{\nu} \| \boldsymbol{\mu})$, it follows that $H(\boldsymbol{\nu} \| \boldsymbol{\mu})$ is the rate function. □

7.3 LARGE DEVIATION PROPERTY FOR MARKOV CHAINS

In this section we extend the results of the previous section from i.i.d. processes to Markov chains. Along the way we introduce an alternate description of Markov chains in terms of 'consistent' distributions, and also introduce a very useful notion known variously as the Kullback-Leibler divergence *rate* between two processes, or the *differential* relative entropy between two consistent distributions.

7.3.1 Consistent Distributions: An Alternate Description of Markov Chains

Suppose $\{\mathcal{X}_t\}$ is a stationary Markov process assuming values in $\mathbb{A}$. Until now we have been describing such a Markov chain is in terms of two entities: (i) The **stationary distribution** $\boldsymbol{\pi} \in \mathcal{M}(\mathbb{A})$, where

$$\pi_i := \Pr\{\mathcal{X}_t = i\}, i \in \mathbb{A}.$$

(ii) The **state transition matrix** $A \in [0,1]^{n \times n}$, where

$$a_{ij} := \Pr\{\mathcal{X}_{t+1} = j | \mathcal{X}_t = i\}, \ \forall i, j \in \mathbb{A}.$$

Note that each row of A is a probability distribution and belongs to $\mathcal{M}(\mathbb{A})$, since the i-th row of A is the conditional distribution of $\mathcal{X}_{t+1}$ given that $\mathcal{X}_t = i$. Thus A is a stochastic matrix.

But there is an alternative description of a Markov chain that is more convenient for present purposes, namely the *vector of doublet frequencies.* To motivate this alternate description, we first introduce the notion of a consistent distribution on $\mathbb{A}^k$. Other authors also use the phrase 'stationary' distribution.

Suppose $\{\mathcal{X}_t\}$ is a stationary stochastic process (not necessarily Markov) assuming values in a finite set $\mathbb{A}$. To make notation more compact, define

$$\mathcal{X}_s^t := \mathcal{X}_s \mathcal{X}_{s+1} \ldots \mathcal{X}_{t-1} \mathcal{X}_t.$$

Clearly this notation makes sense only when $s \leq t$. For each integer $k \geq 1$ and each $\mathbf{i} := (i_1, \ldots, i_k) \in \mathbb{A}^k$, let us define the k-tuple frequencies

$$\mu_{\mathbf{i}} := \Pr\{\mathcal{X}_{t+1}^{t+k} = i_1 \ldots i_k\}.$$

This probability does not depend on t since the process is stationary. Now note that, for each $(k-1)$-tuple $\mathbf{i} \in \mathbb{A}^{k-1}$, the events

$$\{\mathcal{X}_{t+1}^{t+k} = i_1 \ldots i_{k-1} 1\}, \ldots, \{\mathcal{X}_{t+1}^{t+k} = i_1 \ldots i_{k-1} n\}$$

are mutually disjoint, and together generate the event

$$\{\mathcal{X}_{t+1}^{t+k-1} = \mathbf{i}\}.$$

Thus

$$\mu_{\mathbf{i}} = \sum_{j \in \mathbb{A}} \mu_{\mathbf{i}j}, \ \forall \mathbf{i} \in \mathbb{A}^{k-1}.$$

By entirely analogous reasoning, it also follows that

$$\mu_{\mathbf{i}} = \sum_{j \in \mathbb{A}} \mu_{j\mathbf{i}}, \ \forall \mathbf{i} \in \mathbb{A}^{k-1}.$$

This motivates the next definition.

Definition 7.7 *A distribution* $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{A}^2)$ *is said to be* **consistent** *(or stationary) if*

$$\sum_{j \in \mathbb{A}} \nu_{ij} = \sum_{j \in \mathbb{A}} \nu_{ji}, \ \forall i \in \mathbb{A}. \tag{7.17}$$

For $k \geq 3$*, a distribution* $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{A}^k)$ *is said to be* **consistent** *(or stationary) if*

$$\sum_{j \in \mathbb{A}} \nu_{\mathbf{i}j} = \sum_{j \in \mathbb{A}} \nu_{j\mathbf{i}}, \ \forall \mathbf{i} \in \mathbb{A}^{k-1},$$

and in addition, the resulting distribution $\bar{\boldsymbol{\nu}}$ *on* $\mathbb{A}^{k-1}$ *defined by*

$$\bar{\boldsymbol{\nu}}_{\mathbf{i}} := \sum_{j \in \mathbb{A}} \nu_{\mathbf{i}j} \ \forall \mathbf{i} \in \mathbb{A}^{k-1} = \sum_{j \in \mathbb{A}} \nu_{j\mathbf{i}} \ \forall \mathbf{i} \in \mathbb{A}^{k-1}$$

is consistent. Equivalently, a distribution $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{A}^k)$ *is said to be* **consistent** *(or stationary) if*

$$\sum_{\mathbf{j} \in \mathbb{A}^l} \nu_{\mathbf{ij}} = \sum_{\mathbf{j} \in \mathbb{A}^l} \nu_{\mathbf{ji}}, \ \forall \mathbf{i} \in \mathbb{A}^{k-l}, \ \forall l \leq k-1. \tag{7.18}$$

The set of all consistent distributions on $\mathbb{A}^k$ *is denoted by* $\mathcal{M}_c(\mathbb{A}^k)$.

Now let us return to Markov chains. Suppose $\{\mathcal{X}_t\}$ is a stationary Markov chain assuming values in the finite set $\mathbb{A}$. Define the vector $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A}^2)$ by

$$\mu_{ij} = \Pr\{\mathcal{X}_t \mathcal{X}_{t+1} = ij\}, \ \forall i, j \in \mathbb{A}.$$

Then, as per the above discussion, $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^2)$. The vector $\boldsymbol{\mu}$ is called the **vector of doublet frequencies**. The claim is that the doublet frequency vector $\boldsymbol{\mu}$ captures all the relevant information about the Markov chain, and does so in a more natural way. The stationary distribution of the Markov chain is given by

$$\bar{\mu}_i := \sum_{j \in \mathbb{A}} \mu_{ij} = \sum_{j \in \mathbb{A}} \mu_{ji},$$

while the state transition matrix A is given by

$$a_{ij} = \frac{\mu_{ij}}{\bar{\mu}_i}.$$

Dividing by $\bar{\mu}_i$ can be justified by observing that if $\bar{\mu}_i = 0$ for some index i, then the corresponding element i can simply be dropped from the set $\mathbb{A}$. With these definitions, it readily follows that $\bar{\boldsymbol{\mu}}$ is a row eigenvector of A, because

$$(\bar{\boldsymbol{\mu}} A)_j = \sum_{i=1}^{n} \bar{\mu}_i a_{ij} = \sum_{i=1}^{n} \mu_{ij} = \bar{\mu}_j.$$

Note that the above reasoning breaks down if $\boldsymbol{\mu} \in \mathbb{S}_{n^2}$ but $\boldsymbol{\mu} \notin \mathcal{M}_c(\mathbb{A}^2)$.

More generally, suppose $\{\mathcal{X}_t\}$ is an s-step Markov chain, so that

$$E\{\mathcal{X}_t | \mathcal{X}_{t-1} \ldots \mathcal{X}_0\} = E\{\mathcal{X}_t | \mathcal{X}_{t-1} \ldots \mathcal{X}_{t-s}\} \ \forall t.$$

Then the process is completely characterized by its $(s+1)$-tuple frequencies

$$\mu_{\mathbf{i}} := \Pr\{\mathcal{X}_t^{t+s} = \mathbf{i}\}, \ \forall \mathbf{i} \in \mathbb{A}^{s+1}.$$

The probability distribution $\boldsymbol{\mu}$ is consistent and thus belongs to $\mathcal{M}_c(\mathbb{A}^{s+1})$. Now an s-step Markov chain assuming values in $\mathbb{A}$ can also be viewed as a conventional (one-step) Markov chain over the state space $\mathbb{A}^s$. Moreover, if the current state is $i\mathbf{j}$ where $i \in \mathbb{A}, \mathbf{j} \in \mathbb{A}^{s-1}$, then a transition is possible only to a state of the form $\mathbf{j}k, k \in \mathbb{A}$. Thus, even though the state transition matrix has dimension $n^s \times n^s$, each row of the transition matrix can have at most n nonzero elements. The entry in row $i\mathbf{j}$ and column $\mathbf{j}k$ equals

$$\Pr\{\mathcal{X}_t = k | \mathcal{X}_{t-s}^{t-1} = i\mathbf{j}\} = \frac{\mu_{i\mathbf{j}k}}{\bar{\mu}_{i\mathbf{j}}},$$

where as before we define

$$\bar{\mu}_{\mathbf{l}} := \sum_{i \in \mathbb{A}} \mu_{i\mathbf{l}} = \sum_{i \in \mathbb{A}} \mu_{\mathbf{l}i}, \ \forall \mathbf{l} \in \mathbb{A}^s.$$

Next, we introduce two important notions called differential entropy, and differential relative entropy (which can also be referred to as differential Kullback-Leibler divergence).

Suppose $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^k)$ are *consistent* distributions on $\mathbb{A}^k$ for some integer $k \geq 2$. We define $\bar{\boldsymbol{\mu}} \in \mathcal{M}_c(\mathbb{A}^{k-1})$ by

$$\bar{\mu}_{\mathbf{i}} := \sum_{j \in \mathbb{A}} \mu_{\mathbf{i}j} = \sum_{j \in \mathbb{A}} \mu_{j\mathbf{i}}, \ \forall \mathbf{i} \in \mathbb{A}^{k-1}.$$

The overbar serves to remind us that $\bar{\boldsymbol{\mu}}$ is "reduced by one dimension" from $\boldsymbol{\mu}$. Because $\boldsymbol{\mu}$ is a consistent distribution, it does not matter whether the reduction is on the first component or the last. The symbol $\bar{\boldsymbol{\nu}}$ is defined similarly. With this notation we can now define differential entropy etc.

Definition 7.8 *Suppose $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^k)$ for some integer $k \geq 2$. Then*

$$D(\boldsymbol{\mu}) := H(\boldsymbol{\mu}) - H(\bar{\boldsymbol{\mu}}) \tag{7.19}$$

is called the **differential entropy** *of $\boldsymbol{\mu}$, while*

$$D(\boldsymbol{\nu} \| \boldsymbol{\mu}) := H(\boldsymbol{\nu} \| \boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}} \| \bar{\boldsymbol{\mu}}) \tag{7.20}$$

is called the **differential relative entropy** *between $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$, and also the* **differential Kullback-Leibler divergence** *between $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$.*

We can compare and contrast the above definition with Definition 3.7 of entropy and Definition 3.18 of relative entropy. Note that $\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}$ are marginal distributions of $\boldsymbol{\mu}, \boldsymbol{\nu}$ respectively on $\mathbb{A}^{k-1}$. Thus it readily follows that both $D(\cdot)$ and $D(\cdot\|\cdot)$ are nonnegative-valued. Note that we use the symbol H for the (relative) entropy and D for the differential (relative) entropy.

7.3.2 Kullback-Leibler Divergence Rate Between Markov Chains

Suppose $\{\mathcal{X}_t\}_{t \geq 0}, \{\mathcal{Y}_t\}_{t \geq 0}$ are two stationary stochastic processes (not necessarily Markov) on a *common* finite alphabet $\mathbb{A}$. In the present context, this means simply that each process is a sequence of random variables assuming values in $\mathbb{A}$, indexed by the variable t which we can think of as 'time'. Let $\tilde{P}, \tilde{Q}$ denote the 'laws' of the stochastic processes $\{\mathcal{X}_t\}$ and $\{\mathcal{Y}_t\}$ respectively. We have studiously 'avoided the infinite' in this book, but in essence we can think of $\tilde{P}, \tilde{Q}$ as probability measures on the countably infinite cartesian product $\prod_{i=1}^{\infty} \mathbb{A}$. As we shall see below, actually we will have to deal only with joint distributions of finitely many random variables, but it is helpful to have symbols (even if they are not properly defined) to denote the laws of the two stochastic processes. For each integer l, let $\boldsymbol{\phi}^{(l)}$ denote the joint distribution of $\mathcal{X}_1^l = (\mathcal{X}_1, \ldots, \mathcal{X}_l)$ under the law $\tilde{P}$, and similarly let $\boldsymbol{\theta}^{(l)}$ denote the joint distribution of $\mathcal{Y}_1^l$ under the law $\tilde{P}$. Note that both $\boldsymbol{\phi}^{(l)}, \boldsymbol{\theta}^{(l)}$ are distributions on the finite set $\mathbb{A}^l$.

Definition 7.9 *The* **Kullback-Leibler divergence rate** *between the two laws $\tilde{P}, \tilde{Q}$ is defined as the limit, if it exists, as follows:*

$$R(\tilde{Q}\|\tilde{P}) := \lim_{l\to\infty} \frac{1}{l} H(\boldsymbol{\theta}^{(l)}\|\boldsymbol{\phi}^{(l)}), \tag{7.21}$$

Note that the above definition is quite general and does not require either process to be Markovian. Of course, in general, the limit can also fail to exist. However, it is shown in this subsection that, if both the processes $\{\mathcal{X}_t\}, \{\mathcal{Y}_t\}$ are Markov, then not only does the limit exist, but it is easy to compute in closed form.

Theorem 7.10 *Suppose $\{\mathcal{X}_t\}$ is a Markov process with stationary distribution $\pi \in \mathbb{S}_n$ and state transition matrix $A \in [0,1]^{n\times n}$, or (π, A)-Markov for short. Similarly, suppose $\{\mathcal{Y}_t\}$ is $(\boldsymbol{\gamma}, B)$-Markov. Then*

$$R(\tilde{Q}\|\tilde{P}) = \sum_{i\in\mathbb{A}} \gamma_i \sum_{j\in\mathbb{A}} b_{ij} \log(b_{ij}/a_{ij}), \tag{7.22}$$

$$= \sum_{i\in\mathbb{A}} \gamma_i H(\mathbf{b}_i\|\mathbf{a}_i). \tag{7.23}$$

where $\mathbf{a}_i, \mathbf{b}_i$ denote the i-row of A, B respectively. An equivalent expression is

$$R(\tilde{Q}\|\tilde{P}) = D(\boldsymbol{\nu}\|\boldsymbol{\mu}) = H(\boldsymbol{\nu}\|\boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}}\|\bar{\boldsymbol{\mu}}), \tag{7.24}$$

where $\boldsymbol{\nu} = [\gamma_i b_{ij}], \boldsymbol{\mu} = [\pi_i a_{ij}] \in \mathcal{M}_c(\mathbb{A}^2)$ are the doublet frequency vectors of the two Markov processes.

To prove Theorem 7.10, we recall the "chain rule" of relative entropy; see Theorem 3.23.[3]

Lemma 7.11 *Suppose $\mathbb{U} = \{u_1, \ldots, u_r\}, \mathbb{V} = \{v_1, \ldots, v_s\}$ are finite sets, and that $\boldsymbol{\theta}, \boldsymbol{\phi}$ are probability distributions on the product $\mathbb{U} \times \mathbb{V}$. Then*

$$H(\boldsymbol{\theta}\|\boldsymbol{\phi}) = H(\boldsymbol{\theta}_{\mathbb{U}}\|\boldsymbol{\phi}_{\mathbb{U}}) + \sum_{i\in\mathbb{U}} (\boldsymbol{\theta}_{\mathbb{U}})_i H(\boldsymbol{\theta}_{\mathbb{V}|i}\|\boldsymbol{\phi}_{\mathbb{V}|i}). \tag{7.25}$$

Here $\boldsymbol{\phi}_{\mathbb{U}}, \boldsymbol{\theta}_{\mathbb{U}}$ are marginal distributions on $\mathbb{U}$ while $\boldsymbol{\phi}_{\mathbb{V}|i}, \boldsymbol{\theta}_{\mathbb{V}|i}$ are conditional distributions on $\mathbb{V}$ conditioned on the event i. Thus

$$(\boldsymbol{\phi}_{\mathbb{U}})_i = \sum_{j=1}^{s} \boldsymbol{\phi}_{ij}, (\boldsymbol{\phi}_{\mathbb{V}|i})_j = \frac{\phi_{ij}}{\sum_{j'=1}^{s} \phi_{ij'}},$$

and $\boldsymbol{\theta}_{\mathbb{U}}, \boldsymbol{\theta}_{\mathbb{V}|i}$ are defined analogously.

The next result is stated separately because it *does not require* the two processes to be Markov, and is thus of independent interest.

[3] Note that the notation has been changed slightly since the symbol $\mathbb{A}$ has a very specific meaning throughout this chapter.

Lemma 7.12 *Given probability laws $\tilde{P}, \tilde{Q}$ on $\mathbb{A}^\infty$, as before let $\boldsymbol{\phi}^{(l)}, \boldsymbol{\theta}^{(l)}$ denote the joint distributions of the first l coordinates under the two laws, respectively. Now define*

$$\alpha_l := \frac{1}{l} H(\boldsymbol{\theta}^{(l)} \| \boldsymbol{\phi}^{(l)})$$

$$\begin{aligned} \beta_l &:= H(\boldsymbol{\theta}^{(l+1)} \| \boldsymbol{\phi}^{(l+1)}) - H(\boldsymbol{\theta}^{(l)} \| \boldsymbol{\phi}^{(l)}) \\ &= D(\boldsymbol{\theta}^{(l+1)} \| \boldsymbol{\phi}^{(l+1)}), l \geq 1. \end{aligned} \tag{7.26}$$

Then the following statements are equivalent:

1. *The Kullback-Leibler divergence rate $R(\tilde{Q} \| \tilde{P})$ is well-defined.*
2. *The sequence $\{\alpha_l\}$ converges as $l \to \infty$.*
3. *The sequence $\{\beta_l\}$ converges in the Césaro sense as $l \to \infty$.*

Therefore the Kullback-Leibler divergence rate $R(\tilde{Q} \| \tilde{P})$, if it exists, is the Césaro limit of the sequence $\{\beta_l\}$.

Remark: Recall that a sequence of real numbers $\{z_l\}$ is said to **converge in the sense of Césaro** if the sequence of averages $\{y_l\}$ converges, where

$$y_l = \frac{1}{l} \sum_{i=1}^{l} z_i.$$

For instance, the sequence $\{z_l = (-1)^l\}$ does not converge in the conventional sense because it keeps oscillating between 1 and -1. However, it converges in sense of Césaro to 0. Césaro converges is a weaker property than standard convergence because if $z_l \to z^*$, then $y_l \to z^*$ as $l \to \infty$. However, as the above example illustrates, it is possible for y_l to approach a limit even when z_l fails to do so.

Proof. It is a direct consequence of (7.25) that

$$\begin{aligned} l\alpha_l = H(\boldsymbol{\theta}^{(l)} \| \boldsymbol{\phi}^{(l)}) &= H(\boldsymbol{\theta}^{(1)} \| \boldsymbol{\phi}^{(1)}) + \sum_{i=1}^{l-1} H(\boldsymbol{\theta}^{(l+1)} \| \boldsymbol{\phi}^{(l+1)}) \\ &= \sum_{i=1}^{l-1} \beta_i + H(\boldsymbol{\theta}^{(1)} \| \boldsymbol{\phi}^{(1)}). \end{aligned}$$

Therefore

$$\alpha_l = \frac{l-1}{l} \frac{1}{l-1} \sum_{i=1}^{l-1} \beta_i + \frac{H(\boldsymbol{\theta}^{(1)} \| \boldsymbol{\phi}^{(1)})}{l}.$$

Since $H(\boldsymbol{\theta}^{(1)} \| \boldsymbol{\phi}^{(1)})$ is just some constant, the second term on the right side approaches zero as $l \to \infty$. Since $(l-1)/ \to 1$ as $l \to \infty$, the first term approaches the Césaro limit of β_l if any. Hence α_l has a limit if and only if β_l has a Césaro limit. □

Proof. **of Theorem 7.10:** The proof of (7.22) consists of showing that the sequence $\{\beta_l\}$ defined in (7.26) converges in just one step. Suppose $l \geq 2$, and apply the definition (7.26) for β_l. This gives

$$\beta_l = \sum_{\mathbf{i} \in \mathbb{A}^l} (\boldsymbol{\theta})_{\mathbf{i}} H(\boldsymbol{\theta}_{\mathbb{A}|\mathbf{i}} \| \boldsymbol{\phi}_{\mathbb{A}|\mathbf{i}}).$$

Let us expand these conditional probabilities, keeping in mind that the two processes are Markov. Suppose $\mathbf{i} = i_1 \ldots i_l, j \in \mathbb{A}$. Then

$$\begin{aligned}(\boldsymbol{\theta}_{\mathbb{A}|\mathbf{i}})_j &= \Pr\{\mathcal{X}_{l+1} = j | \mathcal{X}_1^l = i_1 \ldots i_l\} \\ &= \Pr\{\mathcal{X}_{l+1} = j | \mathcal{X}_l = i_l\} \text{ by the Markov property} \\ &= b_{i_l j}.\end{aligned}$$

Similarly

$$(\boldsymbol{\phi}_{\mathbb{A}|\mathbf{i}})_j = a_{i_l j}.$$

Therefore

$$\beta_l = \sum_{\mathbf{i} \in \mathbb{A}^l} (\boldsymbol{\theta})_{\mathbf{i}} \sum_{j \in \mathbb{A}} b_{i_l j} \log \frac{b_{i_l j}}{a_{i_l j}}.$$

Next, let us partition $\mathbf{i} \in \mathbb{A}^l$ as $\mathbf{i} = \mathbf{i}_1 i_l$ where $\mathbf{i}_1 \in \mathbb{A}^{l-1}, i_l \in \mathbb{A}$. Then the above expression becomes

$$\begin{aligned}\beta_l &= \sum_{\mathbf{i}_1 \in \mathbb{A}^{l-1}} \sum_{i_l \in \mathbb{A}} (\boldsymbol{\theta})_{\mathbf{i}_1 i_l} \sum_{j \in \mathbb{A}} b_{i_l j} \log \frac{b_{i_l j}}{a_{i_l j}} \\ &= \sum_{i_l \in \mathbb{A}} \left[\sum_{\mathbf{i}_1 \in \mathbb{A}^{l-1}} (\boldsymbol{\theta})_{\mathbf{i}_1 i_l} \right] \sum_{j \in \mathbb{A}} b_{i_l j} \log \frac{b_{i_l j}}{a_{i_l j}} \\ &= \sum_{i_l \in \mathbb{A}} \gamma_{i_l} \sum_{j \in \mathbb{A}} b_{i_l j} \log \frac{b_{i_l j}}{a_{i_l j}}.\end{aligned}$$

In the last step we use the obvious fact that

$$\sum_{\mathbf{i}_1 \in \mathbb{A}^{l-1}} (\boldsymbol{\theta})_{\mathbf{i}_1 i_l} = \gamma_{i_l}, \ \forall i_l \in \mathbb{A}.$$

Since β_l is given by the above formula for *every* $l \geq 2$, the sequence $\{\beta_l\}$ converges in two steps. Replacing the dummy index of summation i_l by i leads to (7.22). The equivalence of (7.22) and (7.23) is obvious.

Finally, to show that these formulae are both equivalent to (7.24), observe that

$$\gamma_i = \bar{\nu}_i, b_{ij} = \frac{\nu_{ij}}{\bar{\nu}_i}, \pi_i = \bar{\mu}_i, a_{ij} = \frac{\mu_{ij}}{\bar{\mu}_i}.$$

Therefore it follows from (7.22) that

$$\begin{aligned} R(\tilde{Q}\|\tilde{P}) &= \sum_{i\in\mathbb{A}} \bar{\nu}_i \sum_{j\in\mathbb{A}} \frac{\nu_{ij}}{\bar{\nu}_i} \log \frac{\nu_{ij}/\bar{\nu}_i}{\mu_{ij}/\bar{\mu}_i} \\ &= \sum_{i\in\mathbb{A}} \sum_{j\in\mathbb{A}} \nu_{ij} \log \frac{\nu_{ij}/\bar{\nu}_i}{\mu_{ij}/\bar{\mu}_i} \\ &= \sum_{i\in\mathbb{A}} \sum_{j\in\mathbb{A}} \nu_{ij} \log \frac{\nu_{ij}}{\mu_{ij}} - \sum_{i\in\mathbb{A}} \left[\sum_{j\in\mathbb{A}} \nu_{ij} \right] \log \frac{\bar{\nu}_i}{\bar{\mu}_i} \\ &= H(\boldsymbol{\nu}\|\boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}}\|\bar{\boldsymbol{\mu}}) \end{aligned}$$

because

$$\sum_{j\in\mathbb{A}} \nu_{ij} = \bar{\nu}_i.$$

□

To complete the discussion, suppose the laws $\tilde{P}, \tilde{Q}$ correspond to s-step Markov chains, and observe in passing that an s-step Markov chain is also an s'-step Markov chain whenever $s' > s$. Thus if $\tilde{P}, \tilde{Q}$ correspond to multi-step Markov chains but possibly with different memory lengths, we can choose s to be the larger of the memories of the two processes. In such a case, the sequence $\{\beta_l\}$ converges after s steps. The proof is the same as that for Theorem 7.10 with suitable modifications.

Theorem 7.13 *Suppose $\tilde{P}, \tilde{Q}$ correspond to s-step Markov chains, and let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$ denote the corresponding $(s+1)$-tuple frequency vectors under the laws $\tilde{P}, \tilde{Q}$ respectively. Then*

$$\beta_l = \beta_s = D(\boldsymbol{\nu}\|\boldsymbol{\mu}),\ \forall l \geq s.$$

Thus the sequence $\{\beta_l\}$ converges after s steps, and as a result

$$R(\tilde{Q}\|\tilde{P}) = D(\boldsymbol{\nu}\|\boldsymbol{\mu}) = H(\boldsymbol{\nu}\|\boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}}\|\bar{\boldsymbol{\mu}}). \tag{7.27}$$

The proof is omitted and is left as an exercise.

7.3.3 The Rate Function for Doubleton Frequencies

Suppose $\{\mathcal{X}_t\}$ is a Markov process assuming values in a finite set $\mathbb{A}$. Given an observation $x_1^l = x_1 \ldots x_l$, we can form an empirical distribution $\boldsymbol{\phi}(x_1^l) \in \mathbb{S}_n$ in analogy with (7.1); that is,

$$\phi_j(x_1^l) := \frac{1}{l} \sum_{t=1}^{l} I_{\{x_t = j\}},\ \forall j \in \mathbb{A}. \tag{7.28}$$

Thus $\boldsymbol{\phi}$ is an approximation to the stationary distribution $\boldsymbol{\pi}$ of the Markov chain.

As shown earlier, the Markov process $\{\mathcal{X}_t\}$ is completely characterized by its vector of doublet frequencies $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^2)$. If $\{\mathcal{X}_t\}$ is a Markov chain, so is the stochastic process consisting of doublets $\{(\mathcal{X}_t, \mathcal{X}_{t+1})\}$. So, given a sample path $\mathbf{x}_1^l$, we can estimate the vector of doublet frequencies using this sample path. Since this point is germane to the subsequent discussion, it is worth describing how precisely the doublet frequency vector is estimated. Given the sample path $x_1^l = x_1 \ldots x_l$, we can define

$$\theta_{ij}(x_1^l) := \frac{1}{l-1} \sum_{t=1}^{l-1} I_{\{\mathcal{X}_t \mathcal{X}_{t+1} = ij\}}. \tag{7.29}$$

This procedure produces a vector $\boldsymbol{\theta} \in \mathbb{S}_{n^2}$ which is a measure on $\mathbb{A}^2$, and can be interpreted as an empirical estimate for $\boldsymbol{\mu}$, the true but unknown vector of doublet frequencies. The difficulty however is that the distribution $\boldsymbol{\theta}$ is *not consistent* in general. If we define $\bar{\boldsymbol{\theta}} \in \mathbb{S}_n$ by

$$\bar{\theta}_i := \frac{1}{l-1} \sum_{t=1}^{l-1} I_{\{\mathcal{X}_t = i\}},$$

then it is certainly true that

$$\bar{\theta}_i = \sum_{j=1}^{n} \theta_{ij}.$$

However, in general

$$\sum_{j=1}^{n} \theta_{ji} \neq \bar{\theta}_i.$$

Hence $\boldsymbol{\theta} \in \mathbb{S}_{n^2}$ is *not* a consistent distribution in general. Moreover, there is no simple relationship between $\bar{\boldsymbol{\theta}} \in \mathbb{S}_n$ and $\boldsymbol{\phi} \in \mathbb{S}_n$ defined in (7.28).

On the other hand, if $\mathbf{x}_l = x_1$ so that the sample path is a cycle, then $\boldsymbol{\theta} \in \mathcal{M}_c(\mathbb{A}^2)$. This suggests that we should use only cyclic sample paths to construct empirical estimates of doublet frequencies, or to carry the argument a bit farther, that we must *create cycles*, artificially if necessary, in the sample path. Accordingly, given a sample path $x_1^l = x_1 \ldots x_l$, we construct the empirical estimate $\boldsymbol{\nu} = \boldsymbol{\nu}(x_1^l)$ as follows:

$$\nu_{ij}(x_1^l) := \frac{1}{l} \sum_{t=1}^{l} I_{\{x_t x_{t+1} = ij\}}, \ \forall i, j \in \mathbb{A}, \tag{7.30}$$

where x_{l+1} is taken as x_1. If we compare (7.30) with (7.29), we see that we have in effect augmented the original sample path x_1^l by adding a “ghost” transition from x_l back to x_1 so as to create a cycle, and used this artificial sample path of length $l+1$ to construct the empirical estimate. The advantage of doing so is that the resulting vector $\boldsymbol{\nu}$ is *always consistent*, unlike $\boldsymbol{\theta}$ in (7.29) which may not be consistent in general.

At this juncture it is worth pointing out that *it is entirely up to us* just how we go about constructing the empirical estimate on the basis of a given a sample path. Thus $\boldsymbol{\nu}$ is as “valid” an estimate as $\boldsymbol{\theta}$. Moreover, if $\boldsymbol{\nu}$ is a better-behaved estimate than $\boldsymbol{\theta}$, then we should use it!

It should be intuitively obvious that $\boldsymbol{\nu}(x_1^l)$ is consistent, but we show this formally.

Lemma 7.14 *The measure $\boldsymbol{\nu}(x_1^l) \in \mathcal{M}(\mathbb{A}^2)$ defined in (7.30) is consistent and thus belongs to $\mathcal{M}_c(\mathbb{A}^2)$.*

Proof. Let us define

$$\phi_i := \sum_{j \in \mathbb{A}} \nu_{ij}, \ \forall i \in \mathbb{A}. \tag{7.31}$$

Then clearly $\boldsymbol{\phi} \in \mathbb{S}_n$. Now it is claimed that

$$\sum_{j \in \mathbb{A}} \nu_{ji} = \phi_i, \ \forall i \in \mathbb{A}, \tag{7.32}$$

thus showing that $\boldsymbol{\nu}$ is consistent. To establish (7.32), observe from (7.31) that $\boldsymbol{\phi}$ is obtained by counting the number of times that i occurs as the first symbol in $x_1x_2, x_2x_3, \ldots, x_{l-1}x_l, x_lx_1$, and then dividing by l. Similarly the quantity $\sum_{j \in \mathbb{A}} \nu_{ji}$ is obtained by counting the number of times that i occurs as the second symbol in $x_1x_2, x_2x_3, \ldots, x_{l-1}x_l, x_lx_1$, and then dividing by l. Now, for $2 \leq t \leq l-1$, x_t is the first symbol in x_tx_{t+1}, and the second symbol in $x_{t-1}x_t$. Next, x_1 is the first symbol in x_1x_2 and the second symbol in the ghost transition x_lx_1. Similarly x_l is the second symbol in $x_{l-1}x_l$ and the first symbol in the ghost transition x_lx_1. Thus $\boldsymbol{\nu}$ is consistent. □

To derive the rate function for this situation, we again use the method of types. For a given integer l, the sample space of all possible sample paths is clearly $\mathbb{A}^l$. With each sample path $x_1^l \in \mathbb{A}^l$, we associate a corresponding empirical distribution $\hat{\boldsymbol{\mu}}(x_1^l)$ defined as $\boldsymbol{\nu}(\mathbf{x}_1^l)$ of (7.30). We again define $x_1^l, y_1^l \in \mathbb{A}^l$ to be equivalent if they lead to the same empirical distribution, that is, if $\hat{\boldsymbol{\mu}}(x_1^l) = \hat{\boldsymbol{\mu}}(y_1^l)$. Let us define $\mathcal{E}(l, n, 2)$ to be the subset of $\mathcal{M}_c(\mathbb{A}^2)$ that can be generated as empirical measures from a sample path of length l over an alphabet of size n. Note that earlier we had introduced the symbol $\mathcal{E}(l, n)$ for the set of all empirical measures in $\mathbb{S}_n$ that can be generated from a sample path of length l over an alphabet of size n. So it is clear that $\mathcal{E}(l, n, 2) \subseteq \mathcal{E}(l, n^2)$. However, not every empirical measure in $\mathcal{E}(l, n^2)$ will be *consistent*. This is why we introduce a new symbol $\mathcal{E}(l, n, 2)$.

As before, for each $\boldsymbol{\nu} \in \mathcal{E}(l, n, 2)$, define

$$T(\boldsymbol{\nu}, l) := \{x_1^l \in \mathbb{A}^l : \hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\}.$$

Then $T(\boldsymbol{\nu}, l) \subseteq \mathbb{A}^l$ is once again called the **type class** of $\boldsymbol{\nu}$. We will once again address the following questions (in no particular order):

- What is the cardinality of $\mathcal{E}(l, n, 2)$? In other words, how many distinct *consistent* empirical measures $\hat{\boldsymbol{\mu}}(x_1^l)$ can be generated as x_1^l varies over $\mathbb{A}^l$?

- For a given $\boldsymbol{\nu} \in \mathcal{E}(l, n, 2)$, what is the cardinality of the associated type class $T(\boldsymbol{\nu}, l)$?

- What is the (log) likelihood of each sample path in $T(\boldsymbol{\nu}, l)$, and how is it related to $\boldsymbol{\nu}$?

We have seen in Section 7.2 that, if the process is i.i.d., and we estimate one-dimensional marginal using (7.1), then every sample path in $T(\boldsymbol{\nu}, l)$ has exactly the same (log) likelihood. However, if the process is not i.i.d., then this statement is no longer true: Different sample paths in $T(\boldsymbol{\nu}, l)$ can have different (log) likelihoods. Nevertheless, it is possible to adapt the arguments from Section 7.2 to derive a rate function for the present situation.

We state at once the main result of this subsection. The proof is given in stages.

Theorem 7.15 *Suppose $\{\mathcal{X}_t\}$ is a stationary Markov chain assuming values in a finite alphabet $\mathbb{A}$. Let $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^2)$ denote the vector of doublet frequencies corresponding to this Markov chain, and let $\boldsymbol{\nu}(x_1^l) \in \mathcal{M}_c(\mathbb{A}^2)$ denote the empirical distribution constructed as in (7.30). Suppose $\mu_{ij} > 0 \ \forall i, j \in \mathbb{A}$. Then the $\mathcal{M}_c(\mathbb{A}^2)$-valued process $\{\boldsymbol{\nu}(x_1^l)\}$ satisfies the LDP with the rate function*

$$I(\boldsymbol{\nu}) := D(\boldsymbol{\nu}\|\boldsymbol{\mu}) = H(\boldsymbol{\nu}\|\boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}}\|\bar{\boldsymbol{\mu}}). \tag{7.33}$$

The proof of the theorem is given through a couple of preliminary lemmas. Each $\boldsymbol{\nu} \in \mathcal{E}(l, n, 2)$ is of the form $\nu_{ij} = l_{ij}/l$ for some integer l_{ij}. Moreover, the corresponding reduced distribution $\bar{\boldsymbol{\nu}}$ over $\mathbb{A}$ is given by $\bar{\nu}_i = \bar{l}_i/l$ where

$$\bar{l}_i = \sum_{j=1}^{n} l_{ij} = \sum_{j=1}^{n} l_{ji}, \ \forall i.$$

Throughout the proof, l denotes the length of the sample path and $l_{ij}, \bar{l}_i$ denote these integers.

Lemma 7.16 *With all notation as above, the following statements hold:*

1. *For each l, we have*
$$|\mathcal{E}(l, n, 2)| \leq (l+1)^{n^2}. \tag{7.34}$$

2. *The countable set $\cup_l \mathcal{E}(l, n, 2)$ is dense in $\mathcal{M}_c(\mathbb{A}^2)$.*

Proof. Suppose $\boldsymbol{\nu} \in \mathcal{E}(l, n, 2)$. Then each component ν_{ij} has $l+1$ possible values, namely $0/l, 1/l, \ldots, (l-1)/l, l/l$. Thus the maximum number of possible vectors in $\mathcal{E}(l, n, 2)$ is $(l+1)^{n^2}$. This proves (7.34) and establishes the first claim.

Let $\mathcal{M}_c(\mathbb{A}^2)^{(l)}$ denote the subset of $\mathcal{M}_c(\mathbb{A}^2)$ consisting of those distributions $\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)$ where each entry is a rational number of the form l_{ij}/l. Now it is obvious that the countable set $\cup_l \mathcal{M}_c(\mathbb{A}^2)^{(l)}$ is dense in $\mathcal{M}_c(\mathbb{A}^2)$. This is because every real number can be approximated arbitrarily closely by a rational number. The proof of the second claim is complicated by the fact that $\mathcal{E}(l, n, 2)$ is a *strict subset* of $\mathcal{M}_c(\mathbb{A}^2)^{(l)}$. To see this, choose *positive* integers $l_1, \ldots, l_n$ that add up to l, and let

$$\nu_{ij} = \begin{cases} l_i/l & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Then $\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)^{(l)}$ as can be easily verified: It is consistent, and every entry is a rational number with denominator l. However $\boldsymbol{\nu} \notin \mathcal{E}(l, n, 2)$, because in any sample path x_1^l that contains each of the n symbols of $\mathbb{A}$, there will be at least one pair ij where $i \neq j$. Hence $\boldsymbol{\nu}(x_1^l)$ cannot have all 'off-diagonal' terms equal to zero. Nevertheless, the off-diagonal entries can be made arbitrarily small by choosing l sufficiently large. Thus $\cup_l \mathcal{E}(l, n, 2)$ is dense in $\cup_l \mathcal{M}_c(\mathbb{A}^2)^{(l)}$, which in turn is dense in $\mathcal{M}_c(\mathbb{A}^2)$. This establishes the second claim of the lemma. □

The next lemma is much more crucial.

Lemma 7.17 *Suppose $\boldsymbol{\nu} \in \mathcal{E}(l, n, 2)$. Then the cardinality of the type class $T(\boldsymbol{\nu}, l)$ is bounded by*

$$(el)^{-2n^2} \exp[lD(\boldsymbol{\nu})] \leq |T(\boldsymbol{\nu}, l)| \leq l \exp[lD(\boldsymbol{\nu})]. \tag{7.35}$$

Proof. Suppose $\boldsymbol{\nu} \in \mathbb{S}_{n^2}$ is a distribution on $\mathbb{A}^2$, *not necessarily consistent*, such that every component of $l\boldsymbol{\nu}$ is an integer for some integer l (or equivalently, every entry in $\boldsymbol{\nu}$ is of the form l_{ij}/l for some integers l_{ij}, l). Then we can associate a directed graph with $\boldsymbol{\nu}$ as follows: The graph has n nodes labelled as 1 through n, an $l\nu_{ij}$ edges from node i to node j. With this definition, it is easy to see that $\sum_{j \in \mathbb{A}} l\nu_{ji}$ is the in-degree of node i, while $\sum_{j \in \mathbb{A}} l\nu_{ij}$ is the out-degree of node i. Therefore the distribution $\boldsymbol{\nu}$ is consistent, and thus belongs to $\mathcal{M}_c(\mathbb{A}^2)^{(l)}$, if and only if each node has the same in-degree and out-degree. Further, $\boldsymbol{\nu}$ belongs to $\mathcal{E}(l, n, 2)$, and not just to $\mathcal{M}_c(\mathbb{A}^2)^{(l)}$, if and only if (i) the graph is strongly connected, and (ii) each node has the same in-degree and out-degree. This is why a measure $\boldsymbol{\nu}$ with $\nu_{ij} = 0$ for all $i \neq j$ cannot belong to $\mathcal{E}(l, n, 2)$ – it satisfies the second criterion but not the first. Therefore such a distribution can belong to $\mathcal{M}_c(\mathbb{A}^2)^{(l)}$ but not to $\mathcal{E}(l, n, 2)$. If the graph is also connected in addition to having the property that each node has the same in-degree and out-degree, then the number of distinct sample paths that generate this empirical distribution $\boldsymbol{\nu}$ is equal to the number of distinct Eulerian circuits in the graph. Recall that, in this connection, an Eulerian circuit is a loop that uses each edge exactly once (and thus traverses each node at least once). Thus, by counting the number of Eulerian circuits, we can compute the cardinality of the type class $|T(\boldsymbol{\nu}, l)|$. This is what is done in [59]. □

At last we come to the proof of the main theorem.

Proof. **of Theorem 7.15:** Suppose we have a sample path x_1^l. Let us compute its likelihood in terms of the properties of the corresponding empirical distribution $\boldsymbol{\nu}(x_1^l)$. We have

$$\Pr\{\mathcal{X}_1^l = x_1^l\} = \Pr\{\mathcal{X}_1 = x_1\} \cdot \prod_{t=1}^{l-1} \Pr\{\mathcal{X}_{t+1} = x_{t+1} | \mathcal{X}_t = x_t\}.$$

Hence[4]

$$\begin{aligned}\Pr\{\mathcal{X}_1^l = x_1^l\} &= \bar{\mu}(x_1) \cdot \prod_{t=1}^{l-1} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)} \\ &= \bar{\mu}(x_1) \cdot \prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)} \cdot \frac{\bar{\mu}(x_l)}{\mu(x_l x_1)} \\ &= \frac{\bar{\mu}(x_1)\bar{\mu}(x_l)}{\mu(x_l x_1)} \cdot \prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)}, \end{aligned} \tag{7.36}$$

where as before we take $x_{l+1} = x_1$. Now, since $\mu_{ij} > 0$ for all i, j, there exist constants $\underline{c}$ and $\bar{c}$ such that

$$\underline{c} \leq \frac{\bar{\mu}_i \bar{\mu}_j}{\mu_{ij}} \leq \bar{c}, \ \forall i, j.$$

Of course these constants depend on $\boldsymbol{\mu}$, but the point is that they do not depend on the empirical measure $\boldsymbol{\nu}(x_1^l)$.

Next we examine the product term in (7.36). We have

$$\log \left[\prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)} \right] = \sum_{t=1}^{l} [\log \mu(x_t x_{t+1}) - \log \bar{\mu}(x_t)].$$

When we do the above summation, we observe that the pair $x_t x_{t+1}$ occurs exactly $l_{ij} = l[\boldsymbol{\nu}(x_1^l)]_{ij}$ times, while x_t occurs exactly $\bar{l}_i = l[\bar{\boldsymbol{\nu}}(x_1^l)]_i$ times. Therefore

$$\begin{aligned}\log \left[\prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)} \right] &= l \sum_{i \in \mathbb{A}} \sum_{j \in \mathbb{A}} \nu_{ij} \log \mu_{ij} - l \sum_{i \in \mathbb{A}} \bar{\nu}_i \log \bar{\mu}_i \\ &= -l[J(\boldsymbol{\nu}, \boldsymbol{\mu}) - J(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}})],\end{aligned}$$

where we write $\boldsymbol{\nu}$ and $\bar{\boldsymbol{\nu}}$ for the more precise $\boldsymbol{\nu}(x_1^l)$ and $\bar{\boldsymbol{\nu}}(x_1^l)$. Substituting this into (7.36) shows that the likelihood of each sample path can be bounded as follows:

$$\log \Pr\{\mathcal{X}_1^l = x_1^l\} + l[J(\boldsymbol{\nu}, \boldsymbol{\mu}) - J(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}})] \in [\log \underline{c}, \log \bar{c}]. \tag{7.37}$$

In large deviation theory, the quantity of interest is the log of the likelihood that a particular empirical estimate will occur, normalized by the length of the observation. Accordingly, let us denote the empirical distribution generated by a sample path as $\hat{\boldsymbol{\mu}}(x_1^l)$, and define

$$\delta(l, \boldsymbol{\nu}) := \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\}.$$

Now we know from (7.37) that the log likelihood of each sample path within the type class $T(\boldsymbol{\nu}, l)$ looks like $l[J(\boldsymbol{\nu}, \boldsymbol{\mu}) - J(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}})]$, and we know from (7.35)

[4] In the interests of clarity, in the proof we write $\mu(x_s x_t)$ instead of $\mu_{x_s x_t}$, and $\bar{\mu}(x_t)$ instead of $\bar{\mu}_{x_t}$. However, we continue to use the subscript notation if the arguments are simple indices such as i and j.

that $\log |T(\boldsymbol{\nu})|$ looks like $lD(\boldsymbol{\nu})$. Combining these two facts leads to

$$\begin{aligned}\delta(l,\boldsymbol{\nu}) &\leq D(\boldsymbol{\nu}) - J(\boldsymbol{\nu},\boldsymbol{\mu}) + J(\bar{\boldsymbol{\nu}},\bar{\boldsymbol{\mu}}) + o(1/l)\\ &= H(\boldsymbol{\nu}) - H(\bar{\boldsymbol{\nu}}) - J(\boldsymbol{\nu},\boldsymbol{\mu}) + J(\bar{\boldsymbol{\nu}},\bar{\boldsymbol{\mu}}) + o(1/l)\\ &= -H(\boldsymbol{\nu}\|\boldsymbol{\mu}) + H(\bar{\boldsymbol{\nu}}\|\bar{\boldsymbol{\mu}}) + o(1/l)\\ &= -D(\boldsymbol{\nu}\|\boldsymbol{\mu}) + o(1/l). \end{aligned} \tag{7.38}$$

Similarly we get

$$\delta(l,\boldsymbol{\nu}) \geq -D(\boldsymbol{\nu}\|\boldsymbol{\mu}) + o(1/l). \tag{7.39}$$

The remainder of the proof is entirely analogous to that of Theorem 7.6. Let $\Gamma \subseteq \mathcal{M}_c(\mathbb{A}^2)$ be any set of consistent distributions on $\mathbb{A}^2$. Then

$$\begin{aligned}\Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} &= \sum_{\boldsymbol{\nu}\in\mathcal{E}(l,n,2)\cap\Gamma} \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\}\\ &\leq |\mathcal{E}(l,n,2)\cap\Gamma| \sup_{\boldsymbol{\nu}\in\mathcal{E}(l,n,2)\cap\Gamma} \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) = \boldsymbol{\nu}\}.\end{aligned}$$

Hence

$$\frac{1}{l}\log \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} \leq \frac{1}{l}\log|\mathcal{E}(l,n,2)| + \sup_{\boldsymbol{\nu}\in\Gamma}\delta(l,\boldsymbol{\nu}).$$

Since $|\mathcal{E}(l,n,2)|$ is polynomial in l, the first term approaches zero as $l \to \infty$. Next, from (7.39) it follows that the second term approaches $-D(\boldsymbol{\nu}\|\boldsymbol{\mu})$ as $l \to \infty$. Combining these two facts shows that

$$\limsup_{l\to\infty} \frac{1}{l}\log \Pr\{\hat{\boldsymbol{\mu}}(x_1^l) \in \Gamma\} \leq \sup_{\boldsymbol{\nu}\in\Gamma} -D(\boldsymbol{\nu}\|\boldsymbol{\mu}) = -\inf_{\boldsymbol{\nu}\in\Gamma} D(\boldsymbol{\nu}\|\boldsymbol{\mu}).$$

This establishes the right inequality in (7.4).

To establish the left inequality, suppose $\boldsymbol{\nu}$ is an interior point of Γ. Then there is an open ball $\mathcal{B}(\boldsymbol{\nu})$ in $\mathcal{M}_c(\mathbb{A}^2)$ that contains $\boldsymbol{\nu}$. Since $\cup_l \mathcal{E}(l,n,2)$ is dense in $\mathcal{M}_c(\mathbb{A}^2)$, there exist a sequence of integers l_k and corresponding elements $\boldsymbol{\nu}_{l_k} \in \mathcal{E}(l_k,n,2)\cap\Gamma$ such that $l_k \to \infty$ and $\boldsymbol{\nu}_{l_k} \to \boldsymbol{\nu}$. Hence

$$\Pr\{\hat{\boldsymbol{\mu}}(x_1^{l_k}) \in \Gamma\} \geq \Pr\{\hat{\boldsymbol{\mu}}(x_1^{l_k}) = \boldsymbol{\nu}_{l_k}\},$$

$$\begin{aligned}\frac{1}{l_k}\log \Pr\{\hat{\boldsymbol{\mu}}(x_1^{l_k}) \in \Gamma\} &\geq \delta(l_k,\boldsymbol{\nu}_{l_k})\\ &\geq -D(\boldsymbol{\nu}_{l_k}\|\boldsymbol{\mu}) + o(1/l_k)\\ &\to -D(\boldsymbol{\nu}\|\boldsymbol{\mu}) \text{ as } k\to\infty.\end{aligned}$$

Hence it follows that

$$\liminf_{l\to\infty} \frac{1}{l}\log \Pr\{\boldsymbol{\nu}(x_1^l) \in \Gamma\} \geq -D(\boldsymbol{\nu}\|\boldsymbol{\mu}),\ \forall \boldsymbol{\nu} \in \Gamma^o.$$

Since the above inequality holds for every $\boldsymbol{\nu} \in \Gamma^o$, we can conclude that

$$\liminf_{l\to\infty} \frac{1}{l}\log \Pr\{\boldsymbol{\nu}(x_1^l) \in \Gamma\} \geq \sup_{\boldsymbol{\nu}\in\Gamma^o} -D(\boldsymbol{\nu}\|\boldsymbol{\mu}) = -\inf_{\boldsymbol{\nu}\in\Gamma^o} D(\boldsymbol{\nu}\|\boldsymbol{\mu}).$$

This establishes that the relationships in (7.4) hold with $I(\boldsymbol{\nu}) = D(\boldsymbol{\nu}\|\boldsymbol{\mu})$, thus showing that $D(\boldsymbol{\nu}\|\boldsymbol{\mu})$ is the rate function. □

In conclusion, it may be remarked that when the samples $\{x_t\}$ come from an i.i.d. process, we have exact formulae for both the size of each type class, and the likelihood of each sample within a type class. In the case where the samples come from a Markov process, we have only bounds. *However*, the 'correction terms' in these bounds approach zero as the number of samples approaches infinity, thus allowing us to deduce the rate function for doubleton frequencies in a straight-forward fashion.

7.3.4 The Rate Function for Singleton Frequencies

In this subsection, we first introduce a very important technique known as the 'contraction principle', which permits us to derive rate functions for *functions* of the empirically estimated frequencies. Using the contraction principle, we then derive the rate function for singleton frequencies of a Markov chain. The contraction principle directly leads to a very appealing formula. By applying duality theory, we then derive another formula that is equivalent to this one.

Theorem 7.18 (The Contraction Principle) *Suppose the stochastic process $\{\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)\}$ assuming values in $\mathbb{S}_m$ satisfies the large deviation property with the rate function $I(\cdot) : \mathbb{S}_m \to \mathbb{R}_+$. Suppose that $f : \mathbb{S}_m \to \mathbb{S}_k$ is continuous. Then the stochastic process $\{f[\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)]\}$ satisfies the large deviation property with the rate function $J(\cdot) : \mathbb{S}_k \to \mathbb{R}_+$ defined by*

$$J(\boldsymbol{\theta}) := \inf_{\boldsymbol{\nu} \in f^{-1}(\boldsymbol{\theta})} I(\boldsymbol{\nu}), \ \forall \boldsymbol{\theta} \in \mathbb{S}_k. \tag{7.40}$$

Remarks:

1. The rate function for the new stochastic process $\{f[\hat{\boldsymbol{\mu}}(\mathbf{x}_1^l)]\}$ has a very intuitive interpretation. The rate function of $\boldsymbol{\theta} \in \mathbb{S}_k$ is the 'slowest' value amongst the original rate function $I(\boldsymbol{\nu})$ as $\boldsymbol{\nu}$ varies over all preimages of $\boldsymbol{\theta}$, that is, over all $\boldsymbol{\nu} \in \mathbb{S}_m$ that map into $\boldsymbol{\theta}$.

2. In the general case, where the stochastic process $\{\mathcal{X}_l\}$ assumes values in $\mathbb{R}^m$, and $f : \mathbb{R}^m \to \mathbb{R}^k$, it is necessary to add the assumption that the original rate function $I(\cdot)$ is 'good', that is, all level sets of $I(\cdot)$ are compact. See for example [35], p. 126, Theorem 4.2.1. However, in the restricted situation being studied here, where the rate function has a compact set as its domain, we need not worry about this additional condition.

Proof. The proof consists of two steps. First, it is shown that the function $J(\cdot)$ is lower semi-continuous. Second, instead of establishing the two relationships in (7.4), we instead establish (7.2) and (7.3). As pointed out earlier, for lower semi-continuous functions, (7.4) is equivalent to (7.2) and (7.3).

To show that the function $J(\cdot)$ is lower semi-continuous, we begin by observing that, since the original rate function $I(\cdot)$ is lower semi-continuous and the set $\mathbb{S}_m$ is compact, the infimum in (7.40) is actually a minimum.[5] Thus for each $\boldsymbol{\theta} \in \mathbb{S}_k$, there exists a $\boldsymbol{\nu} \in \mathbb{S}_m$ such that $f(\boldsymbol{\nu}) = \boldsymbol{\theta}$ and $J(\boldsymbol{\theta}) = I(\boldsymbol{\nu})$. Now suppose $\{\boldsymbol{\theta}_i\}$ is a sequence in $\mathbb{S}_k$ that converges to $\boldsymbol{\theta}^* \in \mathbb{S}_k$. The objective is to show that

$$J(\boldsymbol{\theta}^*) \leq \liminf_{i\to\infty} I(\boldsymbol{\theta}_i). \tag{7.41}$$

Towards this end, let us choose, for each index i, a $\boldsymbol{\nu}_i \in \mathbb{S}_m$ such that $f(\boldsymbol{\nu}_i) = \boldsymbol{\theta}_i$ and $I(\boldsymbol{\nu}_i) = J(\boldsymbol{\theta}_i)$. Now, since $\mathbb{S}_m$ is compact[6] it follows that $\{\boldsymbol{\nu}_i\}$ contains a convergent subsequence. Let us renumber this subsequence again as $\{\boldsymbol{\nu}_i\}$, and let $\boldsymbol{\nu}^* \in \mathbb{S}_m$ denote its limit. Now, since f is a continuous map, it follows that

$$f(\boldsymbol{\nu}^*) = \lim_{i\to\infty} f(\boldsymbol{\nu}_i) = \lim_{i\to\infty} \boldsymbol{\theta}_i = \boldsymbol{\theta}^*.$$

Hence $\boldsymbol{\nu}^* \in f^{-1}(\boldsymbol{\theta}^*)$. The definition of $J(\cdot)$ plus the lower semi-continuity of $I(\cdot)$ together imply that

$$J(\boldsymbol{\theta}^*) = \inf_{\boldsymbol{\nu}\in f^{-1}(\boldsymbol{\theta}^*)} I(\boldsymbol{\nu}) \leq I(\boldsymbol{\nu}^*) \leq \liminf_{i\to\infty} I(\boldsymbol{\nu}_i) = \liminf_{i\to\infty} J(\boldsymbol{\theta}_i).$$

Hence (7.41) is established and $J(\cdot)$ is shown to be lower semi-continuous.

Next, suppose $\Omega \subseteq \mathbb{S}_k$ is an open set. Since the map f is continuous, the preimage $f^{-1}(\Omega)$ is also open. Therefore

$$\Pr\{f(\hat{\boldsymbol{\mu}}_l) \in \Omega\} = \Pr\{\hat{\boldsymbol{\mu}}_l \in f^{-1}(\Omega)\}.$$

As a consequence,

$$\begin{aligned}\liminf_{l\to\infty} \frac{1}{l} \log \Pr\{f(\hat{\boldsymbol{\mu}}_l) \in \Omega\} &= \liminf_{l\to\infty} \frac{1}{l} \log \Pr\{\hat{\boldsymbol{\mu}}_l \in f^{-1}(\Omega)\} \\ &\geq - \inf_{\boldsymbol{\nu}\in f^{-1}(\Omega)} I(\boldsymbol{\nu}) \\ &= - \inf_{\boldsymbol{\theta}\in\Omega} J(\boldsymbol{\theta}).\end{aligned}$$

This establishes that (7.2) holds with I replaced by J and Γ replaced by Ω. The proof of (7.3) is entirely similar and follows upon noting that if $\Omega \subseteq \mathbb{S}_k$ is a closed set, then so is $f^{-1}(\Omega)$. □

To apply the contraction principle to derive the rate function for singleton frequencies, let us define a map $\mathbf{f} : \mathcal{M}_c(\mathbb{A}^2) \to \mathbb{S}_n$ by

$$[\mathbf{f}(\boldsymbol{\nu})]_i := \sum_{j\in\mathbb{A}} \nu_{ij} = \sum_{j\in\mathbb{A}} \nu_{ji}.$$

Thus $\mathbf{f}$ maps a consistent distribution $\boldsymbol{\nu}$ on $\mathbb{A}^2$ onto its one-dimensional marginal $\bar{\boldsymbol{\nu}}$. Moreover, if we construct $\boldsymbol{\nu}(x_1^l)$ for a sample x_1^l using the

[5] In the case of a general stochastic process assuming values in $\mathbb{R}^m$, we invoke the 'goodness' of $I(\cdot)$.

[6] In the general case, since the original rate function $I(\cdot)$ is a 'good' rate function …

formula (7.30), then the corresponding $\mathbf{f}[\boldsymbol{\nu}(x_1^l)]$ is the usual empirical distribution of singleton frequencies. Thus if $\mathbf{f}[\boldsymbol{\nu}(x_1^l)] = \boldsymbol{\phi}(x_1^l)$, then

$$[\boldsymbol{\phi}(x_1^l)]_i = \frac{1}{l} \sum_{t=1}^{l} I_{\{x_t = i\}},$$

which is the same as (7.28). Now, by invoking the contraction principle, we can readily conclude the following:

Theorem 7.19 *The $\mathbb{S}_n$-valued process $\boldsymbol{\phi}(x_1^l)$ satisfies the LDP with the rate function*

$$J(\boldsymbol{\phi}) := \inf_{\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)} D(\boldsymbol{\nu} \| \boldsymbol{\mu}) \ \textit{s.t.} \ \bar{\boldsymbol{\nu}} = \boldsymbol{\phi}. \tag{7.42}$$

Recall that

$$D(\boldsymbol{\nu} \| \boldsymbol{\mu}) = H(\boldsymbol{\nu} \| \boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}} \| \bar{\boldsymbol{\mu}}).$$

Hence we can also write

$$J(\boldsymbol{\phi}) = \left[\inf_{\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)} H(\boldsymbol{\nu} \| \boldsymbol{\mu}) \text{ s.t. } \bar{\boldsymbol{\nu}} = \boldsymbol{\phi} \right] - H(\boldsymbol{\phi} \| \bar{\boldsymbol{\mu}}). \tag{7.43}$$

The problem of minimizing $H(\boldsymbol{\nu} \| \boldsymbol{\mu})$ where $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^2)$ subject to the constraint that $\bar{\boldsymbol{\nu}} = \boldsymbol{\phi}$ is a special case of the following more general problem: Suppose $\mathbb{A}, \mathbb{B}$ are finite sets (not necessarily of the same size), and $\boldsymbol{\mu}$ is a distribution on $\mathbb{A} \times \mathbb{B}$. Suppose $\boldsymbol{\phi}, \boldsymbol{\psi}$ are distributions on $\mathbb{A}, \mathbb{B}$ respectively. Then the problem is: Minimize the relative entropy $H(\boldsymbol{\nu} \| \boldsymbol{\mu})$ subject to the constraints $\boldsymbol{\nu}_{\mathbb{A}} = \boldsymbol{\phi}$ and $\boldsymbol{\nu}_{\mathbb{B}} = \boldsymbol{\psi}$.

In the somewhat uninteresting case where $\boldsymbol{\mu}$ is itself a product measure of the form $\boldsymbol{\mu}_{\mathbb{A}} \times \boldsymbol{\mu}_{\mathbb{B}}$, the solution is easy: $\boldsymbol{\nu} = \boldsymbol{\phi} \times \boldsymbol{\psi}$. But in general no closed-form solution is available.

Theorem 7.19 gives the rate function as the infimum of a convex minimization problem. The reformulation (7.43) makes it obvious that the objective function is convex in $\boldsymbol{\nu}$ since $-H(\boldsymbol{\phi} \| \bar{\boldsymbol{\mu}})$ is just an additive constant. Now by using duality theory, we obtain an alternate formula for the rate function for singleton frequencies.

Theorem 7.20 *Suppose $\boldsymbol{\phi} \in \mathbb{S}_n$ and $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^2)$. Then*

$$\left\{ \inf_{\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)} H(\boldsymbol{\nu} \| \boldsymbol{\mu}) \ \textit{s.t.} \ \bar{\boldsymbol{\nu}} = \boldsymbol{\phi} \right\} = \left\{ H(\boldsymbol{\phi} \| \bar{\boldsymbol{\mu}}) + \sup_{\mathbf{u} > \mathbf{0}} \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(\mathbf{u}A)_i} \right\}, \tag{7.44}$$

where as before $a_{ij} = \mu_{ij} / \bar{\mu}_i$ is the state transition matrix of the Markov chain associated with the doublet frequency vector $\boldsymbol{\mu}$. Therefore an alternate formula for the rate function $J(\cdot)$ is

$$J(\boldsymbol{\phi}) = \sup_{\mathbf{u} > \mathbf{0}} \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(\mathbf{u}A)_i}. \tag{7.45}$$

Proof. Since $H(\boldsymbol{\nu}\|\boldsymbol{\mu})$ is a convex function of $\boldsymbol{\nu}$ and the constraint $\bar{\boldsymbol{\nu}} = \boldsymbol{\phi}$ is linear, it follows that the value of this infimum is the same as the supremum of the dual problem; in other words, there is no duality gap.

To formulate the dual problem, we study the following more general problem, and persist with it as long as we can. The problem is:

$$\inf_{\boldsymbol{\nu}\in\mathbb{S}_{nm}} \sum_{i=1}^{n}\sum_{j=1}^{m} \nu_{ij}\log\frac{\nu_{ij}}{\mu_{ij}} \text{ s.t. } \sum_{j=1}^{m}\nu_{ij} = \phi_i \ \forall i, \text{ and } \sum_{i=1}^{n}\nu_{ij} = \psi_j, \ \forall j. \quad (7.46)$$

Right at the very end we will put $n = m$ and $\boldsymbol{\phi} = \boldsymbol{\psi}$, which will incidentally automatically ensure that $\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^2)$.

The Lagrangian of the above constrained problem is

$$\begin{aligned} L(\boldsymbol{\nu},\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\nu_{ij}\log\frac{\nu_{ij}}{\mu_{ij}} \\ &+ \sum_{i=1}^{n}\left[\phi_i - \sum_{j=1}^{m}\nu_{ij}\right]\alpha_i + \sum_{j=1}^{m}\left[\psi_j - \sum_{i=1}^{n}\nu_{ij}\right]\beta_j, \end{aligned}$$

where $\boldsymbol{\alpha},\boldsymbol{\beta}$ are the vectors of Lagrange multipliers. Then

$$\frac{\partial L}{\partial \nu_{ij}} = \log\frac{\nu_{ij}}{\mu_{ij}} + 1 - \alpha_i - \beta_j.$$

Thus, at the optimum, we have

$$\log\frac{\nu^*_{ij}}{\mu_{ij}} = \alpha_i + \beta_j - 1,$$

or

$$\nu^*_{ij} = \mu_{ij}\exp(\alpha_i + \beta_j - 1).$$

Thus

$$\begin{aligned} L^*(\boldsymbol{\alpha},\boldsymbol{\beta}) &:= \inf_{\boldsymbol{\nu}} L(\boldsymbol{\nu},\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\nu^*_{ij}(\alpha_i+\beta_j-1) + \sum_{i=1}^{n}\left[\phi_i - \sum_{j=1}^{m}\nu^*_{ij}\right]\alpha_i + \sum_{j=1}^{m}\left[\psi_j - \sum_{i=1}^{n}\nu^*_{ij}\right]\beta_j, \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m}\nu^*_{ij} + \sum_{i=1}^{n}\phi_i\alpha_i + \sum_{j=1}^{m}\psi_j\beta_j \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}e^{\alpha_i+\beta_j-1} + \sum_{i=1}^{n}\phi_i\alpha_i + \sum_{j=1}^{m}\psi_j\beta_j. \end{aligned}$$

By duality theory, the infimum in (7.46) is the unconstrained supremum of $L^*(\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha},\boldsymbol{\beta}$.

Next, let us reparametrize the problem. We have

$$L^*(\boldsymbol{\alpha},\boldsymbol{\beta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}e^{\alpha_i+\beta_j-1} + \sum_{i=1}^{n}\phi_i\alpha_i + \sum_{j=1}^{m}\psi_j\beta_j.$$

Now define

$$\exp(\alpha_i) =: v_i, \alpha_i = \log v_i, \exp(\beta_j - 1) =: w_j, \beta_j = \log w_j + 1,$$

and observe that since $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are unconstrained, the corresponding vectors $\mathbf{v}, \mathbf{w}$ are constrained to be strictly positive; that is, $\mathbf{v} > \mathbf{0}, \mathbf{w} > \mathbf{0}$. In a bit of sloppy notation, we continue to refer to the resulting function as $L^*(\mathbf{v}, \mathbf{w})$. Now

$$L^*(\mathbf{v}, \mathbf{w}) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij} v_i w_j + \sum_{i=1}^{n} \phi_i \log v_i + \sum_{j=1}^{m} \psi_j \log w_j + \sum_{=1}^{m} \psi_j .$$

Next, observe that

$$\sup_{\mathbf{v}>\mathbf{0}, \mathbf{w}>\mathbf{0}} L^*(\mathbf{v}, \mathbf{w}) = \sup_{\mathbf{v}>\mathbf{0}} \sup_{\mathbf{w}>\mathbf{0}} L^*(\mathbf{v}, \mathbf{w}).$$

So let us fix $\mathbf{v} > \mathbf{0}$ and define

$$L^{**}(\mathbf{v}) := \sup_{\mathbf{w}>\mathbf{0}} L^*(\mathbf{v}, \mathbf{w}).$$

To compute $L^{**}(\mathbf{v})$, note that

$$\frac{\partial L}{\partial w_j} = -\sum_{i=1}^{n} \mu_{ij} v_i + \frac{\psi_j}{w_j}.$$

Hence at the optimum we have

$$w_j^* = \frac{\psi_j}{\sum_{i=1}^{n} \mu_{ij} v_i}.$$

Thus

$$\begin{aligned} L^{**}(\mathbf{v}) &= L^*(\mathbf{v}, \mathbf{w}^*) \\ &= -\sum_{j=1}^{m} \psi_j + \sum_{i=1}^{n} \phi_i \log v_i + \sum_{j=1}^{m} \psi_j \log \frac{\psi_j}{\sum_{i=1}^{m} v_i \mu_{ij}} + \sum_{j=1}^{m} \psi_j . \end{aligned}$$

Now observe that the first and last terms on the right side cancel out. At last let us use the facts that $n = m$ and $\boldsymbol{\phi} = \boldsymbol{\psi}$. Thus, after making these substitutions, and interchanging the indices i and j in the last summation, we get

$$L^{**}(\mathbf{v}) = \sum_{i=1}^{n} \phi_i \log \frac{\phi_i v_i}{\sum_{j=1}^{m} v_j \mu_{ji}}.$$

Let us now make one last change of variables by defining

$$v_i = u_i / \bar{\mu}_i, u_i = v_i \bar{\mu}_i, \ \forall i.$$

With this change of variables, $\mathbf{u}$ is also constrained to be a strictly positive vector. Also

$$v_j \mu_{ji} = v_j \frac{\mu_{ji}}{\bar{\mu}_j} = u_j a_{ji},$$

where a_{ij} are the elements of the state transition matrix of the Markov chain. Next,

$$\sum_{j=1}^{n} v_j \mu_{ji} = \sum_{j=1}^{n} u_j a_{ji} = (\mathbf{u}A)_i, \ \forall i.$$

And finally, retaining the same symbol L^{**}, we get

$$\begin{aligned} L^{**}(\mathbf{u}) &= \sum_{i=1}^{n} \phi_i \left[\log(\phi_i / \bar{\mu}_i) + \log \frac{u_i}{(\mathbf{u}A)_i} \right] \\ &= H(\boldsymbol{\phi} \| \bar{\boldsymbol{\mu}}) + \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(\mathbf{u}A)_i}. \end{aligned}$$

Therefore the solution to the original minimization problem is

$$H(\boldsymbol{\phi} \| \bar{\boldsymbol{\mu}}) + \inf_{\mathbf{u} > \mathbf{0}} \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(\mathbf{u}A)_i}.$$

This proves (7.44). □

7.3.5 Multi-Step Markov Chains

The results in the two previous subsections dealt with conventional one-step Markov chains. However, the methods can be extended readily to multi-step processes, and that is done in the present subsection.

Let us begin by reprising earlier discussions. Suppose $\{\mathcal{X}_t\}$ is an s-step Markov process, so that

$$E\{\mathcal{X}_t | \mathcal{X}_{t-1} \ldots \mathcal{X}_0\} = E\{\mathcal{X}_t | \mathcal{X}_{t-1} \ldots \mathcal{X}_{t-s}\} \ \forall t.$$

Then the process is completely characterized by the $(s+1)$-tuple frequency vector

$$\mu_{\mathbf{i}} := \Pr\{\mathcal{X}_t^{t+s} = \mathbf{i}\}, \ \forall \mathbf{i} \in \mathbb{A}^{s+1}.$$

Note that the frequency vector $\boldsymbol{\mu}$ is consistent and thus belongs to $\mathcal{M}_c(\mathbb{A}^{s+1})$. Since an s-step Markov process over $\mathbb{A}$ can be viewed as a conventional (one-step) Markov process over the state space $\mathbb{A}^s$, we can identify the stationary distribution

$$\bar{\mu}_{\mathbf{i}} := \sum_{j \in \mathbb{A}} \mu_{i\mathbf{j}} = \sum_{j \in \mathbb{A}} \mu_{j\mathbf{i}}, \ \forall \mathbf{i} \in \mathbb{A}^s,$$

while the transition probabilities are given by

$$\Pr\{\mathcal{X}_t = j | \mathcal{X}_{t-s}^{t-1} = \mathbf{i}\} = \frac{\mu_{\mathbf{i}j}}{\mu_{\mathbf{i}}}.$$

Suppose x_1^l is a sample path of length l of an s-step Markov chain. To construct a *consistent* empirical measure on the basis of this sample path, we define the augmented sample path $\tilde{x}_1^l := x_1 \ldots x_l x_1 \ldots x_s = x_1^l \cdot x_1^s \in \mathbb{A}^{l+s}$. Here the symbol $\cdot$ denotes the concatenation of two strings. The above augmentation is the s-step generalization of adding a single ghost transition from x_l to x_1 in the case of one-step Markov chains. In this case we are adding s ghost transitions. Then we define

$$\nu_{\mathbf{i}} := \frac{1}{l} \sum_{t=1}^{l} I_{\{x_t^{t+s} = \mathbf{i}\}}, \ \forall \mathbf{i} \in \mathbb{A}^{s+1}. \tag{7.47}$$

Compare (7.47) to (7.30). Then the resulting empirical measure $\boldsymbol{\nu}(x_1^l)$ belongs to $\mathcal{M}_c(\mathbb{A}^{s+1})$. For this empirical measure we can state the following result.

Theorem 7.21 *Suppose $\{\mathcal{X}_t\}$ is a stationary s-step Markov assuming values in the finite set $\mathbb{A}$, with the $(s+1)$-tuple frequency vector $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$. Define $\boldsymbol{\nu}(x_1^l) \in \mathcal{M}_c(\mathbb{A}^{s+1})$ as in (7.47). Then the $\mathcal{M}_c(\mathbb{A}^{s+1})$-valued process $\{\boldsymbol{\nu}(x_1^l)\}$ satisfies the LDP with the rate function*

$$I(\boldsymbol{\nu}) := D(\boldsymbol{\nu} \| \boldsymbol{\mu}) = H(\boldsymbol{\nu} \| \boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}} \| \bar{\boldsymbol{\mu}}). \tag{7.48}$$

Proof. Since the proof of Theorem 7.21 closely parallels that of Theorem 7.19, we only sketch it. In analogy with earlier notation, let $\mathcal{E}(l, n, s+1)$ denote the set of empirical measures in $\mathcal{M}_c(\mathbb{A}^{s+1})$ that can possibly arise from a sample path of length l. The set $T(\boldsymbol{\nu})$ denotes the type class of $\boldsymbol{\nu}$, that is, the set of sample paths $x_1^l \in \mathbb{A}^l$ that lead to the empirical measure $\boldsymbol{\nu} \in \mathcal{E}(l, n, s+1)$. With this notation we can mimic all the steps involved in Proving Theorem 7.19. First of all, it is easy to see that

$$|E(l, n, s+1)| \leq (l+1)^{s+1}. \tag{7.49}$$

Compare with (7.34). Second, the countable set $\cup_l \mathcal{E}_{l,s}$ is dense in $\mathcal{M}_c(\mathbb{A}^{s+1})$. Next, we can estimate the cardinality of the type class $T(\boldsymbol{\nu})$ by constructing a graph with n^s nodes, where each node represents a string in $\mathbb{A}^s$. Suppose $\boldsymbol{\nu} \in \mathcal{E}(l, n, s+1)$, and for each string $\mathbf{j} \in \mathbb{A}^{s+1}$, let $l_{\mathbf{j}}$ denote the integer $l\nu_{\mathbf{j}}$. Now since the process under study is an s-step Markov process, a transition is possible only from a state $i\mathbf{j}, i \in \mathbb{A}, \mathbf{j} \in \mathbb{A}^{s-1}$ to another state $\mathbf{j}k, k \in \mathbb{A}$. So in the directed graph, we draw $l_{\mathbf{ij}k}$ directed edges from the node $i\mathbf{j}$ to the node $\mathbf{j}k$. Note that, even though there are n^s nodes in the graph, each node can have outgoing edges to at most n other nodes. Since the empirical measure $\boldsymbol{\nu}$ is consistent, as before each node has the same in-degree and out-degree. Moreover, the number of Eulerian circuits is the number of distinct sample paths that leads to the empirical measure $\boldsymbol{\nu}$. So we can compute the size of the type class $T(\boldsymbol{\nu}, s)$ by counting the number of Eulerian circuits in this graph. This problem is also studied in [59]. The bounds corresponding to (7.35) are

$$(el)^{-2n^{s+1}} e^{lD(\boldsymbol{\nu})} \leq |T(\boldsymbol{\nu})| \leq (l-1) e^{lD(\boldsymbol{\nu})}. \tag{7.50}$$

Finally, we need to estimate the log likelihood of each sample path within the type class $T(\boldsymbol{\nu})$. We can simply mimic the arguments leading to (7.36) to get

$$\Pr\{\mathcal{X}_1^l = x_1^l\} = \bar{\mu}(x_1^s) \cdot \prod_{t=l-s}^{l} \frac{\bar{\mu}(x_t^{t+s-1})}{\mu(x_t^{t+s})} \cdot \prod_{t=1}^{l} \frac{\mu(x_t^{t+s})}{\bar{\mu}(x_t^{t+s-1})}.$$

Thus (7.37) holds with

$$\underline{c}, \bar{c} = \min, \max \prod_{t=l-s}^{l} \frac{\bar{\mu}(x_t^{t+s-1})}{\mu(x_t^{t+s})}.$$

This in turn permits us to prove analogs of (7.38) and (7.39). The last part of the proof of Theorem 7.19 goes through unchanged. □

Theorem 7.21 presents the rate function for $(s+1)$-tuple frequencies of an s-step Markov chain. Using the contraction principle, it is possible to obtain the rate function for the frequencies of k-tuples where $1 \leq k \leq s$. For this purpose, we introduce a new symbol. Suppose $\boldsymbol{\nu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$ and that $1 \leq r \leq s$. Then $B^r(\boldsymbol{\nu}) \in \mathcal{M}_c(\mathbb{A}^{s+1-r})$ is defined by

$$[B^r(\boldsymbol{\nu})]_{\mathbf{i}} := \sum_{\mathbf{j} \in \mathbb{A}^r} \nu_{\mathbf{ij}} = \sum_{\mathbf{j} \in \mathbb{A}^r} \nu_{\mathbf{ji}} \ \forall \mathbf{i} \in \mathbb{A}^{s+1-r}.$$

Thus $B^1(\boldsymbol{\nu}) = \bar{\boldsymbol{\nu}}$ as defined earlier, and $B^r(\boldsymbol{\nu})$ is just $\boldsymbol{\nu}$ "barred" r times. Because $\boldsymbol{\nu}$ is consistent, it does not matter whether the summation is on the first component or the last component.

Now suppose we are given a sample path x_1^l and we construct the empirical measure $\boldsymbol{\nu}(x_1^l)$ as in (7.47). Then it is clear that

$$[B^r(\boldsymbol{\nu})]_{\mathbf{i}} = \frac{1}{l} \sum_{t=1}^{l} I_{\{x_t^{t+s-r} = \mathbf{i}\}}, \ \forall \mathbf{i} \in \mathbb{A}^{s+1-r}. \tag{7.51}$$

Theorem 7.22 *Suppose $\{\mathcal{X}_t\}$ is a stationary s-step Markov assuming values in the finite set $\mathbb{A}$, with the $(s+1)$-tuple frequency vector $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$. Define $\boldsymbol{\nu}(x_1^l) \in \mathcal{M}_c(\mathbb{A}^{s+1})$ as in (7.47), and define the $\mathcal{M}_c(\mathbb{A}^k)$-valued process $\{B^{s+1-k}(\boldsymbol{\nu}(x_1^l))\}$ as in (7.51) with $r = s+1-k$. Then this process satisfies the LDP with the rate function*

$$\begin{aligned} I(\phi) &:= \inf D(\boldsymbol{\nu} \| \boldsymbol{\mu}) \ \textit{s.t.} \ B^{(s+1-k)}(\boldsymbol{\nu}) = \phi \\ &= \inf H(\boldsymbol{\nu} \| \boldsymbol{\mu}) - H(\bar{\boldsymbol{\nu}} \| \bar{\boldsymbol{\mu}}) \\ &\textit{s.t.} \ B^{(s+1-k)}(\boldsymbol{\nu}) = \phi. \end{aligned} \tag{7.52}$$

Corollary 7.23 *Let $\{\mathcal{X}_t\}$ be as in Theorem 7.22 and define*

$$a_{\mathbf{i}j} := \frac{\mu_{\mathbf{i}j}}{\bar{\mu}_{\mathbf{i}}} = \Pr\{\mathcal{X}_t = j | \mathcal{X}_{t-s}^{t-1} = \mathbf{i}\}, \ \forall \mathbf{i} \in \mathbb{A}^s, j \in \mathbb{A}.$$

Then the s-tuple empirical frequency vector $\{\bar{\boldsymbol{\nu}}(x_1^l)\}$ satisfies the LDP with the rate function

$$I(\phi) := \sup_{\mathbf{u} \in \mathbb{R}^s, \mathbf{u} > \mathbf{0}} \sum_{\mathbf{i} \in \mathbb{A}^{s-1}} \sum_{j \in \mathbb{A}} \phi_{\mathbf{i}} \log \frac{u_{\mathbf{i}j}}{\sum_{k \in \mathbb{A}} u_{k\mathbf{i}} a_{\mathbf{i}j}}. \tag{7.53}$$

The proofs are omitted as they are entirely analogous to those of the corresponding theorems in Section 7.3.4.

7.3.6 Rate Functions for Time-Reversed Markov Chains

Suppose $\{\mathcal{X}_t\}_{t=-\infty}^{\infty}$ is a Markov chain assuming values in $\mathbb{A}$. Then $\{\mathcal{X}_{-t}\}_{t=-\infty}^{\infty}$ is the corresponding **time-reversed** Markov chain. If the original chain has the stationary distribution π, then so does the time-reversed Markov chain. If the original Markov chain has the state transition matrix $A = [a_{ij}]$, then the time-reversed Markov chain has the state transition matrix $A^{(r)} = [a_{ij}^{(r)}]$ where

$$a_{ij}^{(r)} = \frac{\pi_i a_{ij}}{\pi_j}. \tag{7.54}$$

But there is a much simpler way to think of a time-reversed Markov chain. Given a k-tuple $\mathbf{i} = i_1 \ldots i_k \in \mathbb{A}^k$, let $\mathbf{i}^{(r)}$ denote $i_k i_{k-1} \ldots i_2 i_1$. Thus $\mathbf{i}^{(r)}$ is just $\mathbf{i}$ written backwards. Recall that an s-step Markov chain is completely characterized by the vector of frequencies of $(s+1)$-tuples, call it $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$. Then the time-reversed Markov chain has the $(s+1)$-tuple frequency vector defined by

$$[\mu^{(r)}]_{\mathbf{i}} = \mu_{\mathbf{i}}^{(r)}, \ \forall \mathbf{i} \mathbb{A}^{s+1}. \tag{7.55}$$

In particular, if a conventional (one-step) Markov chain has the doublet frequency vector $\mu_{ij}, i, j \in \mathbb{A}$, then its time-reversed version has the doublet frequency vector μ_{ji}. Therefore a one-step Markov chain is reversible (equal to its time-reversed version) if and only if

$$\mu_{ij} = \mu_{ji}, \ \forall i, j \in \mathbb{A}.$$

With the above observation we can readily derive various rate functions of time-reversed Markov chains in terms of the corresponding rate functions of the original chain.

Theorem 7.24 *Suppose $\{\mathcal{X}_t\}_{t=-\infty}^{\infty}$ is a stationary s-step Markov assuming values in the finite set $\mathbb{A}$, with the $(s+1)$-tuple frequency vector $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$. Let $I_k(\cdot)$ denote the rate function defined over $\mathcal{M}_c(\mathbb{A}^k)$ for empirical frequencies of k-tuples defined in accordance with (7.51). Then for the time-reversed Markov process, the empirical frequencies of k-tuples satisfy the LDP with the rate function $I_k^{(r)} : \mathcal{M}_c(\mathbb{A}^k) \to \mathbb{R}_+$ defined by*

$$I_k^{(r)}(\phi) := I_k(\phi^{(r)}), \ \forall \phi \in \mathcal{M}_c(\mathbb{A}^k). \tag{7.56}$$

Corollary 7.25 *Suppose $\{\mathcal{X}_t\}$ is a (one-step) Markov chain. Then the empirical frequencies of singletons of both the "forward" and time-reversed Markov chain have exactly the same rate function.*

The proof of Theorem 7.24 is quite easy and makes use of two facts. First, if $\boldsymbol{\mu} \in \mathcal{M}_c(\mathbb{A}^{s+1})$, then

$$\sum_{j \in \mathbb{A}} \mu_{\mathbf{i}j} = \sum_{j \in \mathbb{A}} \mu_{j\mathbf{i}}, \ \forall \mathbf{i} \in \mathbb{A}^s.$$

In other words, whether we project a consistent distribution on the first component or the last, we get exactly the same answer. Second, given any two distributions $\boldsymbol{\nu}, \boldsymbol{\mu}$ on $\mathbb{A}^k$, the relative entropy $H(\boldsymbol{\nu} \| \boldsymbol{\mu})$ is invariant under every permutation of the indices.

Chapter Eight

BLAST Theory

BLAST (Basic Local Alignment Search Tool) is a widely used statistical method for finding similarities between sequences of symbols from finite alphabets. While the theory is completely general, the most widely used applications are to comparing sequences of nucelotides and sequences of amino acids. Though the letter B in BLAST stands for 'basic,' in fact the theory itself is anything but basic. The objective of this chapter therefore is to present an accessible treatment of the theory.

The theory of BLAST was developed through a series of papers co-authored by Samuel Karlin; see [68, 65, 66, 67, 33, 34]. The notation and problem formulations across these papers are not always consistent, making it very difficult for the non-expert reader to navigate through these papers. It is hoped that the present chapter will assist somewhat in this process. The treatment here follows [33, 34]. The reader is cautioned that there are several modifications of the theory presented here; these modifications do not always have a theoretical justification. In the interests of brevity and clarity, we treat here only the most 'basic' version of BLAST theory.

The chapter is organized as follows. In Section 8.1, we discuss the problem of optimal gapped alignment between two sequences. Though an 'exact' solution to this problem can be found using dynamic programming, it turns out that an approximate solution is often good enough in many situations. This was one of the motivations for the development of BLAST theory. In Section 8.2, we present the problem that BLAST theory addresses, state the main results without proof, and show how these main results can be applied in practice. In Section 8.3, we present the proofs of all the main results. A reader who is not interested in knowing how the theorems that underlie BLAST are proved can skip this section.

8.1 THE GAPPED SEQUENCE ALIGNMENT PROBLEM

8.1.1 Problem Formulation

Suppose we are given two strings $\mathbf{x} = x_1 \ldots x_k$ and $\mathbf{y} = y_1 \ldots y_l$ over a common finite alphabet $\mathcal{N}$. Often, though not always, it is the case that one of the strings is much shorter than the other, say $k \ll l$. In such a case, determining whether $\mathbf{x}$ is a *perfect* substring of $\mathbf{y}$ is computationally straight-forward. Indeed, text editors address precisely this problem. Thus it is easy to determine whether or not there exists an index j such that $y_{j+i} = x_i$ for $i = 1, \ldots, k$. The problem remains tractable even if we introduce some ‘wild card’ entries. Thus, a text editor that searches for the string $x_1 \ldots x_s * x_{s+1} \ldots x_k$ within $\mathbf{y}$ looks for indices j_1 and $j_2 \geq j_1$ such that

$$y_{j_1+i} = x_i \text{ for } i = 1, \ldots, s, \text{ and } y_{j_2+i} = x_i \text{ for } i = s+1, \ldots, k.$$

The string $\mathbf{x}$ can in fact be divided into any finite number of segments and the wild cards introduced in-between, and the problem remains tractable. The tractability arises from two factors: First, the locations within the string $\mathbf{x}$ where one or more wild card entries are to be introduced are specified ahead of time. Second, we insist on a *perfect* match between the symbols of the two strings. If we were to change either of these requirements then the problem becomes more difficult.

Now we state the so-called ‘optimal gapped alignment’ problem. Suppose $\mathbf{x} = x_1 \ldots x_k$ is a string over some finite alphabet $\mathcal{N}$ and $\mathbf{y} = y_1 \ldots y_l$ is a string over another finite alphabet $\mathcal{M}$, which may or may not be the same as $\mathcal{N}$. The problem is to determine an optimal ‘gapped’ alignment between the two strings. Before stating the problem formally, we motivate it through a simple example. Suppose $\mathcal{N} = \mathcal{M} = \{A, C, G, T\}$, the set of nucleotides, and let

$$\mathbf{x} = ACACTGT, \mathbf{y} = TAGACGGAGCTTCAC.$$

Then these two strings can be imperfectly aligned as shown below, with the dash indicating a ‘gap’:

$$\begin{array}{cccccccccccccccc} & & & A & C & - & - & A & C & - & T & G & T & & & \\ T & A & G & A & C & G & G & A & G & C & T & - & T & A & A & C \end{array}$$

By judiciously introducing gaps into the two sequences, we are able to achieve ‘perfect’ matches between those symbols within each string that do not lie opposite a gap, with just one mismatch. To measure the quality of the gapped alignment, we introduce a ‘scoring’ function $S : \mathcal{N} \times \mathcal{M} \to \mathbb{R}$ that assigns a real number score $S(i, j)$ to each pair $(i, j) \in \mathcal{N} \times \mathcal{M}$. We can also think of $S(i, j)$ as representing the ‘similarity’ between the symbols i and j instead of ‘score’. Thus in the case where $\mathcal{M} = \mathcal{N}$ meaning that both strings $\mathbf{x}$ and $\mathbf{y}$ are over the same alphabet, we would expect $S(i, j)$ to be large and positive whenever $i = j$, and to be much smaller and possibly negative if $i \neq j$. Note that it is *not* assumed that the matrix $S(i, j)$ is symmetric. We also need to define ‘gap scores’ $S(i, -)$ and $S(-, j)$. Note that in this problem it makes no sense to put one gap opposite another. We can avoid the situation by defining $S(-, -) = -\infty$. In this way, the scoring function S can be extended to $(\mathcal{N} \cup \{-\}) \times (\mathcal{M} \cup \{-\})$.

At this point we need to distinguish between ‘global’ alignment and ‘local’ alignment. In global alignment, we would insist that the end points of the two strings must match, after being augmented by gaps if necessary. Thus, in the example above, we would be forced to place gaps above TAG to the left of the $\mathbf{x}$ string, and above CAC to the right of the $\mathbf{x}$ string. This makes no sense if, as is often the case in practice, one of the strings is significantly longer than the other. In such a case one would study *local* alignment, where the scoring function is counted only between the end points of the shorter string. In either case, the total score of a gapped alignment is the sum of the individual pairwise scores $S(i,j), S(i,-)$ or $S(-,j)$ as we traverse from one end to the other. The optimal gapped alignment problem is to determine the alignment that results in the highest score.

8.1.2 Solution Via Dynamic Programming

The problem of optimal gapped alignment can be solved using dynamic programming. The principle of optimality, which we have encountered in Section 8.2, applies here too: If any alignment is optimal, then any subset thereof must also be optimal for the appropriate substrings. Otherwise we could take out that particular part of the alignment, replace it with a better alignment, and improve the overall score. Clearly this property is a consequence of the *additive nature* of the total score.

Using the principle of optimality, we can give a simple recursive scheme for solving the problem of optimal gapped alignment. Let $\mathbf{x}, \mathbf{y}$ be the two strings to be aligned (not necessarily over the same alphabet), using the scoring function $S : \mathcal{N} \times \mathcal{M} \rightarrow \mathbb{R}$. For simplicity let us suppose that $S(-,j) = S(i,-) = -\gamma$. Let $\mathbf{x}$ have length k and let $\mathbf{y}$ have length l. Suppose we begin aligning from the ends of the two strings, and let $P^*(i,j)$ denote the highest possible score that can be achieved by optimally aligning from the end of $\mathbf{x}$ until position i, and from the end of $\mathbf{y}$ until position j. Think of $P^*(i,j)$ as the optimal payoff until positions i,j. Now the optimal payoff function satisfies the following recursion:

$$P^*(i,j) = \max \begin{cases} P^*(i,j+1) - \gamma \\ P^*(i+1,j) - \gamma \\ P^*(i+1,j+1) + S(x_i, y_j) \end{cases} \tag{8.1}$$

This is because, at position i,j there are only three things we can do:

1. We can introduce a gap above the symbol y_j.

2. We can introduce a gap below the symbol x_i.

3. We can match x_i against y_j.

Let us suppose that we have aligned the two sequences optimally before i,j. In the first alternative, the resulting score would be $P^*(i,j+1) - \gamma$, because $P^*(i,j+1)$ is the optimal score of aligning $x_i, \ldots, x_k$ against $y_{j+1}, \ldots, y_l$, and $-\gamma$ is the additional score due to introducing a gap above y_j. Similarly, in the second alternative, the resulting score would be $P^*(i+1,j) - \gamma$. Finally, in the third alternative the resulting score would be $P^*(i+1,j+1) + S(x_i, y_j)$. Now the principle of optimality tells us that the best thing to do would be to maximize amongst these three alternatives. It is of course possible that there is a 'tie' between two alternatives, in which case we arbitrarily choose one of them; this does not affect the discussion to follow. To apply the above formula, we begin at the ends of the two strings $\mathbf{x}$ and $\mathbf{y}$ with the optimal score $P^*(k.l) = 0$, and work backwards.

Example 8.1 The application of (8.1) is illustrated through a 'toy' example. It must be emphasized that in reality the smallest values of l, k for which we would wish to carry out optimal gapped alignment would be of the order of a few hundred.

Suppose the scoring function is given by

$$S = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left[\begin{array}{rrrr} 10 & -3 & -2 & 1 \\ -2 & 8 & 1 & -2 \\ -3 & 1 & 9 & -3 \\ 0 & -3 & -2 & 6 \end{array} \right] \end{array}, \gamma = 1.$$

Suppose the two strings to be aligned are

$$\mathbf{x} = CACGAAT, \mathbf{y} = AGTTCAA.$$

Then we can construct the table of optimal payoff functions as follows:

	C	A	C	G	A	A	T	
A	$32 \leftarrow$	$33 \nwarrow$	$22^* \nwarrow$	$24 \uparrow$	$17 \nwarrow$	$11 \nwarrow$	$0 \uparrow$	$-7 \uparrow$
G	$26 \nwarrow$	$24 \uparrow$	$23 \uparrow$	$25 \nwarrow$	$15 \uparrow$	$7 \uparrow$	$1 \uparrow$	$-6 \uparrow$
T	$24^* \uparrow$	$25^* \uparrow$	$24 \uparrow$	$18 \uparrow$	$16 \uparrow$	$8 \uparrow$	$2^* \nwarrow$	$-5 \uparrow$
T	$25 \leftarrow$	$26 \nwarrow$	$25 \uparrow$	$19 \uparrow$	$17 \uparrow$	$9 \uparrow$	$3 \nwarrow$	$-4 \uparrow$
C	$24^* \nwarrow$	$25 \leftarrow$	$26 \nwarrow$	$20 \nwarrow$	$18 \uparrow$	$10 \uparrow$	$-1 \uparrow$	$-3 \uparrow$
A	$15 \leftarrow$	$16^* \leftarrow$	$17 \leftarrow$	$18 \leftarrow$	$19 \nwarrow$	$11 \nwarrow$	$0^* \uparrow$	$-2 \uparrow$
A	$4 \leftarrow$	$5 \nwarrow$	$6 \leftarrow$	$7 \leftarrow$	$8 * \leftarrow$	$9 \nwarrow$	$1 \nwarrow$	$-1 \uparrow$
	$-7 \leftarrow$	$-6 \leftarrow$	$-5 \leftarrow$	$-4 \leftarrow$	$-3 \leftarrow$	$-2 \leftarrow$	$-1 \leftarrow$	0

In constructing this table, we need to keep track of the optimal choice that we made at each square of the matrix. Thus a left arrow ← indicates that the optimal choice was to insert a gap above y_j, while a vertical arrow ↑ indicates that the optimal choice was to insert a gap below x_i. The diagonal arrow ↖ indicates that the optimal choice was to match x_i against y_j. A red color for the number indicates a 'tie.'

The above procedure gives the optimal gapped alignment of the two sequences, namely

$$\begin{array}{cccccccccc} C & A & - & - & - & C & G & A & A & T \\ - & A & G & T & T & C & - & A & A & - \end{array}, P^* = 32.$$

However, the same matrix can also be used to determine the optimal gapped alignment from the ends of the two strings to any pair of intermediate points, as shown below.

$$\begin{array}{cccccc} - & - & G & A & A & T \\ T & T & C & A & A & - \end{array}, P^* = 18.$$

$$\begin{array}{cccccccc} C & G & - & - & - & A & A & T \\ A & G & T & T & C & A & A & - \end{array}, P^* = 22.$$

The solution using dynamic programming was first introduced into the biology community by Needleman and Wunsch [82] to solve problems of optimal *global* alignment of two strings. It is possible to make simple modifications to the algorithm to solve the problem of optimal *local* alignment. This was done by Smith and Waterman; see [101].

From the above discussion, it is easy to see that the complexity of optimal gapped alignment using the above procedure $O(kl)$, where k, l are the lengths of $\mathbf{x}, \mathbf{y}$ respectively. If both strings are of comparable length, then the complexity is quadratic in the length. There are several improvements available that trade off storage for time or vice versa, but these need not concern us here. It is also possible to modify (8.1) to incorporate more sophisticated scoring functions. For example, one can have different penalties for gap *creation* versus gap *extension*; it is believed by biologists that the gap extension penalty should be much smaller than the gap creation penalty. Similarly, it possible to make the scoring function depend not only on the two symbols being matched, but also on their positions within the two strings. The modifications required are relatively straight-forward, and the reader is referred to [54] for more detailed discussion.

8.2 BLAST THEORY: STATEMENTS OF MAIN RESULTS

8.2.1 Problem Formulations

The fundamental objective of BLAST theory is to align sequences as well as possible, and then make a determination as to the level of statistical significance of the alignment. Thus one computes a 'maximal segmental score' of the alignment between the two sequences, and tests to see whether the maximal segmental score could have been obtained purely as a matter of chance. If the match is better than could be explained by chance, then one would be able to conclude that the two sequences do indeed show some similarity. Thus, in order to apply the theory, one needs to be able to compute two things: The expected maximal segmental score for sequences of a given length, and the 'tail probability distribution' of the likelihood that the maximal segmental score will exceed this expected value. In the remainder of the chapter, we derive answers to these and other related questions.

The fundamental theories of BLAST are developed in a series of papers written by Samuel Karlin along with several coauthors [68, 65, 66, 67, 33, 34]. We note here that neither the problem formulation nor the notation is constant across these papers. Thus we begin by surveying the various problem formulations.

Suppose $\mathbb{A}$ and $\mathbb{B}$ are finite sets, and that $\mathcal{X}, \mathcal{Y}$ are random variables assuming values in $\mathbb{A}$ and $\mathbb{B}$ respectively, with probability distributions $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ respectively. We can also consider the 'product' random variable $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ that assumes values in the product set $\mathbb{A} \times \mathbb{B}$ and has the product distribution $\boldsymbol{\mu} := \boldsymbol{\phi} \times \boldsymbol{\psi}$. Note that the marginal distributions of $\boldsymbol{\mu}$

are given by $\boldsymbol{\mu}_{\mathbb{A}} = \boldsymbol{\phi}$ and $\boldsymbol{\mu}_{\mathbb{B}} = \boldsymbol{\psi}$.

Now suppose we draw i.i.d. samples of $\mathcal{X}$ and $\mathcal{Y}$ of length l according to their respective laws; call the sample paths $x_1, \ldots, x_l$ and $y_1, \ldots, y_l$ respectively. As discussed in Section 8.1, let us define a scoring function $F : \mathbb{A} \times \mathbb{B} \to \mathbb{R}$. Then we can define the cumulative score of the sample paths as

$$\sum_{i=1}^{l} F(x_i, y_i).$$

However, this cumulative score is not of interest to us. Rather, we are interested in the 'maximal segmental score.' This can be defined in one of two ways. If we insist that the starting points of the two segments must coincide, then we examine the quantity

$$R_l := \max_{L \geq 0, 0 \leq i \leq l-L} \sum_{k=1}^{L} F(x_{i+k}, y_{i+k}). \tag{8.2}$$

The quantity R_l examines all subsequences of length L within the two sample paths, and then computes only the *segmental score* over this segment of length L; then the maximum of all these segmental scores over all possible segment length L becomes R_l. If we *don't* insist that the starting points of the two segments must coincide, or in other words, if we allow the two segments to be shifted with respect to each other, then we look at the quantity

$$M_l := \max_{L \geq 0, 0 \leq i,j \leq l-L} \sum_{k=1}^{L} F(x_{i+k}, y_{j+k}). \tag{8.3}$$

To repeat, the main difference between the quantities R_l and M_l is that in defining R_l, we insist that the starting points of the two segments being aligned must coincide, whereas in defining M_l, we permit the starting points of the two segments to be shifted with respect to each other. The reader is cautioned that in [65, 67], the quantity R_l is referred to as M_l; thus the notation changes from [65, 67] to [33, 34].

When we examine the issue of maximal scores, we can ask four distinct questions:

1. Given sample paths of length l, what is the expected value of M_l or R_l?

2. Let L_l denote the length of a maximal scoring segment. What is the expected value of L_l? That is, how long is a maximally scoring segment on average, from a sample path of length l?

3. What is the empirical distribution of the symbols x_{i+k}, y_{j+k} in a maximally scoring segment?

4. What is the tail probability distribution of the quantities M_l and R_l beyond their expected values? In other words, suppose M_l exceeds its

expected value by some ϵ. Can we compute the likelihood that this has happened purely due to chance? This would give us the *significance* of the high-scoring segment.

In the sequel, we will answer all of these questions.

8.2.2 The Moment Generating Function

Suppose $X = \{x_1, \ldots, x_n\}$ is a subset of the real numbers $\mathbb{R}$ (and not an abstract set of labels, as in other places in this book). Suppose $\mathcal{X}$ is a random variable assuming values in the set X with the distribution $\boldsymbol{\mu}$. Thus μ_i denotes $\Pr\{\mathcal{X} = x_i\}$ for all i. It can be assumed without loss of generality that $\mu_i > 0$ for all i, because if $\mu_i = 0$ for some i, then the corresponding element x_i can simply be deleted from the set X. For each positive integer k, the quantity

$$M_k(\mathcal{X}) := \sum_{i=1}^{n} x_i^k \mu_i = E[\mathcal{X}^k, \boldsymbol{\mu}]$$

is called the **k-th moment** of the random variable $\mathcal{X}$. In particular, the quantity $M_1(\mathcal{X})$ is just the mean of the random variable $\mathcal{X}$, while $M_2(\mathcal{X}) - [M_1(\mathcal{X})]^2$ is the variance of $\mathcal{X}$. Note that, since X is a finite set, all of the summations above are also finite; as a result, $M_k(\mathcal{X})$ is well-defined for every integer $k \geq 1$. Next, the function

$$\text{mgf}(\lambda; \mathcal{X}) := E[\exp(\lambda \mathcal{X}), \boldsymbol{\mu}] = \sum_{i=1}^{n} \mu_i \exp(\lambda x_i)$$

is called the **moment-generating function (mgf)** of the random variable $\mathcal{X}$. Note that

$$\left[\frac{d^k \text{mgf}(\lambda; \mathcal{X})}{d\lambda^k} \right]_{\lambda=0} = \left[\sum_{i=1}^{n} \mu_i x_i^k \exp(\lambda x_i) \right]_{\lambda=0} = \sum_{i=1}^{n} \mu_i x_i^k = M_k(\mathcal{X})$$

for every integer $k \geq 1$. This explains the nomenclature. Note that $\text{mgf}(0; \mathcal{X}) = 1$ for every random variable $\mathcal{X}$.

Next we define the so-called **logarithmic moment generating function** $\Lambda(\lambda; \mathcal{X})$ by

$$\Lambda(\lambda; \mathcal{X}) := \log \text{mgf}(\lambda; \mathcal{X}) = \log E[\exp(\lambda \mathcal{X}), \boldsymbol{\mu}].$$

Since the function log is concave, it follows from Jensen's inequality that

$$\log E[\exp(\lambda \mathcal{X}), \boldsymbol{\mu}] \geq E[\log \exp(\lambda \mathcal{X}), \boldsymbol{\mu}] = \lambda M_1(\mathcal{X}), \ \forall \lambda.$$

A very useful property of the mgf and its logarithm are brought out next.

Lemma 8.1 *For a fixed nontrivial random variable $\mathcal{X}$, both* $\text{mgf}(\lambda; \mathbf{x})$ *and* $\Lambda(\lambda; \mathcal{X})$ *are strictly convex functions of* λ.

Here by a 'nontrivial' random variable, we mean a random variable that assumes at least two distinct values. (Otherwise the 'random varible' would be just a constant!)

Proof. For a fixed random variable $\mathcal{X}$, we have

$$\text{mgf}(\lambda; \mathcal{X}) = \sum_{i=1}^{n} \mu_i \exp(\lambda x_i)$$

is a linear combination of strictly convex functions $\lambda \mapsto \exp(\lambda x_i)$. Hence $\text{mgf}(\lambda; \mathcal{X})$ is also a strictly convex function of λ. To show that $\Lambda(\lambda; \mathcal{X})$ is also strictly convex in λ, let us, for the purposes of this proof alone, use η to denote the moment generating function and η' to denote $d\eta/d\lambda$. Then we have

$$\Lambda = \log \eta, \Lambda' = \frac{\eta'}{\eta}, \Lambda'' = \frac{\eta\eta'' - (\eta')^2}{\eta^2}.$$

Hence the strict convexity of Λ follows if it can be established that

$$\eta\eta'' > (\eta')^2 \ \forall \lambda.$$

For this purpose, note that

$$\eta = \sum_{i=1}^{n} \mu_i \exp(\lambda x_i), \eta' = \sum_{i=1}^{n} \mu_i x_i \exp(\lambda x_i) \eta'' = \sum_{i=1}^{n} \mu_i x_i^2 \exp(\lambda x_i).$$

So the inequality that we desire to establish can be written as

$$(E[\mathcal{X} e^{\lambda \mathcal{X}}, \boldsymbol{\mu}])^2 < E[e^{\lambda \mathcal{X}}, \boldsymbol{\mu}] \cdot E[\mathcal{X}^2 e^{\lambda \mathcal{X}}, \boldsymbol{\mu}].$$

Now we make use of Schwarz' inequality, which says in this setting that

$$(E[fg, \boldsymbol{\mu}])^2 \leq E[f^2, \boldsymbol{\mu}] \cdot E[g^2, \boldsymbol{\mu}],$$

with equality if and only if f and g are multiples of each other. Apply Schwarz' inequality with the choices

$$f = \exp(\lambda \mathcal{X}/2), g = \mathcal{X} \exp(\lambda \mathcal{X}/2),$$

and observe that if $\mathcal{X}$ is a nontrivial random variable, then f and g are *not* multiples of each other. The desired inequality follows. □

Now we state a very useful fact about the moment generating function.

Lemma 8.2 *Suppose $\mathcal{X}$ is a random variable assuming values in a finite set $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$, with the probability distribution $\boldsymbol{\mu}$, where $\mu_i > 0$ for all i. Suppose in addition that*

(i) X contains both positive and negative numbers. In other words, there exist indices i and j such that $x_i > 0$ and $x_j < 0$ (and by assumption $\mu_i, \mu_j > 0$).

(ii) $E[\mathcal{X}, \boldsymbol{\mu}] \neq 0$.

Under these conditions, there exists a unique $\lambda^* \neq 0$ *such that* $\text{mgf}(\lambda^*; \mathcal{X}) = 1$ *or equivalently* $\Lambda(\lambda^*; \mathcal{X}) = 0$*. Moreover,* λ^* *has sign opposite to that of* $E[\mathcal{X}, \boldsymbol{\mu}]$*.*

Proof. Note that $\text{mgf}(0; \mathcal{X})$ always equals 1. Now

$$\text{mgf}'(0; \mathcal{X}) = \left[\frac{d\mu(\lambda; \mathcal{X})}{d\lambda} \right]_{\lambda=0} = E[\mathcal{X}, P] \neq 0$$

by assumption. We have already seen from Lemma 8.1 that the mgf is a strictly convex function. Finally, it follows from Condition (i) again that

$$\mu(\lambda; \mathcal{X}) \to \infty \text{ as } \lambda \to \pm\infty.$$

This is because at least one of the x_i is positive and at least one is negative. From this information we conclude that the equation $\mu(\lambda; \mathcal{X}) = 1$ has precisely two solutions, one of which is $\lambda = 0$, and the other one, denoted by λ^*, has sign opposite to that of $E[\mathcal{X}, P]$. □

Note that if the set X consists of only nonnegative or only nonpositive numbers, then the above lemma is false. In particular, if $x_i \geq 0$ for all i, then $\mu(\lambda; \mathcal{X}) \to 0$ as $\lambda \to -\infty$, and the only solution of $\mu(\lambda; \mathcal{X}) = 1$ is $\lambda = 0$. The situation when every element of X is nonpositive is similar.

The vector $\boldsymbol{\theta}$ defined by

$$\theta_i = \exp(\lambda^* x_i) \mu_i$$

belongs to $\mathbb{S}_n$ because clearly $\theta_i > 0$ for all i, and in addition,

$$\sum_{i=1}^{n} \theta_i = \text{mgf}(\lambda^*; \mathcal{X}) = 1.$$

The distribution $\boldsymbol{\theta}$ is referred to as the **conjugate distribution** of $\boldsymbol{\mu}$ with respect to the random variable $\mathcal{X}$. Note that $\boldsymbol{\theta}$ depends on both the distribution $\boldsymbol{\mu}$ and the corresponding values of the random variable $\mathcal{X}$.

8.2.3 Statement of Main Results

In this subsection, we state the main results of BLAST theory. Specifically, we answer the four questions raised earlier. Recall that we are given a probability distribution $\boldsymbol{\mu} = \boldsymbol{\phi} \times \boldsymbol{\psi}$ on the product set $\mathbb{A} \times \mathbb{B}$ and a scoring function $F : \mathbb{A} \times \mathbb{B} \to \mathbb{R}$. So we can think of F as a real-valued random variable that assumes the value $F(x_i, y_j)$ with probability $\mu_{ij} = \phi_i \psi_j$. To simplify notation, let us denote $F(x_i, y_j)$ by F_{ij}.

In order to state these results, we introduce one 'standing assumption'.

$$E[F, \boldsymbol{\mu}] < 0, \text{ and } \exists i, j \text{ s.t. } F(x_i, y_j) > 0. \tag{8.4}$$

Thus the standing assumption states that there is at least one pair (x_i, y_j) for which the score F_{ij} is positive, but the expected value of the score over all pairs is negative. With this standing assumption, it follows from Lemma 8.2 that there exists a unique number $\lambda^* > 0$ such that $E[\exp(\lambda^* F, \boldsymbol{\mu}] = 1$.

As before, let $\boldsymbol{\theta}$ denote the conjugate distribution of $\boldsymbol{\mu}$ with respect to F, that is

$$\theta_{ij} = \exp(\lambda^* F_{ij})\mu_{ij}, \ \forall i, j. \tag{8.5}$$

The results are stated next. Note that all results are asymptotic; that is, they apply as $l \to \infty$. In the theorem, the notation $a_l \sim b_l$ as $l \to \infty$ means that the ratio $a_l/b_l \to 1$ as $l \to \infty$.

Theorem 8.3 *Let all symbols be as defined above. Then*

1. $R_l \sim (\ln l)/\lambda^*$ *as* $l \to \infty$.
2. *Let* $\boldsymbol{\theta}$ *denote the conjugate distribution of* $\boldsymbol{\mu}$*, as defined in (8.5). Then the length of a maximal scoring segment* L_l *is asymptotically equal to* $l/H(\boldsymbol{\theta}\|\boldsymbol{\mu})$.
3. *On any maximal scoring segment, the empirical distribution of* $(\mathcal{X}, \mathcal{Y})$ *is asymptotically equal to* $\boldsymbol{\theta}$.

Now we come to the quantity M_l, the maximum segmental score when we *don't* insist that the starting points of the two segments must match.

Theorem 8.4 *Let all symbols be as defined above. Suppose in addition that the two sets* $\mathbb{A}, \mathbb{B}$ *are the same, that the marginal distributions* $\boldsymbol{\phi}, \boldsymbol{\psi}$ *are the same, and that the scoring function* F *is symmetric in the sense that* $F(x_i, y_j) = F(y_j, x_i)$. *Then*

1. $M_l \sim (2 \ln l)/\lambda^*$ *as* $l \to \infty$.
2. *Let* $\boldsymbol{\theta}$ *denote the conjugate distribution of* $\boldsymbol{\mu}$*, as defined in (8.5). Then the length of a maximal scoring segment* L_l *is asymptotically equal to* $2l/H(\boldsymbol{\theta}\|\boldsymbol{\mu})$.
3. *On any maximal scoring segment, the empirical distribution of* $(\mathcal{X}, \mathcal{Y})$ *is asymptotically equal to* $\boldsymbol{\theta}$.

Whereas it is possible to analyze the asymptotic behavior of R_l without making any additional assumptions about the two sets $\mathbb{A}, \mathbb{B}$, the two marginal distributions, or the scoring function, we are obliged to make some assumptions in order to get nice statements about M_l. Actually, any reader who takes the trouble to go through the detailed proofs in the next section will see that these assumptions are strictly speaking *not necessary* for the bulk of the analysis. They are made only to enable us to give 'closed-form' formulas, but if one is willing to forgo these, there is no need to make these additional assumptions.

Taken together, these theorems establish the following facts:

- On a sample of length l, the expected maximal segmental score if we allow different starting points for the two segments is twice the expected maximal segmental score if we insist that the starting points must match.

- The expected length of a maximal scoring segment is asymptotically twice as long in the case of M_l compared to R_l
- In either case, the empirical distribution of the symbols (x_i, y_j) on a maximal scoring segment is given by the conjugate distribution $\boldsymbol{\theta}$.

In order to state the next theorem, we need to introduce the notion of a 'lattice' random variable. We say that a random variable assuming values in a finite (or even a countably infinite) subset S of the real numbers $\mathbb{R}$ is a 'lattice' random variable if S is contained in an arithmetic progression. In other words, a random variable is a lattice random variable if all its possible values are of the form $a + md$ for real numbers a, d and integers m. In this case, the set $\{a + kd, k = 0, \pm 1, \pm 2, ...\}$ is the associated lattice. In the present instance, the random variable of interest is $F(\mathcal{X}, \mathcal{Y})$, the score as $\mathcal{X}, \mathcal{Y}$ vary over their respective sets. We shall first state the theorems and then discuss their implications.

Theorem 8.5 *Suppose the score $F(\mathcal{X}, \mathcal{Y})$ is not a lattice random variable. Then there exists a constant K^*, which can be estimated and in some cases computed explicitly, such that for all $x > 0$, the following inequality holds:*

$$\lim_{l \to \infty} \Pr\{R_l - \frac{\log l}{\lambda^*} \leq x\} = \exp(-K^* \exp(-\lambda^* x)). \tag{8.6}$$

In case the score $F(\mathcal{X}, \mathcal{Y})$ is a lattice random variable, the following statement is true:

$$\lim_{l \to \infty} \Pr\{R_l - \frac{\log l}{\lambda^*} \leq x_l\} \cdot \exp(K^* \exp(-\lambda^* x_l)) = 1 \tag{8.7}$$

whenever $\{x_l\}$ is a bounded sequence such that $x_l - \log l / \lambda^*$ belongs to the lattice for each value of l.

Theorem 8.6 *Suppose the score $F(\mathcal{X}, \mathcal{Y})$ is not a lattice random variable. Then there exists a constant K^*, which can be estimated and in some cases computed explicitly, such that for all $x > 0$, the following inequality holds:*

$$\lim_{l \to \infty} \Pr\{M_l - \frac{2 \log l}{\lambda^*} \leq x\} = \exp(-K^* \exp(-l^* x)). \tag{8.8}$$

In case the score $F(\mathcal{X}, \mathcal{Y})$ is a lattice random variable, the following statement is true:

$$\lim_{l \to \infty} \Pr\{M_l - \frac{2 \log l}{\lambda^*} \leq x_l\} \cdot \exp(K^* \exp(-\lambda^* x_l)) = 1 \tag{8.9}$$

whenever $\{x_l\}$ is a bounded sequence such that $x_l - \log l / \lambda^*$ belongs to the lattice for each value of l.

Note that (8.6) is the same as (8.7), just written differently. Thus the difference is that if $F(\mathcal{X}, Y)$ is a non-lattice variable, then we have a tail probability estimate for *every* value of x, whereas if $F(\mathcal{X}, \mathcal{Y})$ is a lattice random variable, we have a tail probability estimate only for *some* values of

x. Similar remarks apply to (8.8) and (8.9). This topic is discussed further in the next few paragraphs.

Equation (8.6) gives an extremely precise estimate of the rate at which the tail probability that R_l exceeds its expected value $\log l/\lambda^*$ by an amount x decays as x increases. The distribution on the right side of (8.6) is called a **Gumbel distribution (of Type I)**. Note that as $x \to \infty$, the exponential term $\exp(-l^* x)$ approaches zero, as result of which the right side of (8.6) approaches one. More precisely, suppose x is sufficiently large that

$$K^* \exp(-l^* x) \ll 1, \text{ or equivalently } x \gg \frac{\log K^*}{\lambda^*}.$$

Then, using the approximation $\exp(-\alpha) \approx 1 - \alpha$ when α is small, we can rewrite (8.6) as

$$\lim_{l\to\infty} \Pr\{R_l - \frac{\log l}{\lambda^*} > x\} \approx K^* \exp(-l^* x) \text{ whenever } x \gg \frac{\log K^*}{\lambda^*}. \tag{8.10}$$

Thus, while the formula (8.6) is extremely precise, in practice the tail probability decays exponentially with respect to x. Similar remarks apply to the tail estimate of M_l as well.

Note that all of the above equations from (8.6) through (8.9) give estimates of the *absolute excess* of the score from a maximal segment beyond its expected value. However, it is obvious that if we replace the scoring function F by some multiple cF, then we still have the same problem, and the maximal segments would still be the same. However, the expected value of the maximal segmental score would be scaled by the same factor c, and λ^* gets replaced by λ^*/c. Thus, in order to make the estimates more meaningful and 'scale-free', we should perhaps look at the excess *as a fraction of the expected value*, and not as an absolute excess. Accordingly, suppose

$$x = \alpha \frac{\log l}{\lambda^*}.$$

Then, after some routine algebra, the counterpart of (8.6) is

$$\lim_{l\to\infty} \Pr\{R_l - \frac{\log l}{\lambda^*} \leq \alpha \frac{\log l}{\lambda^*}\} = \exp(-K^* l^{-\alpha}). \tag{8.11}$$

Moreover, if

$$l \gg \frac{1}{\alpha} \log \frac{1}{K^*},$$

then

$$\Pr\{R_l - \frac{\log l}{\lambda^*} > \alpha \frac{\log l}{\lambda^*}\} \approx K^* l^{-\alpha}.$$

Similar modifications of (8.7) through (8.9) are routine and are left to the reader.

Now let us discuss the implications of the lattice vs. non-lattice variable. This discussion unfortunately borders on the pedantic, and the reader would miss very little by assuming that the relationship (8.10) always holds. However, since we have attempted to make mathematically precise statements in this text, we discuss this issue.

Quite often one would assign integer values to the scoring function; in other words, $F(x_i, y_j)$ is always assigned an integer value, as in Example 8.1 for instance. This would make $F(\mathcal{X}, \mathcal{Y})$ a lattice random variable with the spacing d equal to one. More generally, suppose all entries of $F(\mathcal{X}, \mathcal{Y})$ are specified to k significant decimal places. Then clearly every element of $F(\mathcal{X}, \mathcal{Y})$ is of the form $k_{ij} \cdot 10^{-k}$ for suitable integers k_{ij}, which would again make $F(\mathcal{X}, \mathcal{Y})$ a lattice random variable with the spacing $d = 10^{-k}$. The only way to assure that $F(\mathcal{X}, \mathcal{Y})$ is a non-lattice random variable is to specify the values of F to infinite precision, which makes no sense in practice. In this case, one cannot use the exact formulas (8.6) and (8.8). However, given any x, it is clear that we can always find a bounded sequence $\{x_n\}$ such that $x_n - d \leq x \leq x_n$ (where d is the lattice spacing) and $x_n + \log l/\lambda^*$ is a lattice point. Note that

$$\Pr\{R_l - \frac{\log l}{\lambda^*} > x_l - d\} \leq \Pr\{R_l - \frac{\log l}{\lambda^*} > x\} \leq \Pr\{R_l - \frac{\log l}{\lambda^*} > x_n\}.$$

Moreover, the two extreme probabilities do indeed satisfy the Gumbel type of tail probability estimate of the form (8.6). Hence, for all practical purposes, we need not worry about the distinction between lattice and non-lattice random variables.

8.2.4 Application of Main Results

A very nice discussion of the application of the above theorems to detecting similarity of sequences can be found in [3]. There are two distinct ways in which the theorems can be used. First, suppose the scoring function F is specified, and we are given two sample paths x_1^l, y_1^l of known statistics. Suppose we compute the maximal segmental score M_l. Now we wish to know whether the two sample paths are similar or not. If M_l significantly exceeds the expected score $2 \log l/\lambda^*$ by some quantity x, then we can compute the likelihood that this has happened purely by chance, by using the tail probability estimates in Theorem 8.6. This would allow us to say, with appropriate confidence, whether or not the segmental score M_l connotes sequence similarity.

The second application is to 'reverse engineer' the scoring function F itself. In a given problem, suppose we do not know what the 'right' scoring function is, but we *do have* at hand several pairs of similar sequences. So the problem then becomes one of choosing a scoring function F in such a way that these known answers are 'automatically' generated by the theory. In this context, Theorems 8.3 and 8.4 are useful. These theorems tell us that, on maximal scoring segments, the empirical distribution looks like $\boldsymbol{\theta}$. So we can proceed as follows: Suppose we are given several pairs of high-scoring segments. By definition, the two segments are of equal length. So we can just concatenate the various segments to generate one really huge segment of the form (x_1^l, y_1^l). And then:

- Construct the probability distribution $\boldsymbol{\phi}$ as the empirical distribution of the symbols in $\mathbb{A}$ amongst x_1^l. Similarly, construct the probability distribution $\boldsymbol{\psi}$ as the empirical distribution of the symbols in $\mathbb{B}$ amongst y_1^l.

- Similarly, compute the probability distribution $\boldsymbol{\theta}$ on $\mathbb{A} \times \mathbb{B}$ as the *joint empirical distribution* of the pair (a_i, b_j) on the sequence (x_1^l, y_1^l).
- Using $\boldsymbol{\mu} = \boldsymbol{\phi} \times \boldsymbol{\psi}$ and $\boldsymbol{\theta}$ as defined above, define the scoring function

$$F_{ij} := \frac{1}{c} \log \frac{\theta_{ij}}{\mu_{ij}} = \frac{1}{c} \log \frac{\theta_{ij}}{\phi_i \cdot \psi_j}. \tag{8.12}$$

  Here c is any constant that we choose. As discussed earlier, if we scale the scoring function uniformly by any constant, the problem remains unchanged.

It might be mentioned that the above approach to constructing scoring functions has been used extensively in the computational biology community.

8.3 BLAST THEORY: PROOFS OF MAIN RESULTS

In this section, we present the proofs of Theorems 8.3 and 8.4. Those who wish only to *use* BLAST theory can perhaps skip reading this section, but those who aspire to understand the basis of BLAST theory and to generalize it to other contexts may perhaps benefit from reading this section. Unfortunately the proofs of Theorems 8.5 and 8.6 are beyond the scope of the book. The interested reader is referred to [31] for the proof of Theorem 8.5 and [34] for the proof of Theorem 8.6. Even the proofs of Theorems 8.3 and 8.4 are quite complicated, as can be seen from the contents of this section. However, on the basis of these proofs, it may perhaps be possible to relax some of the assumptions underlying these two theorems, such as that the two sample paths x_1^l an y_1^l come from i.i.d. processes, and instead replace them by, say, sample paths of Markov processes.

The principal source for this section is the seminal paper by Dembo, Karlin and Zeitouni [33]. To assist the reader in following this paper, throughout this section we mention the corresponding theorem or lemma number from [33], and also highlight departures from their notation if any.

Let us recall the framework of the problem under study. We are given sets $\mathbb{A}, \mathbb{B}$ and probability distributions $\boldsymbol{\phi}$ on $\mathbb{A}$ and $\boldsymbol{\psi}$ on $\mathbb{B}$; for convenience let $\boldsymbol{\mu} := \boldsymbol{\phi} \times \boldsymbol{\psi}$ denote the corresponding product distribution on $\mathbb{A} \times \mathbb{B}$. We define the quantities

$$R_l := \max_{L \geq 0, 0 \leq i \leq l-L} \sum_{k=1}^{L} F(x_{i+k}, y_{i+k}).$$

and

$$M_l := \max_{L \geq 0, 0 \leq i,j \leq l-L} \sum_{k=1}^{L} F(x_{i+k}, y_{j+k}).$$

The objective is to answer the following questions:

1. Given sample paths of length l, what is the expected value of M_l or R_l?
2. Let L_l denote the length of a maximal scoring segment. What is the expected value of L_l? That is, how long is a maximally scoring segment on average, from a sample path of length l?
3. What is the empirical distribution of the symbols x_{i+k}, y_{j+k} in a maximally scoring segment?

Answering the last question raised earlier, namely to describe the tail probability distribution of the maximal segmental score beyond its expected value, is beyond the scope of this book. The interested reader is referred to the papers by Dembo and Karlin [31, 32, 67].

Given the set $\mathbb{C} = \mathbb{A} \times \mathbb{B}$, let $\mathcal{M}(\mathbb{C})$ denote the set of all probability distributions on this set. Thus $\mathcal{M}(\mathbb{C})$ can be identified with the nm-dimensional simplex $\mathbb{S}_{nm}$. If $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{C})$, we use the symbols $\boldsymbol{\nu}_{\mathbb{A}}, \boldsymbol{\nu}_{\mathbb{B}}$ to denote its marginal distributions on $\mathbb{A}, \mathbb{B}$ respectively.

To state and prove the theorems, we introduce a great deal of preliminary notation. First, for any $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{C})$, we define

$$H^*(\boldsymbol{\nu}\|\boldsymbol{\mu}) := \max\{0.5H(\boldsymbol{\nu}\|\boldsymbol{\mu}), H(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}}), H(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}})\}.$$

It is clear that $H^*(\boldsymbol{\nu}\|\boldsymbol{\mu}) > 0$ unless $\boldsymbol{\nu} = \boldsymbol{\mu}$. Next, for any given set $U \subseteq \mathcal{M}(\mathbb{C})$, define

$$J(\boldsymbol{\nu}) := \frac{E[F, \boldsymbol{\nu}]}{H^*(\boldsymbol{\nu}\|\boldsymbol{\mu})}, J(U) := \sup_{\boldsymbol{\nu} \in U} \max\{J(\boldsymbol{\nu}), 0\}.$$

Now by the standing hypothesis we have that $E[F, \boldsymbol{\mu}] < 0$. Hence, as $\boldsymbol{\nu} \to \boldsymbol{\mu}$, the quantity $E[F, \boldsymbol{\nu}]$ approaches some negative number, while $H^*(\boldsymbol{\nu}\|\boldsymbol{\mu})$ approaches zero. As a result $J(\boldsymbol{\nu}) \to -\infty$ as $\boldsymbol{\nu} \to \boldsymbol{\mu}$. If the set J does not contain any $\boldsymbol{\nu}$ such that $J(\boldsymbol{\nu}) > 0$, then by definition $J(U)$ is taken to equal zero.

The third piece of notation we need is the following: Suppose we are given the sample paths x_1^l, y_1^l, and that $L \geq 0, i, j \leq l - L$. Then we define $\hat{\boldsymbol{\mu}}(i, j, L)$ to be the pairwise empirical distribution on $\mathbb{C}$ defined by the sample $(x_{i+1}^{i+L}, y_{i+1}^{i+L})$. Similarly, given the sample path we define $\hat{\boldsymbol{\mu}}(i, \cdot, L)$ to be the empirical distribution on $\mathbb{A}$ defined by the sample x_{i+1}^{i+L}, and analogously we define $\hat{\boldsymbol{\mu}}(\cdot, j, L)$ to be the empirical distribution on $\mathbb{B}$ defined by the sample y_{i+1}^{i+L}. Since the set $\mathbb{C}$ has nm elements and the sample path has length L, it follows that the empirical distribution $\hat{\boldsymbol{\mu}}(i, j, L)$ belongs to the set $\mathcal{E}(L, nm)$ defined in Chapter 7. In the same way, $\hat{\boldsymbol{\mu}}(i, \cdot, L) \in \mathcal{E}(L, n)$ and $\hat{\boldsymbol{\mu}}(\cdot, j, L) \in \mathcal{E}(L, m)$.

Given a set $U \subseteq \mathcal{M}(\mathbb{C})$, define

$$M_l^U := \max\{\sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) : L \geq 0, 0 \leq i, j \leq l - L, \hat{\boldsymbol{\mu}}(i, j, L) \in U\}. \tag{8.13}$$

Thus M_l^U is exactly the same as M_l, except that the maximum is taken only over those segments of length L such that the corresponding pair empirical distribution $\hat{\boldsymbol{\mu}}(i,j,L)$ belongs to U. Other segments where $\hat{\boldsymbol{\mu}}(i,j,L)$ does not belong to U are not included in the computation of the maximum.

Now we state a 'global' theorem from which the desired results follow.

Theorem 8.7 *With the notation as above, we have that*

$$P_{\boldsymbol{\mu}}^l\{\limsup_{l\to\infty} \frac{M_l^U}{\log l} \leq J(U)\} \to 0 \text{ as } l \to \infty, \tag{8.14}$$

$$P_{\boldsymbol{\mu}}^l\{\liminf_{l\to\infty} \frac{M_l^U}{\log l} \geq J(U^o)\} \to 0 \text{ as } l \to \infty, \tag{8.15}$$

where U^o denotes the interior of the set U.

This theorem is a slightly weakened version of Theorem 3 of [33]. Their theorem states the following:

Theorem 8.8 *With the notation above, we have that*

$$J(U^o) \leq \liminf_{l\to\infty} \frac{M_l^U}{\log l} \leq \limsup_{l\to\infty} \frac{M_l^U}{\log l} \leq J(U), \tag{8.16}$$

where both inequalities hold almost surely.

What Theorem 8.7 claims is called 'convergence in probability', and we *can* prove *this* theorem using the methods we have developed thus far. It is good enough to allow us to study the behavior of the BLAST algorithm. We have no machinery to discuss 'almost sure convergence' as we have scrupulously 'avoided the infinite' in this book, so as to keep the exposition both rigorous as well as not so advanced. For those with the appropriate background, going from Theorem 8.7 to Theorem 8.8 is fairly straight-forward using the Borel-Cantelli lemma.

The proof of Theorem 8.7 proceeds via a series of preliminary steps. Throughout the sequel, the symbol λ^* denotes the unique positive solution to the equation $E[\exp(\lambda^* F), \boldsymbol{\mu}] = 1$, and $\boldsymbol{\theta}$ denotes the conjugate distribution of $\boldsymbol{\mu}$ as defined in (8.5). Also, to keep the notation simple, we will simply use the symbol P to denote the probability measure $P_{\boldsymbol{\mu}}^l$.

Lemma 8.9 *(Lemma 1 of [33]) Choose any $\lambda_0 \in (0, \lambda^*)$. Then, whenever*

$$L \geq -\frac{5 \log l}{\Lambda(\lambda_0)} =: L_0(l),$$

we have that

$$P\{\sup_{L \geq L_0(l)} \sup_{0 \leq i,j \leq L} \sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) \geq 0\} \leq 1/l^2.$$

Remark: This lemma allows us to focus our attention to segments of length $L \leq L_0(l)$, because the likelihood of having a positive segmental score on very long segments becomes vanishingly small as $l \to \infty$. Note that moment generating function $E[\exp(\lambda F), \boldsymbol{\mu}] \in (0,1)$ for $0 < \lambda < \lambda^*$. Hence the logarithmic moment generating function $\Lambda(\lambda) < 0$ for $0 < \lambda < \lambda^*$, and $L_0(l) > 0$. Note that hereafter we suppress the dependence of L_0 on l.

Proof. Since there are at most l^3 possible choices of i, j, L, the conclusion follows if it can be shown that

$$P\{\sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) \geq 0\} \leq 1/l^5 \; \forall L \geq L_0.$$

To establish this inequality, apply Markov's inequality as in Corollary 2.26 with $\epsilon = 0$. This leads to

$$\begin{aligned} P\{\sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) \geq 0\} &\leq E\left[\exp\left(\lambda_0 \sum_{k=1}^{L} F(x_{i+k}, y_{i+k})\right), \mathcal{P}_{\boldsymbol{\mu}}^L\right] \\ &\leq E\left[\prod_{k=1}^{L} \exp(\lambda_0 F(x_{i+k}, y_{i+k})), \mathcal{P}_{\boldsymbol{\mu}}^L\right] \\ &= (E[\exp(\lambda_0 F(x,y)), P_{\boldsymbol{\mu}}])^L \\ &= (\exp(\Lambda(\lambda_0))^L = \exp(L\Lambda(\lambda_0)) \\ &\leq \exp(-5\log l) = 1/l^5. \end{aligned}$$

□

Lemma 8.10 *(Lemma 2 of [33]) Let $\bar{M}_l^U$ be the same as M_l^U, except that $L \leq L_0(l)$. Suppose $J_U > 0$. Then for all $t > 1$, we have that*

$$P\{\bar{M}_l^U \geq tJ_U \log l\} \leq \frac{(L_0+1)^{nm}}{l^{t-1}}.$$

Remark: This lemma is a counterpart to Lemma 8.9. In that lemma we could say that the maximal segmental length on very long segment (defined as L exceeding L_0) will not even exceed zero, with high probability. In the present lemma we examine the complementary situation where put an upper bound of L_0 on the length of the segments, and call the resulting maximal segmental length as $\bar{M}_l^U$.

Proof. For each $\boldsymbol{\nu} \in \mathcal{E}(L, nm)$, let $A(\boldsymbol{\nu}, L)$ denote the event

$$A(\boldsymbol{\nu}, L) := \{\exists i, j, 0 \leq i, j \leq l - L, \hat{\boldsymbol{\mu}}(i,j,L) = \boldsymbol{\nu}\}.$$

We recognize that $A(\boldsymbol{\nu}, L)$ is the event that some segment of length L generates the empirical distribution $\boldsymbol{\nu}$. Now an upper bound for the likelihood of the event $A(\boldsymbol{\nu}, L)$ is readily available from the earlier discussion in Chapter 7, specifically from the method of types upper bound. If $\hat{\boldsymbol{\mu}}(i,j,L) = \boldsymbol{\nu}$, then definitely the marginals also match, so that $\hat{\boldsymbol{\mu}}_{\mathbb{A}}(i,j,L) = \boldsymbol{\nu}_{\mathbb{A}}$ and $\hat{\boldsymbol{\mu}}_{\mathbb{B}}(i,j,L) = \boldsymbol{\nu}_{\mathbb{B}}$. For any fixed i, j, we have the method of types bounds

$$P\{\hat{\boldsymbol{\mu}}(i,j,L) = \boldsymbol{\nu}\} \leq \exp(-LH(\boldsymbol{\nu}\|\boldsymbol{\mu})),$$

$$P\{\hat{\boldsymbol{\mu}}_{\mathbb{A}}(i,j,L) = \boldsymbol{\nu}_{\mathbb{A}}\} \leq \exp(-LH(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}})),$$

$$P\{\hat{\boldsymbol{\mu}}_{\mathbb{B}}(i,j,L) = \boldsymbol{\nu}_{\mathbb{B}}\} \leq \exp(-LH(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}})).$$

Now there are l^2 possible choices for the pair (i,j), and l choices for i and j respectively. So we can write

$$\begin{aligned} P\{A(\boldsymbol{\nu},L)\} \leq \min\{ & l^2 \exp(-LH(\boldsymbol{\nu}\|\boldsymbol{\mu})), \\ & l \exp(-LH(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}})), l \exp(-LH(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}})), 1\}. \end{aligned}$$

The last term of 1 reflects the obvious fact that every probability is bounded above by one. Now let us note that

$$l^2 \exp(-LH(\boldsymbol{\nu}\|\boldsymbol{\mu})) = [l \exp(-0.5H(\boldsymbol{\nu}\|\boldsymbol{\mu}))]^2,$$

and use the inequality $\min\{a^2, 1\} \leq a$ for all $a > 0$. This leads to the final estimate

$$P\{A(\boldsymbol{\nu},L)\} \leq l \exp(-LH^*(\boldsymbol{\nu}\|\boldsymbol{\mu})). \tag{8.17}$$

We shall make use of this inequality many times in the remainder of the proof. Returning to the above, we conclude that, as a consequence,

$$LH^*(\boldsymbol{\nu}\|\boldsymbol{\mu}) \geq t \log l \;\Rightarrow\; P\{A(\boldsymbol{\nu},L)\} \leq l^{-(t-1)}. \tag{8.18}$$

Next, observe that along any segment of length L, we have

$$\sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) = LE[F, \hat{\boldsymbol{\mu}}(i,j,L)].$$

So the event $\{\bar{M}_l^U \geq tJ_U \log l\}$ is contained in the union of the events

$$\hat{\boldsymbol{\mu}}(i,j,L) = \boldsymbol{\nu} \in U \cap \mathcal{E}(L, nm), LE[F, \boldsymbol{\nu}] \geq tJ(\boldsymbol{\nu}) \log l. \tag{8.19}$$

But the latter inequality implies that

$$LH(\boldsymbol{\nu}\|\boldsymbol{\mu}) = L\frac{E[F,\boldsymbol{\nu}]}{J(\boldsymbol{\nu})} \geq t \log l.$$

In turn, from the definition of the function H^*, the above inequality implies that

$$LH^*(\boldsymbol{\nu}\|\boldsymbol{\mu}) \geq t \log l.$$

So whenever the events in (8.19) hold, we can infer from (8.18) that

$$P\{\sum_{k=1}^{L} F(x_{i+k}, y_{i+k}) \geq tJ_U \log l\} \leq l^{-(t-1)} \;\forall \boldsymbol{\nu} \in U \cap \mathcal{E}(L, nm).$$

Since $|\mathcal{E}(L, nm)| \leq (L+1)^{nm} \leq (L_0+1)^{nm}$, the desired conclusion follows. □

Lemma 8.11 *(Lemma 3 of [33]) For any $\boldsymbol{\nu} \in \mathcal{E}(L, nm)$ and any $l \geq L$, we have*

$$\begin{aligned} 1 - P\{A(\boldsymbol{\nu},L)\} \leq & \, 4(L+1)^{nm+1} l_*^{-2} \exp(LH(\boldsymbol{\nu}\boldsymbol{\mu})) \\ & + (L+1)^{nm} l_*^{-1} \exp(LH(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}})) \\ & + (L+1)^{nm} l_*^{-1} \exp(LH(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}})), \end{aligned} \tag{8.20}$$

where $l_* = l\lfloor l/L \rfloor$ is the largest integer multiple of L not exceeding l.

Proof. Clearly, for a fixed $\boldsymbol{\nu}, L$, the quantity $P\{A(\boldsymbol{\nu}, l)\}$ is a nondecreasing function of l. So we can just replace l by l_*; that is, we can assume that $l = ML$ for some integer M.

Divide the sample path $x_1^l = x_1^{ML}$ into M blocks of length L each, and do the same for y_1^l. Given $\boldsymbol{\nu} \in \mathcal{E}(L, nm)$, let N_x denote the number of times that the empirical distribution of a block of length L precisely equals $\boldsymbol{\nu}_{\mathbb{A}}$. In other words, let

$$N_x := \sum_{i=1}^{M} I_{\{\hat{\boldsymbol{\mu}}((i-1)L+1,\cdot,L)=\boldsymbol{\nu}_{\mathbb{A}}\}}.$$

Define the integer N_y analogously. Let p_x denote the probability that the empirical distribution of a block of length L is equal to $\boldsymbol{\nu}_{\mathbb{A}}$. We can even write down a formula for p_x, but it is not necessary. Now N_x is the sum of M independent binary variables, each of which assumes a value of 1 with the probability p_x. Hence N_x is a random variable assuming values in the range $\{0, 1, \ldots, M\}$ with a binomial distribution corresponding to p_x. Similar remarks apply to N_y and p_y.

Now let B_{ij} be the event that, on the i-th $\mathcal{X}$-block with empirical distribution $\boldsymbol{\nu}_{\mathbb{A}}$ and the j-th $\mathcal{Y}$-block with empirical distribution $\boldsymbol{\nu}_{\mathbb{B}}$, the joint empirical distribution of $(\mathcal{X}, \mathcal{Y})$ equals $\boldsymbol{\nu}$. Let p denote the probability that the event B_{ij} occurs. Thus

$$p = \{\hat{\boldsymbol{\mu}}(i, j, L) = \boldsymbol{\nu} | \hat{\boldsymbol{\mu}}_{\mathbb{A}} = \boldsymbol{\nu}_{\mathbb{A}} \& \hat{\boldsymbol{\mu}}_{\mathbb{B}} = \boldsymbol{\nu}_{\mathbb{B}}\}.$$

Let us define

$$W = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} I_{B_{ij}}.$$

Then the above definitions imply that

$$E[W | N_x, N_y] = p N_x N_y.$$

Also, the variance of W is bounded by

$$\text{var}[W | N_x, N_y] = N_x N_y \text{var}[B_{ij} | N_x, N_y] = N_x N_y p(1-p) \leq p N_x N_y.$$

So by Chebycheff's inequality, it follows that

$$P\{W = 0 | N_x, N_y\} \leq \frac{\text{var}[W | N_x, N_y]}{(E[W | N_x, N_y])^2} \leq \frac{1}{p N_x N_y}.$$

We can also write

$$P\{W = 0\} = E[P\{W = 0 | N_x \geq 1 \& N_y \geq 1\}] + P\{N_x = 0\} + P\{N_y = 0\}.$$

Now it is clear that

$$P\{N_x = 0\} = (1 - p_x)^M, P\{N_y = 0\} = (1 - p_y)^M.$$

As for the first term, we have

$$\begin{aligned} E[P\{W = 0 \mid N_x \geq 1 \& N_y \geq 1\}] &\leq E\left[\frac{1}{p N_x N_y} \middle| N_x \geq 1 \& N_y \geq 1 \right] \\ &= \frac{1}{p} E[1/N_x | N_x \geq 1] \cdot E[1/N_y | N_y \geq 1], \end{aligned}$$

where we take advantage of the fact that N_x and N_y are independent random variables. Now it can be readily verified from the binomial distribution of N_x that

$$E[1/N_x|N_x \geq 1] = \sum_{i=1}^{M} \frac{1}{i} \left(\begin{array}{c} M \\ i \end{array} \right) p_x^i (1-p_x)^{M-i} \leq \frac{2}{Mp_x}.$$

Similarly

$$E[1/N_y|N_y \geq 1] \leq \frac{2}{Mp_y}.$$

This leads to

$$P\{W=0\} \leq \frac{4}{M^2 pp_x p_y} + (1-p_x)^M + (1-p_y)^M.$$

A cruder estimate is

$$P\{W=0\} \leq \frac{4}{M^2 pp_x p_y} + \frac{1}{Mp_x} + \frac{1}{Mp_y}. \tag{8.21}$$

Now observe from the method of types lower bounds that

$$pp_x p_y = P\{\hat{\boldsymbol{\mu}}(1,1,L) = \boldsymbol{\nu}\} \geq (L+1)^{-nm} \exp(-LH(\boldsymbol{\nu}\|\boldsymbol{\mu})), \tag{8.22}$$

$$p_x = P\{\hat{\boldsymbol{\mu}}(1,\cdot,L) = \boldsymbol{\nu}_{\mathbb{A}}\} \geq (L+1)^{-nm} \exp(-LH(\boldsymbol{\nu}_{\mathbb{A}}\|\boldsymbol{\mu}_{\mathbb{A}})), \tag{8.23}$$

$$p_y = P\{\hat{\boldsymbol{\mu}}(1,\cdot,L) = \boldsymbol{\nu}_{\mathbb{B}}\} \geq (L+1)^{-nm} \exp(-LH(\boldsymbol{\nu}_{\mathbb{B}}\|\boldsymbol{\mu}_{\mathbb{B}})). \tag{8.24}$$

Now observe that the event $\{W > 0\}$ implies the event $\{A(\boldsymbol{\nu}, L)\}$. Hence

$$P\{A(\boldsymbol{\nu}, L)\} \geq P\{W > 0\} = 1 - P\{W = 0\},$$

which in turn implies that

$$1 - P\{A(\boldsymbol{\nu}, L)\} \leq P\{W = 0\}.$$

The desired conclusion now follows from substituting the bounds from (8.22) through (8.24) into (8.21). □

Lemma 8.12 *(Lemma 4 of [33]) Suppose $J_{U^o} > 0$. Then for each $t < 1$, there exists an integer $l_0 = l_0(t)$*[1] *such that*

$$P\{M_l^U \leq t \log l J_{U^o}\} \leq \frac{1}{l^{(1-t)/2}}, \ \forall l \geq l_0.$$

Proof. Suppose $t < 1$, and choose a number $\tau \in (t, (1+t)/2)$. Because $\tau > t$, it follows that

$$\frac{\tau + t}{2\tau} = 0.5 + 0.5 \frac{t}{\tau} < 1.$$

Given the set U, choose $\tilde{\boldsymbol{\nu}} \in U^o$ such that

$$J(\tilde{\boldsymbol{\nu}}) > \frac{\tau + t}{2\tau} J_{U^o}.$$

[1] Not to be confused with L_0

Define

$$k_l = \lceil \tau \log l / H^*(\tilde{\boldsymbol{\nu}} \| \boldsymbol{\mu}) \rceil,$$

and note that eventually k_l eventually less than any constant multiple of l, because $k_l \sim$ const. $\cdot \log l$. Hence, for large enough l, we can always choose $\tilde{\boldsymbol{\nu}}_l \in \mathcal{E}(k_l, nm)$ such that $\rho(\boldsymbol{\nu}, \tilde{\boldsymbol{\nu}}_l) \leq nm/k_l$, where ρ denotes the total variation metric. Now let c denote an upper bound for the function $F(x,y), x \in \mathbb{A}, y \in \mathbb{B}$. Then

$$\begin{aligned} k_l E[F, \tilde{\boldsymbol{\nu}}_l] &\geq k_l(E[F, \boldsymbol{\nu}] - c\rho(\boldsymbol{\nu}, \tilde{\boldsymbol{\nu}}_l)) \\ &\geq k_l E[F, \boldsymbol{\nu}] - cnm. \end{aligned}$$

Note that cnm is just some constant independent of l. Hence it follows from the definition of the constant k_l that

$$\begin{aligned} k_l E[F, \tilde{\boldsymbol{\nu}}_l] &\geq k_l E[F, \boldsymbol{\nu}] - cnm \\ &\geq \tau \log l \frac{E[F, \boldsymbol{\nu}]}{H^*(\tilde{\boldsymbol{\nu}} \| \boldsymbol{\mu})} - cnm \\ &= \tau \log l J(\tilde{\boldsymbol{\nu}}) - cnm \\ &\geq \frac{\tau + t}{2} \log l J_{U*o} - cnm. \end{aligned}$$

It is clear that $\tilde{\boldsymbol{\nu}}_l \to \tilde{\boldsymbol{\nu}}$ as $l \to \infty$, and $\tilde{\boldsymbol{\nu}} \in U^o$. Hence $\tilde{\boldsymbol{\nu}}_l \in U^o$ for all large enough l. Moreover, since $(\tau + t)/2 > t$, the constant term cnm can be neglected in comparison with the $\log l$ term. Hence for large enough l it can be said that

$$k_l E[F, \tilde{\boldsymbol{\nu}}_l] > t \log l J_{U*o}.$$

Since $\tilde{\boldsymbol{\nu}}_l \in \mathcal{E}(k_l, nm)$ for each l, this means that the event $\{M_l^U \leq t \log l J_{U*o}\}$ is contained in the complement of the event $\{A(\tilde{\boldsymbol{\nu}}_l, k_l)\}$, i.e. that there exists a sequence of length k_l whose empirical distribution is $\tilde{\boldsymbol{\nu}}_l$. Thus

$$P\{M_l^U \leq t \log l J_{U*o}\} \leq 1 - P\{A(\tilde{\boldsymbol{\nu}}_l, k_l)\}. \tag{8.25}$$

Now an upper bound for the right side is given by Lemma 8.11, specifically (8.20). There are three terms in this bound, out of which one contains l^{-2} and thus decays faster than the other two terms which contain l^{-1}. Next, observe that, with l replaced by k_l, we have

$$\exp(k_l H^*(\boldsymbol{\nu} \| \boldsymbol{\mu})) \sim \exp(\tau \log l) = l^\tau.$$

Hence

$$1 - P\{A(\tilde{\boldsymbol{\nu}}_l, k_l)\} \leq \text{const.} l^{-(1-\tau)}.$$

But since $\tau < (1+t)/2$, it follows that $1 - \tau > (1-t)/2$. So we conclude that

$$[1 - P\{A(\tilde{\boldsymbol{\nu}}_l, k_l)\}] l^{(1-t)/2} \to 0 \text{ as } l \to \infty.$$

In view of the bound (8.25), this in turn implies that

$$P\{M_l^U \leq t \log l J_{U*o}\} l^{(1-t)/2} \to 0 \text{ as } l \to \infty.$$

This is the desired conclusion. □

Now we are in a position to prove Theorem 8.7. What has been proven thus far is that:

$$P\{\bar{M}_l^U \geq tJ_U \log l\} \to 0 \text{ as } l \to \infty, \ \forall t > 1,$$

$$P\{M_l^U \leq t \log l J_{U^o}\} \to 0 \text{ as } l \to \infty, \ \forall t < 1.$$

Taken together, these two assertions are precisely the desired conclusions.

Chapter Nine

Hidden Markov Processes

In this chapter, we study a special type of stochastic process that forms the main focus of this book, called a 'hidden' Markov process (HMP).' Some authors also use the expression 'hidden Markov model (HMM).' In this book we prefer to say 'A process $\{\mathcal{Y}_t\}$ *is* a hidden Markov process' or 'A process $\{\mathcal{Y}_t\}$ *has* a hidden Markov model.' We use the two expressions interchangeably.

The chapter is organized as follows: In Section 9.1 we present three distinct types of HMM's, and show that they are all equivalent from the standpoint of their expressive power or modelling ability. In Section 9.2 we study various issues related to the computation of likelihoods in a HMM.

In the remainder of the book, the acronyms HMP for hidden Markov process and HMM for hidden Markov model are used freely.

9.1 VARIOUS TYPES OF HIDDEN MARKOV MODELS AND THEIR EQUIVALENCE

In this section we formulate three distinct types of hidden Markov models, and then show that they are all equivalent from the standpoint of their expressive power. This discussion becomes relevant because each of these models appears in the literature. Hence it is important to realize that, while these models may *appear* to be different, in fact each of the models can be transformed to any of the other two.

9.1.1 Three Different-Looking Models

Definition 9.1 *Suppose* $\{\mathcal{Y}_t\}_{t=1}^{\infty}$ *is a stationary stochastic process assuming values in a finite set* $\mathbb{M} = \{1, \ldots, m\}$.[1] *We say that* $\{\mathcal{Y}_t\}$ *has a* **Type 1 hidden Markov model**, *or a* **HMM of the deterministic function of a Markov chain type**, *if there exists a stationary Markov process* $\{\mathcal{X}_t\}_{t=0}^{\infty}$ *over a finite state space* $\mathbb{N} = \{1, \ldots, n\}$ *and a function* $f : \mathbb{N} \to \mathbb{M}$ *such that* $\mathcal{Y}_t = f(\mathcal{X}_t)$.[2]

[1] As always, we really mean to say that $\mathcal{Y}_t$ assumes values in a finite set $\{y_1, \ldots, y_m\}$ consisting of just abstract labels. We write the set as $\{1, \ldots, m\}$ in the interests of simplifying the notation. However, these elements should be viewed as just labels and not as integers.

[2] Here again, we really mean that $\mathcal{X}_t$ assumes values in a finite set $\{x_1, \ldots, x_n\}$.

From a historical perspective HMM's of Type 1 are the earliest to be introduced into the literature; see [20, 53]. Note that we must have $n \geq m$ in order for the above definition to make sense. If $n < m$, then some elements of the set $\mathbb{M}$ cannot be images of any element of $\mathbb{N}$, and can therefore be deleted from the output space. Observe also that the requirement that $\{\mathcal{X}_t\}$ must have a *finite* state space is crucial. Carlyle [24] has shown that *every* stochastic process $\{\mathcal{Y}_t\}$ over a finite output space can be expressed in the form $\{f(\mathcal{X}_t)\}$ where $\{\mathcal{X}_t\}$ is a Markov process whose state space is *countably infinite.*

Definition 9.2 *Suppose $\{\mathcal{Y}_t\}_{t=1}^{\infty}$ is a stationary stochastic process assuming values in a finite set $\mathbb{M} = \{y_1, \ldots, y_m\}$. We say that $\{\mathcal{Y}_t\}$ has a* **hidden Markov model of Type 2***, or a* **HMM of the random function of a Markov chain type***, if there exist a finite integer n, a pair of matrices $A \in [0,1]^{n \times n}$, $B \in [0,1]^{n \times m}$, and a probability distribution $\pi \in \mathbb{S}_n$, such that the following properties hold:*

1. *A and B are both stochastic matrices. Thus each row of A and each row of B add up to one, or equivalently*
$$A\mathbf{e}_n = \mathbf{e}_n, B\mathbf{e}_m = \mathbf{e}_n. \tag{9.1}$$

2. *π is a stationary distribution of A; that is, $\pi A = \pi$.*

3. *Suppose $\{\mathcal{X}_t\}$ is a homogeneous Markov chain on the state space $\mathbb{N} = \{1, \ldots, n\}$ with the initial distribution π and state transition matrix A. Thus*
$$\Pr\{\mathcal{X}_0 = i\} = \pi_i, \text{ and } \Pr\{\mathcal{X}_{t+1} = j | \mathcal{X}_t = i\} = a_{ij}, \ \forall i, j, t. \tag{9.2}$$
Suppose that, at each time instant t, the random variable $\mathcal{Z}_t$ is selected according to the rule
$$\Pr\{\mathcal{Z}_t = u | \mathcal{X}_t = j\} = b_{ju}, \ \forall j \in \mathbb{N}, u \in \mathbb{M}, t \geq 0. \tag{9.3}$$
Then $\{\mathcal{Z}_t\}$ has the same law as $\{\mathcal{Y}_t\}$.

Thus, in a Type 2 HMM, the current output $\mathcal{Y}_t$ can be viewed as a 'random' function of the current state $\mathcal{X}_t$, according to the rule (9.3).[3] The Type 2 HMM was apparently first introduced into the literature in [13]. Note that, in contrast to a Type 1 HMM, the Type 2 HMM remains meaningful even if $m > n$.

In the engineering world, the expressions HMP and HMM are invariably reserved for a Type 2 HMP or HMM. According to [43], a Type 1 HMM is referred to as a 'Markov source' in the world of communication theory.

[3] Note that the distinction between $\mathcal{Z}_t$ and $\mathcal{Y}_t$ is out of respect to the niceties of mathematical expression and is perhaps pedantic. We really cannot say '$\mathcal{Y}_t$ *is* generated by the rule (9.3).' Instead we say 'A random variable $\mathcal{Z}_t$ selected according to the rule is indistinguishable from $\mathcal{Y}_t$.' Having said that, we will hereafter ignore the distinction.

Definition 9.3 *Suppose $\{\mathcal{Y}_t\}_{t=1}^{\infty}$ is a stationary stochastic process assuming values in a finite set $\mathbb{M} = \{y_1, \ldots, y_m\}$. We say that $\{\mathcal{Y}_t\}$ is a* **hidden Markov model of Type 3**, *or a* **HMM of the joint Markov process type** *if there exist a finite set $\mathbb{N} = \{1, \ldots, n\}$ and a stationary stochastic process $\{\mathcal{X}_t\}$ assuming values in $\mathbb{N}$ such that the following properties hold:*

1. *The joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov.*
2. *In addition*

$$\Pr\{(\mathcal{X}_t, \mathcal{Y}_t)|(\mathcal{X}_{t-1}, \mathcal{Y}_{t-1})\} = \Pr\{(\mathcal{X}_t, \mathcal{Y}_t)|\mathcal{X}_{t-1}\}. \tag{9.4}$$

In other words

$$Pr\{(\mathcal{X}_t, \mathcal{Y}_t) = (j, u)|(\mathcal{X}_{t-1}, \mathcal{Y}_{t-1}) = (i, v)\}$$
$$= Pr\{(\mathcal{X}_t, \mathcal{Y}_t) = (j, u)|\mathcal{X}_{t-1} = i\}, \ \forall i, j \in \mathbb{N}, u, v \in \mathbb{M}. \tag{9.5}$$

Note that in a Type 3 HMM, the associated process $\{\mathcal{X}_t\}$ is Markov by itself. The distinction between a Type 2 HMM and a Type 3 HMM is brought out clearly by comparing (9.3) and (9.4). In a Type 2 HMM, the current output $\mathcal{Y}_t$ is a 'random function' of the *current* state $\mathcal{X}_t$, whereas in a Type 3 HMM, the current output $\mathcal{Y}_t$ is a 'random function' of the *previous* state $\mathcal{X}_{t-1}$. Of course, in a Type 1 HMM, the current output $\mathcal{Y}_t$ is a 'deterministic function' of the current state $\mathcal{X}_t$, in contrast with a Type 2 HMM where $\mathcal{Y}_t$ is a 'random function' of the current state $\mathcal{X}_t$.

Whereas Type 1 and Type 2 HMM's are historical and widely used in the literature, the Type 3 HMM is somewhat nonstandard and appears to have been introduced in [6]. But a Type 3 HMM has two significant advantages over the other types of HMM's. First, as shown in the next subsection, a Type HMM in general requires a smaller state space compared to the other two types of HMM's. Second, when we study realization theory in Chapter 10, the proofs become very streamlined if we use a Type 3 HMM.

9.1.2 Equivalence Between the Three Models

The objective of this subsection is to show that in fact all three types of HMM's are the same in terms of their expressive power. However, when it comes to the 'economy' of the model as measured by the size of the state space of the associated Markov process, the Type 3 HMM is the most economical whereas the Type 1 HMM is the least economical. A Type 2 HMM lies in between.

Theorem 9.4 *The following statements are equivalent:*

(i) The process $\{\mathcal{Y}_t\}$ has a Type 1 HMM (that is, a HMM of the 'deterministic function of a Markov chain' type).

(ii) The process $\{\mathcal{Y}_t\}$ has a Type 2 HMM (that is, a HMM of the 'random function of a Markov chain' type).

(iii) The process $\{\mathcal{Y}_t\}$ has a Type 3 HMM (that is, a HMM of the 'joint Markov process' type).

Proof. **(i) ⇒ (ii)** Clearly every deterministic function of a Markov chain is also a 'random' function of the same Markov chain, with every element of B equal to zero or one. Precisely, since both $\mathbb{N}$ and $\mathbb{M}$ are finite sets, the function f simply induces a partition of the state space $\mathbb{N}$ into m subsets $\mathbb{N}_1, \ldots, \mathbb{N}_m$, where $\mathbb{N}_u := \{i \in \mathbb{N} : f(i) = u\}$. Thus two states in $\mathbb{N}_u$ are indistinguishable through the measurement process $\{\mathcal{Y}_t\}$. Now set $b_{ju} = 1$ if $j \in \mathbb{N}_u$ and zero otherwise.

(ii) ⇒ (iii) If $\{\mathcal{Y}_t\}$ is modelled as a Type 2 HMM with $\{\mathcal{X}_t\}$ as the associated Markov chain, then the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov. Indeed, if we define $(\mathcal{X}_t, \mathcal{Y}_t) \in \mathbb{N} \times \mathbb{M}$, then it readily follows from the HMM conditions that

$$\Pr\{(\mathcal{X}_{t+1}, \mathcal{Y}_{t+1}) = (j, u) | (\mathcal{X}_t, \mathcal{Y}_t) = (i, v)\} = a_{ij} b_{ju},$$

and is therefore independent of v. Now define

$$M^{(u)} := [a_{ij} b_{ju}] \in [0, 1]^{n \times n}.$$

Then the process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov, and its state transition matrix is given by

$$\begin{bmatrix} M^{(1)} & M^{(2)} & \ldots & M^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ M^{(1)} & M^{(2)} & \ldots & M^{(m)} \end{bmatrix}.$$

Finally, note that the probability that $(\mathcal{X}_{t+1}, \mathcal{Y}_{t+1}) = (j, u)$ depends only on $\mathcal{X}_t$ but not on $\mathcal{Y}_t$. Hence the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ satisfies all the conditions required of the Type 3 HMM.

(iii) ⇒ (i) Suppose $\{\mathcal{Y}_t\}$ has a Type 3 HMM, and suppose $\mathcal{X}_t$ is a Markov process such that the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is also Markov. Then clearly $\mathcal{Y}_t = f[(\mathcal{X}_t, \mathcal{Y}_t)]$ for a suitable function f. Hence this is also a Type 1 HMM. □

Up to now we have considered only the 'expressive power' of the various HMM types. However, this is only part of the problem of stochastic modelling. An equally, if not more, important issue is the 'economy' of the representation, that is, the size of the state space of the associated Markov chain. To study this issue, let us use the shorthand expression '$\{\mathcal{Y}_t\}$ has a Type 1 (or 2 or 3) HMM of size n' if the Markov process $\{\mathcal{X}_t\}$ associated with the HMM of the appropriate type evolves over a set $\mathbb{N}$ of cardinality n. Then the next result is almost a direct consequence of the proof of Theorem 9.4, and shows that a Type 3 HMM is the most economical, while a Type 1 HMM is the least economical.

Corollary 9.5 *Suppose $\{\mathcal{Y}_t\}_{t=1}^{\infty}$ is a stationary stochastic process assuming values in a finite set $\mathbb{M} = \{y_1, \ldots, y_m\}$.*

1. *Suppose $\{\mathcal{Y}_t\}$ has a Type 1 HMM of size n. Then it has a Type 2 HMM of size n.*

2. *Suppose $\{\mathcal{Y}_t\}$ has a Type 2 HMM of size n. Then it has a Type 3 HMM of size n.*

The next lemma discusses the extent to which the above statements can be reversed.

Lemma 9.6 *Suppose $\{\mathcal{Y}_t\}_{t=1}^{\infty}$ is a stationary stochastic process assuming values in a finite set $\mathbb{M} = \{y_1, \ldots, y_m\}$.*

(i) Suppose a process $\{\mathcal{Y}_t\}$ has a HMM of the random function of a Markov chain type, and let $\{\mathcal{X}_t\}$ denote the associated Markov chain. Let A and B denote respectively the state transition matrix and output matrix of the HMM. Then $\mathcal{Y}_t$ is a deterministic function of $\mathcal{X}_t$ if and only if every element of the matrix B is either zero or one.

(ii) Suppose a process $\{\mathcal{Y}_t\}$ has a HMM of the joint Markov process type, and let $\{\mathcal{X}_t\}$ denote the associated Markov chain. Define

$$m_{ij}^{(u)} := \Pr\{\mathcal{X}_t = i \& \mathcal{Y}_t = u | \mathcal{X}_{t-1} = i\}, \ \forall i, j \in \mathbb{N}, u \in \mathbb{M}, \tag{9.6}$$

$$M^{(u)} := [m_{ij}^{(u)}, i, j \in \mathbb{N}], \ \forall u \in \mathbb{M}, \tag{9.7}$$

and

$$a_{ij} := \sum_{u \in \mathbb{M}} m_{ij}^{(u)}, \ \forall i, j. \tag{9.8}$$

Then $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$ (and not just $\mathcal{X}_{t-1}$) if and only if the following consistency conditions hold: If $a_{ij} \neq 0$, then the ratio

$$\frac{m_{ij}^{(u)}}{a_{ij}}$$

is independent of i.

Proof. The first statement is obvious. Let us consider the second statement. Suppose the process $\{\mathcal{Y}_t\}$ has a joint Markov process type of HMM, and let $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ be the associated Markov process. Define the matrices $M^{(u)}$ as in (9.6). Then we already know that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_{t-1}$. The aim is to show that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$ (and not just $\mathcal{X}_{t-1}$) if and only if the stated condition holds.

'Only if': By assumption, $\{\mathcal{X}_t\}$ is a Markov process. Moreover, we have that

$$\Pr\{\mathcal{X}_t = j | \mathcal{X}_{t-1} = i\} = \sum_{u \in \mathbb{M}} \Pr\{(\mathcal{X}_t, \mathcal{Y}_t) = (j, u) | \mathcal{X}_{t-1} = i\} = \sum_{u \in \mathbb{M}} m_{ij}^{(u)}.$$

Therefore A is the state transition matrix of the Markov process $\{\mathcal{X}_t\}$. Now suppose that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$, and not just $\mathcal{X}_{t-1}$, and define

$$b_{ju} := \Pr\{\mathcal{Y}_t = u | \mathcal{X}_t = j\}, \ \forall u \in \mathcal{M}, j \in \mathbb{N}.$$

Then we must have $m_{ij}^{(u)} = a_{ij} b_{ju}$ for all i, j, u. If $a_{ij} = 0$ for some i, j, then perforce $m_{ij}^{(u)} = 0 \; \forall u \in \mathcal{M}$. Suppose $a_{ij} \neq 0$. Then it is clear that

$$b_{ju} = \frac{m_{ij}^{(u)}}{a_{ij}} \; \forall i$$

and is therefore independent of i.

'If': This consists of simply reversing the arguments. Suppose the ratio is indeed independent of i, and define b_{ju} as above. Then clearly $m_{ij}^{(u)} = a_{ij} b_{ju}$ and as a result $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$. □

As a simple example, suppose $n = m = 2$,

$$M^{(1)} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}, M^{(2)} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}, A = \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}.$$

Then

$$\frac{m_{11}^{(1)}}{a_{11}} = 5/7, \; \frac{m_{21}^{(1)}}{a_{21}} = 1/2 \neq 5/7.$$

Hence, while $\mathcal{Y}_t$ is a random function of $\mathcal{X}_{t-1}$, it is *not* a random function of $\mathcal{X}_t$.

9.2 COMPUTATION OF LIKELIHOODS

To place the contents of this section in perspective, let us observe that in the HMM world (and it does not matter what the type of the HMM is), we can identify three entities, namely: the model, the output sequence, and the state sequence. Given any two of these entities, we can seek to determine the third. So we can ask three distinct questions.

1. Given a HMM and a state sequence, what is the likelihood of observing a particular output sequence?

2. Given a HMM and an observed output sequence, what is the most likely state sequence?

3. Given observed output and state sequences, what is the most likely HMM?

All three questions are addressed in this section. In Section 9.2.1 we answer the first question. The second question is answered in Section 9.2.2. The third question is addressed through 'realization theory' and turns out to be by far the most difficult and deep question. It is addressed in several stages. First, a standard method known as the Baum-Welch method is given in Section 9.2.3. In the Baum-Welch method, it is assumed that the *size n of the state space of $\{\mathcal{X}_t\}$ is known beforehand*, and the emphasis is only on choosing the best possible *parameter set* for the HMM. Clearly there is some arbitrariness in this problem formulation. Ideally, the size n of the state space of $\{\mathcal{X}_t\}$ should be determined by the data that we wish to model, and not be fixed beforehand for our convenience. Chapters 9 and 9 address this latter problem formulation. In Chapter 9 the so-called 'partial realization problem' is studied, while the so-called 'complete realization problem' is addressed in Chapter 9.

9.2.1 Computation of Likelihoods of Output Sequences

As is common in computer science, let us use the symbol $\mathbb{M}^*$ to denote the set of all *finite* strings over the set $\mathbb{M}$. Thus $\mathbb{M}^*$ consists of all strings of the form $\mathbf{u} = u_1 \dots u_l$, where each $u_i \in \mathbb{M}$ and l is the length of the string. We let $|\mathbf{u}|$ denote the length of the string. Now it is well known that the set $\mathbb{M}^*$ is countably infinite.[4] The objective of this section is to derive a simple formula for the likelihood of observing a string $\mathbf{u} \in \mathbb{M}^*$. In the literature, this simple formula is referred to as the "forward-backward recursion."

To facilitate the various problem statements and their solutions, a bit of notation is introduced. Suppose we are given a stochastic process $\{\mathcal{Y}_t\}$. In principle the process has a time index ranging from $-\infty$ to ∞. In the interests of brevity, we define

$$\mathcal{Y}_k^l := (\mathcal{Y}_k, \mathcal{Y}_{k+1}, \dots, \mathcal{Y}_{l-1}, \mathcal{Y}_l). \tag{9.9}$$

In the above notation, it is supposed that $k \leq l$. If $k = l$, then $\mathcal{Y}_k^l$ becomes just $\mathcal{Y}_k = \mathcal{Y}_l$. The symbol is undefined if $k > l$. Similar notation is used without comment for other stochastic processes (such as $\{\mathcal{X}_t\}$).

Suppose that a process $\{\mathcal{Y}_t\}$ has a Type 3 HMM. Thus there is an associated Markov process $\{\mathcal{X}_t\}$ such that the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov, and in addition (9.4) holds. As in (9.6), let us define the $n \times n$ matrices $M^{(u)} \in [0,1]^{n \times n}$ by

$$m_{ij}^{(u)} := \Pr\{\mathcal{X}_t = i \& \mathcal{Y}_t = u | \mathcal{X}_{t-1} = i\}, \ \forall i, j \in \mathbb{N}, u \in \mathbb{M}.$$

and the stochastic matrix $A \in [0,1]^{n \times n}$ by

$$a_{ij} := \sum_{u \in \mathbb{M}} m_{ij}^{(u)}, \ \forall i, j.$$

Then A is the state transition matrix of the Markov process $\{\mathcal{X}_t\}$. If the HMM is of Type 2 rather than Type 3, then as shown in the proof of Theorem 9.4, we have that

$$m_{ij}^{(u)} = a_{ij} b_{ju}.$$

To complete the specification of the Markov process $\{\mathcal{X}_t\}$, we need to specify the stationary distribution π, since in general the matrix A could have more than one stationary distribution. Suppose π is also specified. Then the dynamics of the two processes $\{\mathcal{X}_t\}$ and $\{\mathcal{Y}_t\}$ are completely specified.

[4] If we permit strings of *infinite* length, then the resulting set is uncountably infinite.

Now suppose $\mathbf{u} \in \mathbb{M}^*$ and that $|\mathbf{u}| = l$. We wish to compute the likelihood

$$\mathbf{f_u} := \Pr\{(\mathcal{Y}_t, \mathcal{Y}_{t+1}, \ldots, \mathcal{Y}_{t+l-1}) = \mathbf{u}\}. \tag{9.10}$$

Using the notation of (9.9), we can also write

$$\mathbf{f_u} := \Pr\{\mathcal{Y}_t^{t+l-1} = \mathbf{u}\}. \tag{9.11}$$

In other words, we wish to compute the probability of observing the sequence of outputs $\mathbf{u} = u_1 \ldots u_l$ in exactly that order. Note that, since the process $\{\mathcal{Y}_t\}$ is stationary, the number $f_{\mathbf{u}}$ as defined in (9.10) is independent of t, the time at which the observations start. So we might as well take $t = 1$. We refer to the quantity $f_{\mathbf{u}}$ as the **frequency** of the string $\mathbf{u}$.

To compute the frequency $f_{\mathbf{u}}$ corresponding to a particular $\mathbf{u} \in \mathbb{M}^*$, let us observe that if $\mathbf{i} := (i_0, i_1, \ldots, i_l) \in \mathbb{N}^{l+1}$ is a particular sequence of states, then

$$\begin{aligned}\Pr\{\mathcal{Y}_1^l = \mathbf{u} \& \mathcal{X}_0^l = \mathbf{i}\} &= \Pr\{\mathcal{X}_0 = i_0\} \\ &\quad \cdot \prod_{t=1}^{l} \Pr\{(\mathcal{Y}_t, \mathcal{X}_t) = (u_t, i_t) | \mathcal{X}_{t-1} = i_{t-1}\} \\ &= \pi_{i_0} \prod_{t=1}^{l} m_{i_{t-1} i_t}^{(u_t)}.\end{aligned}$$

Now to compute $f_{\mathbf{u}}$, we can observe that

$$\begin{aligned}f_{\mathbf{u}} &= \Pr\{\mathcal{Y}_1^l = \mathbf{u}\} \\ &= \sum_{\mathbf{i} \in \mathbb{N}^{l+1}} \Pr\{\mathcal{Y}_1^l = \mathbf{u} \& \mathcal{X}_0^l = \mathbf{i}\},\end{aligned}$$

by summing over all possible state sequences $\mathbf{i}$. So in principle we could compute the above probability for each sequence $\mathbf{i} \in \mathbb{N}^{l+1}$ and then add them up. This gives the expression

$$f_{\mathbf{u}} = \sum_{(i_0, \ldots, i_l) \in \mathbb{N}^{l+1}} \pi_{i_0} \prod_{t=1}^{l} m_{i_{t-1} i_t}^{(u_t)}. \tag{9.12}$$

If the above equation is interpreted literally and we sum over all possible state sequences, this computation requires $O(n^{l+1})$ computations. Clearly this is a very silly way to compute $f_{\mathbf{u}}$.

Instead let us note that (9.12) can be interpreted as a matrix product. In fact it equals

$$f_{\mathbf{u}} = \pi M^{(u_1)} \cdots M^{(u_l)} \mathbf{e}_n. \tag{9.13}$$

The equivalence of the two formulae (9.12) and (9.13) can be seen easily by expanding the right side of (9.12) as

$$f_{\mathbf{u}} = \sum_{i_0} \sum_{i_1} \cdots \sum_{i_l} \pi_{i_0} m_{i_0 i_1}^{(u_1)} \ldots m_{i_{l-1} i_l}^{(u_l)}.$$

Now if we simply start from the leftmost term in (9.13) and multiply π by $M^{(u_1)}$, and then multiply the resulting row vector by $M^{(u_2)}$, and so on, and finally multiply the resulting row vector by $\mathbf{e}_n$, then the complexity is $O(ln^2)$, since multiplying an $1 \times n$ row vector by an $n \times n$ matrix requires $O(n^2)$ operations, and we need to do this $l+1$ times.[5]

The very useful formula (9.13) can be given a very nice interpretation, which is referred to in some HMM literature as the 'forward-backward recursion.' Observe that, given a sequence $\mathbf{u} \in \mathbb{M}^*$ of length l, the formula (9.13) for $f_{\mathbf{u}}$ can be written as

$$f_{\mathbf{u}} = \alpha(\mathbf{u}, k)\beta(\mathbf{u}, k),$$

where

$$\alpha(\mathbf{u}, k) = \pi M^{(u_1)} \cdots M^{(u_k)} = \pi \prod_{t=1}^{k} M^{(u_t)}, \tag{9.14}$$

$$\beta(\mathbf{u}, k) = M^{(u_{k+1})} \cdots M^{(u_l)} \mathbf{e}_n = \prod_{t=k+1}^{l} M^{(u_t)} \mathbf{e}_n. \tag{9.15}$$

These formulas are valid for *every* k between 1 and l, provided we take the empty product as the identity matrix. These formulas can now be given a very simple interpretation in terms of conditional probabilities, which is the basis of the so-called 'forward-backward recursion.' Observe that $\alpha(\mathbf{u}, k)$ is a $1 \times n$ row vector, whereas $\beta(\mathbf{u}, k)$ is an $n \times 1$ column vector.

Lemma 9.7 *With $\alpha(\mathbf{u}, k)$ and $\beta(\mathbf{u}, k)$ defined as in (9.14) and (9.15) respectively, we have*

$$\alpha_i(\mathbf{u}, k) = \Pr\{(\mathcal{Y}_1, \ldots, \mathcal{Y}_k) = (u_1, \ldots u_k) \& \mathcal{X}_k = i\}. \tag{9.16}$$

$$\beta_i(\mathbf{u}, k) = \Pr\{\mathcal{X}_k = i \& (\mathcal{Y}_{k+1}, \ldots, \mathcal{Y}_l) = (u_{k+1}, \ldots, u_l)\}. \tag{9.17}$$

The proof is obvious from the formulas (9.14) and (9.15) respectively, and is left as an exercise to the reader.

9.2.2 Computation of Likelihoods of State Sequences: The Viterbi Algorithm

In this subsection, we study the following question: Suppose we are given a HMM, and an observation $\mathbf{u} \in \mathcal{M}^l$. Thus we know the parameters of an HMM, and are given an observation $\mathcal{Y}_1 = u_1, \ldots, \mathcal{Y}_l = u_l$. The question is: What is the "most likely" state sequence $\mathcal{X}_0, \ldots, \mathcal{X}_l$, where "most likely" is interpreted in the sense of the maximum *a posteriori* estimate as defined in Section 2.2.4. Thus the problem is: Find

$$\text{Arg} \max_{\mathbf{i} \in \mathbb{N}^{l+1}} \Pr\{\mathcal{X}_0^l = \mathbf{i} | \mathcal{Y}_1^l = \mathbf{u}\}.$$

[5] Since $l+1$ and l are of the same order of magnitude, we can write $O(ln^2)$ instead of $O((l+1)n^2)$.

In words, the problem is to find $\mathbf{i} \in \mathbb{N}^{l+1}$ that maximizes the *a posteriori* probability $\Pr\{\mathcal{X}_0^l = \mathbf{i} | \mathcal{Y}_1^l = \mathbf{u}\}$. The first, but very crucial, step is to observe that

$$\Pr\{\mathcal{X}_0^l = \mathbf{i} | \mathcal{Y}_1^l = \mathbf{u}\} = \frac{\Pr\{\mathcal{X}_0^l = \mathbf{i} \& \mathcal{Y}_1^l = \mathbf{u}\}}{\Pr\{\mathcal{Y}_1^l = \mathbf{u}\}}.$$

Now the variable of optimization here is $\mathbf{i} \in \mathbb{N}^{l+1}$, which *does not appear* in the denominator. So we can treat the denominator as a constant and simply maximize the numerator with respect to $\mathbf{i}$. In other words, the problem is to find

$$\operatorname{Arg} \max_{\mathbf{i} \in \mathbb{N}^{l+1}} \Pr\{\mathcal{X}_0^l = \mathbf{i} \& \mathcal{Y}_1^l = \mathbf{u}\}.$$

An explicit expression for the right side can be deduced from (9.12), namely

$$\Pr\{\mathcal{X}_0^l = \mathbf{i} \& \mathcal{Y}_1^l = \mathbf{u}\} = \pi_{i_0} \prod_{t=1}^{l} m_{i_{t-1} i_t}^{(u_t)}.$$

Maximizing the right side with respect to $i_0, \ldots, i_l$ is a very difficult problem if we try solve it directly. The well-known **Viterbi algorithm** is a systematic approach that allows us to break down the single (and seemingly intractable) optimization problem into a *sequence* of optimization problems, each of which is tractable.

Towards this end, let us fix a time $t \leq l$, a state $j \in \mathbb{N}$, and define

$$\begin{aligned} \gamma(t, j; \mathbf{i}) &:= \Pr\{\mathcal{X}_0^{t-1} = \mathbf{i} \& \mathcal{X}_t = j \& \mathcal{Y}_1^t = \mathbf{u}_1^t\} \\ &= \pi_{i_0} \cdot \left[\prod_{s=1}^{t-1} m_{i_{s-1} i_s}^{(u_s)} \right] \cdot m_{i_{t-1} j}^{(u_t)}, \end{aligned} \tag{9.18}$$

$$\gamma^*(t, j) := \max_{\mathbf{i} \in \mathbb{N}^t} \gamma(t, j; \mathbf{i}), \tag{9.19}$$

$$I^*(t, j) := \{\mathbf{i} \in \mathbb{N}^t : \gamma(t, j; \mathbf{i}) = \gamma^*(t, j)\}. \tag{9.20}$$

Thus $I^*(t, j)$ consists of the most likely state sequences $\mathbf{i} \in \mathbb{N}^t$ that lead to state j at time t and match the observation history up to time t. The key result is stated next. It is an instance of the "Principle of Optimality." This principle states that in a sequential optimization problem, any subsequence of an optimal solution is also optimal (under suitable conditions of course).

Theorem 9.8 *Fix t and j, and define $I^*(t, j)$ as in 9.20. Suppose $t \geq 2$ and that $i_0 \ldots i_{t-1} \in I^*(t, j)$. Then*

$$i_0 \ldots i_{t-2} \in I^*(t-1, i_{t-1}). \tag{9.21}$$

More generally, for any $s \leq t - 1$, we have

$$i_0 \ldots i_{s-1} \in I^*(s, i_s). \tag{9.22}$$

Proof. Suppose by way of contradiction that

$$i_0 \dots i_{t-2} \notin I^*(t-1, i_{t-1}).$$

This implies that there exists another sequence $j_0 \dots j_{t-2} \in \mathbb{N}^{t-1}$ such that

$$\gamma(t-1, i_{t-1}; i_0 \dots i_{t-2}) < \gamma(t-1, i_{t-1}; j_0 \dots j_{t-2}).$$

Expanding this inequality leads to

$$\pi_{i_0} \left[\prod_{s=1}^{t-2} m_{i_{s-1} i_s}^{(u_s)} \right] \cdot m_{i_{t-2} i_{t-1}}^{(u_{t-1})} < \pi_{j_0} \left[\prod_{s=1}^{t-2} m_{j_{s-1} j_s}^{(u_s)} \right] \cdot m_{j_{t-2} i_{t-1}}^{(u_{t-1})}.$$

Multiplying both sides by $m_{i_{t-1} j}^{(u_t)}$ shows that

$$\gamma(t, j; \mathbf{i}) < \gamma(t, j; j_0 \dots j_{t-2} i_{t-1}),$$

which contradicts the assumption that $i_0 \dots i_{t-2} i_{t-1} \in I^*(t, j)$. Hence (9.21) is true. The proof of (9.22) is entirely similar. □

Theorem 9.9 *The function $\gamma^*(\cdot, \cdot)$ satisfies the recursive relationship*

$$\gamma^*(t, j) = \max_{i \in \mathbb{N}} \left[\gamma^*(t-1, i) \cdot m_{ij}^{(u_t)} \right], t \le l. \tag{9.23}$$

Proof. Fix $t \le l$ and $j \in \mathbb{N}$. Suppose $i_0 \dots i_{t-1} \in I^*(t, j)$ is an optimal state sequence. Then it follows from Theorem 9.8 that

$$\begin{aligned} \gamma * (t, j) &= \gamma(t, j, i_0 \dots i_{t-1}) \\ &= \gamma(t, i_{t-1}; i_0 \dots i_{t-2}) \cdot m_{i_{t-1} j}^{(u_t)} \text{ from (9.18)} \\ &= \gamma^*(t-1, i_{t-1}) \cdot m_{i_{t-1} j}^{(u_t)} \end{aligned} \tag{9.24}$$

At this point the only variable of optimization left is i_{t-1}, which we can simply relabel as i. Choosing $i = i_{t-1}$ so as to maximize the right side of (9.24) leads to the recursive relationship (9.23). □

Now we have everything in place to state the Viterbi algorithm.

Step 1. (Initialization) For each $i_1 \in \mathbb{N}$, choose $i_0 \in \mathbb{N}$ so as to maximize the product $\pi_{i_0} m_{i_0 i_1}^{(u_1)}$. In case there is more than one optimal i_0 corresponding to a given i_1, choose any one of the optimal i_0. Thus leads to n optimal trajectories $i_0 i_1$, and n corresponding optimal values $\gamma(i, i_1) = \pi_{i_0} m_{i_0 i_1}^{(u_1)}$, one for each $i_1 \in \mathbb{N}$.

Step 2. (Recursion) Suppose $2 \le t \le l$. At time t, for each $\mathbf{i}_{t-1} \in \mathbb{N}$ we have an optimal value $\gamma^*(t-1, i_{t-1})$ and an associated optimal trajectory $i_0 \dots i_{t-2}$. Now, for each $i_t \in \mathbb{N}$, choose $i_{t-1} \in \mathbb{N}$ as the value of i that achieves the maximum in (9.23), with $j = i_t$. If there is more than one value of i_{t-1} that achieves the maximum, choose any one value. Once i_{t-1} is determined, concatenate i_{t-1} to the associated optimal trajectory to i_{t-1}, and call it the optimal trajectory corresponding to i_t. The optimal value $\gamma^*(t, i_t)$ is the maximum value in (9.23).

Step 3. (Completion) Let $t = l$. At this stage we have n optimal values $\gamma^*(l, 1), \dots, \gamma^*(l, n)$, and n associated optimal trajectories. Now choose as

i_l the value of j that maximizes $\gamma^*(l,j)$ and choose the associated optimal trajectory as the most likely state sequence.

The algorithm is illustrated through a "baby" example below, so as to illustate the various steps involved. But now let us analyze the computational complexity of the algorithm. At each time t and for each state $j \in \mathbb{N}$, we need to compute the product $\gamma(t-1,i)m_{ij}^{(u_t)}$ for each $i \in \mathbb{N}$, and then find the largest value of the product. This has complexity $O(n)$ for each $j \in \mathbb{N}$, or $O(n^2)$ in all. Since we need to do this l times in all, the overall complexity is $O(ln^2)$, which is the same as the complexity of computing the frequency $f_{\mathbf{u}}$ for a given $\mathbf{u} \in \mathcal{M}^l$. Note that a frontal attack on the problem by enumerating all possible state trajectories would have complexity $O(n^l)$, which would be unacceptable.

9.2.3 Learning Hidden Markov Models: The Baum-Welch Algorithm

In this subsection we study the problem of "learning" a hidden Markov model (HMM) on the basis of observations. In the literature, two distinct problems are usually mixed up, and we try to disentangle them here.

Suppose we are specified a state space $\mathbb{N} = \{1,\ldots,n\}$, an output space $\mathcal{M} = \{1,\ldots,m\}$ and a *family* of Type 3 HMM's, parametrized by $\lambda \in \Lambda \subseteq \mathbb{R}^d$. Here Λ is the *set* of parameters, d is the *number* of parameters, and λ is a generic symbol denoting a particular parameter (vector if $d > 1$). The family of HMM's is denoted by $\{(\pi(\lambda); M^{(u)}(\lambda), u \in \mathcal{M}), \lambda \in \Lambda\}$. As in Section 9.1, the various quantities are interpreted as follows: $M^{(u)}(\lambda) \in [0,1]^{n\times n}$ for each $u \in \mathcal{M}$, and

$$m_{ij}^{(u)}(\lambda) = \Pr\{\mathcal{X}_t = j \& \mathcal{Y}_t = u | \mathcal{X}_{t-1} = i\},$$

when the parameter is λ. The matrix

$$A(\lambda) := \sum_{u\in\mathcal{M}} M^{(u)}(\lambda) \in [0,1]^{n\times n}$$

is a stochastic matrix, and is the state transition matrix of the Markov chain associated with the HMM. Finally, $\pi(\lambda)A(\lambda) = \pi(\lambda)$, so that $\pi(\lambda)$ is the stationary distribution of the Markov chain. So the only new feature here is that the vector $\pi(\lambda)$ and the matrices $M^{(u)}(\lambda)$ depend on the auxiliary parameter λ.

Now suppose we are given a specific observation $\mathbf{u} \in \mathcal{M}^l$. Thus we have observed that $\mathcal{Y}_1 = u_1, \ldots, \mathcal{Y}_l = u_l$. The problem is to choose, amongst all the given HMM's, a model that maximizes the likelihood of the observed sequence. Now we know from (9.12) that

$$f_{\mathbf{u}}(\lambda) = \sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \pi_{i_0} \prod_{t=1}^{l} m_{i_{t-1}i_t}^{(u_t)}(\lambda). \tag{9.25}$$

So the problem is to *maximize* $f_{\mathbf{u}}(\lambda)$ with respect to $\lambda \in \Lambda$. From Definition 2.23, we see that we are seeking a maximum likelihood estimate of $\lambda \in$

Λ. Once we find a $\lambda^* \in \Lambda$ that maximizes $f_{\mathbf{u}}(\lambda)$, we can say that the corresponding model $(\pi(\lambda^*); M^{(u)}(\lambda^*), u \in \mathcal{M})\}$ is the most likely model.

There is a further refinement possible. Up to now we have assumed that the length l of the observation is *fixed*. Now suppose that the data is generated by a "true but unknown" model $(\pi(\lambda_0); M^{(u)}(\lambda_0), u \in \mathcal{M})\}$, where λ_0 is the "true but unknown" value of the parameter. We take longer and longer observations, and at each length l, we form a corresponding estimate λ_l^*. So the question is: Does $\lambda_l^* \to \lambda_0$ as $l \to \infty$? In other words, does the estimated model converge to the true but unknown model as the length of the observations approaches infinity?

Let us begin with the first problem, namely, finding the most likely model, or in other words, maximizing the right side of (9.25) with respect to λ. It is clear that trying to differentiate this expression with respect to λ is very complicated. Instead, we use an observation first presented in [14] to make the problem tractable. A little bit of notation is presented first.

Let μ denote the uniform probability distribution on hte set $\mathbb{N}^{l+1}$. Thus μ assigns a weight of $n^{-(l+1)}$ to each element $\mathbf{i} \in \mathbb{N}^{l+1}$. With this notation, we can think of the expression in (9.25) as an expected value. Let us define the map $\phi_{\mathbf{u}}(\lambda) : \mathbb{N}^{l+1} \to [0, 1]$ by

$$\begin{aligned} \phi_{\mathbf{u}}(\lambda, \mathbf{i}) &:= \Pr\{\mathcal{X}_0^l = \mathbf{i} \& \mathcal{Y}_1^l = \mathbf{u}\} \\ &= \pi_{i_0} \prod_{t=1}^{l} m_{i_{t-1} i_t}^{(u_t)}(\lambda). \end{aligned}$$

Then it is clear from (9.25) that

$$f_{\mathbf{u}}(\lambda) = n^{l+1} E[\phi_{\mathbf{u}}(\lambda, \mathbf{i}), \mu].$$

Since n^{l+1} is just a constant, the problem of maximizing $f_{\mathbf{u}}(\lambda)$ with respect to λ is equivalent to the problem of maximizing the expected value $E[\phi_{\mathbf{u}}(\lambda, \mathbf{i}), \mu]$ with respect to λ. Actually, since $\mathbf{u}$ is a given *fixed* observation, there is no need to display explicitly the dependence of various quantities on $\mathbf{u}$. Accordingly, we suppress this dependence in all intermediate calculations, and display it only later on, as needed.

Now we come to the theorem first presented in [14]. Note that the theorem applies to *any* function ϕ, not just a function arising from a HMM.

Theorem 9.10 *Suppose $\phi : \Lambda \to \mathbb{N}^{l+1} \times [0, \infty)$ has the property, for a given $\lambda \in \Lambda$, that*

$$\phi(\lambda, \mathbf{i}) = 0 \ \Rightarrow \ \phi(\nu, \mathbf{i}) = 0 \ \forall \nu \in \Lambda.$$

Let μ denote the uniform probability distribution on $\mathbb{N}^{l+1}$, and define

$$g(\lambda) := E[\phi(\lambda, \mathbf{i}), \mu], g(\nu) := E[\phi(\nu, \mathbf{i}), \mu],$$

$$h(\lambda, \nu) := E[\phi(\lambda, \mathbf{i}) \log \phi(\nu, \mathbf{i}), \mu].$$

Then

$$h(\lambda, \nu) - h(\lambda, \lambda) \le g(\lambda) \log \left[\frac{g(\nu)}{g(\lambda)} \right]. \tag{9.26}$$

So in particular it follows that

$$[h(\lambda,\nu) \geq h(\lambda,\lambda)] \Rightarrow [g(\nu) \geq g(\lambda)]. \tag{9.27}$$

Proof. The assumption on $\phi(\lambda,\cdot)$ ensures that the quantity $\phi(\lambda,\mathbf{i})\log\phi(\nu,\mathbf{i})$ is well-defined for all $\mathbf{i}$ (and equals zero if $\phi(\lambda,\mathbf{i}) = 0$). Now we reason as follows: Since μ is the probability distribution on $\mathbb{N}^{l+1}$, it follows that

$$g(\lambda) = n^{-(l+1)} \sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \phi(\lambda,\mathbf{i}),$$

$$g(\nu) = n^{-(l+1)} \sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \phi(\nu,\mathbf{i}),$$

$$\begin{aligned}\log\left[\frac{g(\nu)}{g(\lambda)}\right] &= \log\left[\frac{1}{n^{l+1}g(\lambda)} \sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \phi(\nu,\mathbf{i})\right] \\ &= \log\left[\sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \frac{\phi(\nu,\mathbf{i})}{\phi(\lambda,\mathbf{i})} \cdot \frac{\phi(\lambda,\mathbf{i})}{n^{l+1}g(\lambda)}\right]\end{aligned}$$

Now observe that

$$\sum_{\mathbf{i}\in\mathbb{N}^{l+1}} \frac{\phi(\lambda,\mathbf{i})}{n^{l+1}g(\lambda)} = 1$$

by the definition of $g(\lambda)$. So the vector

$$\mathbf{q}_\lambda := \left[\frac{\phi(\lambda,\mathbf{i})}{n^{l+1}g(\lambda)}, \mathbf{i} \in \mathbb{N}^{l+1}\right] \in \mathbb{S}_{n^{l+1}}$$

is a probability distribution on $\mathbb{N}^{l+1}$. So we can write

$$\log\left[\frac{g(\nu)}{g(\lambda)}\right] = \log E\left[\frac{\phi(\nu,\mathbf{i})}{\phi(\lambda,\mathbf{i})}, \mathbf{q}_\lambda\right].$$

Now, since log is a concave function, we can apply Jensen's inequality to conclude that

$$\begin{aligned}\log\left[\frac{g(\nu)}{g(\lambda)}\right] &\geq E\left[\log\frac{\phi(\nu,\mathbf{i})}{\phi(\lambda,\mathbf{i})}, \mathbf{q}_\lambda\right] \\ &= E[\log\phi(\nu,\mathbf{i}), \mathbf{q}_\lambda] - E[\log\phi(\lambda,\mathbf{i}), \mathbf{q}_\lambda] \\ &= \frac{1}{g(\lambda)}[h(\lambda,\nu) - h(\lambda,\lambda)]\end{aligned}$$

This establishes (9.26), from which (9.27) follows readily. □

Chapter Ten

Hidden Markov Processes: Complete Realization Problem

In this chapter we continue our study of hidden Markov processes begun in the previous chapter. The focus of study here is the so-called complete realization problem, which can be stated as follows: Suppose $\mathbb{M} = \{1, \ldots, m\}$ is a finite set[1] and that $\{\mathcal{Y}_t\}_{t \geq 0}$ is a stationary stochastic process assuming values in $\mathbb{M}$. We wish to derive necessary and/or sufficient conditions for $\{\mathcal{Y}_t\}$ to be a hidden Markov process. As shown in Theorem 9.4, the existence of each of the three types of HMMs is equivalent: The process $\{\mathcal{Y}_t\}$ has any one kind of a HMM realization if and only if it has all three. Thus we shall use whichever HMM makes it easy to prove what we wish to prove.

In summary, it is quite easy to prove a universal *necessary* condition for the given process to have a HMM. But this condition is *not sufficient in general.* Unfortunately the demonstration of this fact is rather long. One can in principle present a 'necessary and sufficient condition', but as pointed out by Anderson [6], the 'necessary and sufficient condition' is virtually a restatement of the problem to be solved and does not shed any insight on the problem. However, if one adds the requirement that the process $\{\mathcal{Y}_t\}$ is also 'mixing', then it is possible to present conditions that are 'almost necessary and sufficient' for the process to have a HMM realization.

10.1 A UNIVERSAL NECESSARY CONDITION

10.1.1 The Hankel Matrix

In this subsection we introduce a very useful matrix which we refer to as a 'Hankel' matrix, because it has some superficial similarity to a Hankel matrix. Given the set $\mathbb{M}$ in which the stochastic process $\{\mathcal{Y}_t\}$ assumes its values, let us define some lexical ordering of the elements in $\mathbb{M}$. The specific order itself does not matter, and the reader can verify that all of the discussion in the present chapter is insensitive to the specific lexical ordering used. For each integer l, the set $\mathbb{M}^l$ has cardinality m^l and consists of l-tuples. These can be arranged either in first-lexical order (flo) or last-lexical order (llo). First-lexical order refers to indexing the first element,

[1] As always, we really mean that $\mathbb{M} = \{m_1, \ldots, m_n\}$, a collection of abstract symbols, and we write $\mathbb{M} = \{1, \ldots, m\}$ only to simplify notation.

then the second, and so on, while last-lexical order refers to indexing the last element, then the next to last, and so on. For example, suppose $m = 2$ and that $\mathbb{M} = \{1, 2\}$ in the natural order. Then

$$\begin{aligned} \mathbb{M}^3 \text{ in llo } &= \{111, 112, 121, 122, 211, 212, 221, 222\}, \\ \\ \mathbb{M}^3 \text{ in flo } &= \{111, 211, 121, 221, 112, 212, 122, 222\}. \end{aligned}$$

Given any finite string $\mathbf{u} \in \mathbb{M}^l$, its frequency $f_{\mathbf{u}}$ is defined by

$$f_{\mathbf{u}} := \Pr\{(\mathcal{Y}_{t+1}, \mathcal{Y}_{t+2}, \ldots, \mathcal{Y}_{t+1}) = (u_1, u_2, \ldots, u_l)\}.$$

Since the process $\{\mathcal{Y}_t\}$ is assumed to be stationary, the above probability is independent of t. Moreover, as seen earlier, the frequency vector is *consistent*; that is

$$f_{\mathbf{u}} = \sum_{v \in \mathbb{M}} f_{\mathbf{u}v} = \sum_{w \in \mathbb{M}} f_{w\mathbf{u}}, \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.1}$$

More generally,

$$f_{\mathbf{u}} = \sum_{\mathbf{v} \in \mathbb{M}^r} f_{\mathbf{uv}} = \sum_{\mathbf{w} \in \mathbb{M}^s} f_{\mathbf{wu}}, \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.2}$$

Given integers $k, l \geq 1$, the matrix $F_{k,l}$ is defined as

$$F_{k,l} = [f_{\mathbf{uv}}, \mathbf{u} \in \mathbb{M}^k \text{ in flo}, \mathbf{v} \in \mathbb{M}^l \text{ in llo}] \in [0,1]^{m^k \times m^l}.$$

Thus the rows of $F_{k,l}$ are indexed by an element of $\mathbb{M}^k$ in flo, while the columns are indexed by an element of $\mathbb{M}^l$ in llo. For example, suppose $m = 2$, and $\mathbb{M} = \{1, 2\}$. Then

$$F_{1,2} = \begin{bmatrix} f_{111} & f_{112} & f_{121} & f_{122} \\ f_{211} & f_{212} & f_{221} & f_{222} \end{bmatrix},$$

whereas

$$F_{2,1} = \begin{bmatrix} f_{111} & f_{112} \\ f_{211} & f_{212} \\ f_{121} & f_{122} \\ f_{221} & f_{222} \end{bmatrix}.$$

In general, for a given integer s, the matrices $F_{0,s}, F_{1,s-1}, \ldots, F_{s-1,1}, F_{s,0}$ all contain frequencies of the m^s s-tuples in the set $\mathbb{M}^s$. However, the dimensions of the matrices are different, and the elements are arranged in a different order. Note that by convention $F_{0,0}$ is taken as the 1×1 matrix 1 (which can be thought of as the frequency of occurence of the empty string).

Given integers $k, l \geq 1$, we define the matrix $H_{k,l}$ as

$$H_{k,l} := \begin{bmatrix} F_{0,0} & F_{0,1} & \ldots & F_{0,l} \\ F_{1,0} & F_{1,1} & \ldots & F_{1,l} \\ \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \ldots & F_{k,l} \end{bmatrix}.$$

Note that $H_{k,l}$ has $1 + m + \ldots + m^k$ rows, and $1 + m + \ldots + m^l$ columns. In general, $H_{k,l}$ is not a 'true' Hankel matrix, since it is not constant along backward diagonals. It is not even 'block Hankel.' However, it resembles a Hankel matrix in the sense that the matrix in the (i, j)-th block consists of frequencies of strings of length $i + j$. Finally, we define H (without any subscripts) to be the infinite matrix of the above form, that is,

$$H := \begin{bmatrix} F_{0,0} & F_{0,1} & \ldots & F_{0,l} & \ldots \\ F_{1,0} & F_{1,1} & \ldots & F_{1,l} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \ldots & F_{k,l} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Through a mild abuse of language we refer to H as the Hankel matrix associated with the process $\{\mathcal{Y}_t\}$.

10.1.2 A Necessary Condition for the Existence of Hidden Markov Models

In this section, it is shown that a process $\{\mathcal{Y}_t\}$ has a HMM *only if* the matrix H has finite rank. Taking some liberties with the English language, we refer to this as the 'finite Hankel rank condition'. Theorem 10.1 below shows that the finite Hankel rank condition is a *universal necessary condition* for a given process to have a HMM realization. However, as shown in the next subsection, the finiteness of the rank of H is only necessary, but not sufficient in general.

Theorem 10.1 *Suppose $\{\mathcal{Y}_t\}$ has a Type 3 HMM with the associated $\{\mathcal{X}_t\}$ process having n states. Then Rank$(H) \leq n$.*

Proof: The definition of a Type 3 HMM implies that the process $\{\mathcal{X}_t\}$ is Markov over a set $\mathbb{N}$ of cardinality n. Define the matrices $M^{(1)}, \ldots, M^{(m)}, A$ as in (9.7) and (9.8) respectively, and let $\boldsymbol{\pi}$ denote the stationary distribution associated with this Markov process. Now suppose $\mathbf{u} \in \mathbb{M}^l$, specificlly that $\mathbf{u} = u_1 \ldots u_l$. Then from the sum of products formula (9.13) derived earlier, we have

$$f_{\mathbf{u}} = \sum_{i=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{j_l=1}^{n} \pi_i m_{ij_1}^{(u_1)} \cdots m_{j_{l-1}j_l}^{(u_l)} = \pi M^{(u_1)} \cdots M^{(u_l)} \mathbf{e}_n. \tag{10.3}$$

Note that

$$\sum_{l \in \mathbb{M}} M^{(l)} = A, \ \pi \left[\sum_{l \in \mathbb{M}} M^{(l)} \right] = \pi, \text{ and } \left[\sum_{l \in \mathbb{M}} M^{(l)} \right] \mathbf{e}_n = \mathbf{e}_n. \tag{10.4}$$

Thus the sum of the matrices $M^{(u)}$ is the state transition matrix of the Markov chain, and π and $\mathbf{e}_n$ are respectively a row eigenvector and a column eigenvector of A corresponding to the eigenvalue 1.

Now let us return to the matrix H. Using (10.3), we see at once that H can be factored as a product KL, where

$$K = \begin{bmatrix} \pi \\ \pi M^{(1)} \\ \vdots \\ \pi M^{(m)} \\ \pi M^{(1)} M^{(1)} \\ \vdots \\ \pi M^{(m)} M^{(m)} \\ \vdots \end{bmatrix},$$

$$L = [\mathbf{e}_n \mid M^{(1)}\mathbf{e}_n| \ldots |M^{(m)}\mathbf{e}_n \mid M^{(1)}M^{(1)}\mathbf{e}_n| \ \ldots \mid M^{(m)}M^{(m)}\mathbf{e}_n| \ldots].$$

In other words, the rows of K consist of $\pi M^{(u_1)} \cdots M^{(u_l)}$ as $\mathbf{u} \in \mathbb{M}^l$ is in flo and l increases, whereas the columns of L consist of $M^{(u_1)} \cdots M^{(u_l)}\mathbf{e}_n$ as $\mathbf{u} \in \mathbb{M}^l$ is in llo and l increases. Now note that the first factor has n columns whereas the second factor has n rows. Hence $\text{Rank}(H) \leq n$.

It has been shown by Sontag [102] that the problem of deciding whether or not a given 'Hankel' matrix has finite rank is undecidable.

10.2 NON-SUFFICIENCY OF THE FINITE HANKEL RANK CONDITION

Let us refer to the process $\{\mathcal{Y}_t\}$ as 'having finite Hankel rank' if $\text{Rank}(H)$ is finite. Thus Theorem 10.1 shows that $\text{Rank}(H)$ being finite is a *necessary* condition for the given process to have a HMM. However, the converse is *not true in general* – it is possible for a process to have finite Hankel rank and yet not have a realization as a HMM. The original example in this direction was given by Fox and Rubin [49]. However, their proof contains an error, in the opinion of this author. In a subsequent paper, Dharmadhikari and Nadkarni [41] quietly and without comment simplified the example of Fox and Rubin and also gave a correct proof (without explicitly pointing out that the Fox-Rubin proof is erroneous). In this section, we present the example of [41] and slightly simplify their proof. It is worth noting that the example crucially depends on rotating a vector by an angle α that is not commensurate with π, that is, α/π is not a rational number. A similar approach is used by Benvenuti and Farina [16], Example 4 to construct a nonnegative impulse response with finite Hankel rank which does not have a finite rank *nonnegative* realization.

Let us begin by choosing numbers $\lambda \in (0, 0.5], \alpha \in (0, 2\pi)$ such that α and π are noncommensurate. In particular, this rules out the possibility that $\alpha = \pi$. Now define

$$h_l := \lambda^l \sin^2(l\alpha/2), \ \forall l \geq 1.$$

Note that we can also write

$$h_l = \lambda^l \frac{(e^{\mathbf{i}l\alpha/2} - e^{-\mathbf{i}l\alpha/2})^2}{4},$$

where (just in this equation) $\mathbf{i}$ denotes $\sqrt{-1}$. Simplifying the expression for h_l shows that

$$h_l = \frac{\lambda^l}{4}(\zeta^l + \zeta^{-l} - 2), \tag{10.5}$$

where $\zeta := e^{\mathbf{i}\alpha}$. Because h_l decays at a geometric rate with respect to l, the following properties are self-evident.

1. $h_i > 0 \; \forall i$. Note that $l\alpha$ can never equal a multiple of π because α and π are noncommensurate.
2. We have that
$$\sum_{i=1}^{\infty} h_i =: \delta < 1. \tag{10.6}$$
3. We have that
$$\sum_{i=1}^{\infty} i h_i < \infty.$$
4. The infinite Hankel matrix
$$\bar{H} := \begin{bmatrix} h_1 & h_2 & h_3 & \dots \\ h_2 & h_3 & h_4 & \dots \\ h_3 & h_4 & h_5 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$
has finite rank of 3.

This last property follows from standard linear system theory. Given a sequence $\{h_i\}_{i \geq 1}$, let us define its z-transform $\tilde{h}(\cdot)$ by[2]

$$\tilde{h}(z) := \sum_{i=1}^{\infty} h_i z^{i-1}.$$

Thanks to an old theorem of Kronecker [75], it is known that the Hankel matrix $\bar{H}$ has finite rank if and only if $\tilde{h}$ is a *rational* function of z, in which case the rank of the Hankel matrix is the same as the degree of the rational function $\tilde{h}(z)$. Now it is a ready consequence of (10.5) that

$$\tilde{h}(z) = \frac{1}{4}\left[\frac{\lambda\zeta}{1 - \lambda\zeta z} + \frac{\lambda\zeta^{-1}}{1 - \lambda\zeta^{-1} z} - 2\frac{\lambda}{1 - \lambda z}\right].$$

[2] Normally in z-transformation theory, the sequence $\{h_i\}$ is indexed starting from $i = 0$, whereas here we have chosen to begin with $i = 1$. This causes the somewhat unconventional-looking definition.

Hence the infinite matrix $\bar{H}$ has rank 3.

The counterexample is constructed by defining a Markov process $\{\mathcal{X}_t\}$ with a countable state space and another process $\{\mathcal{Y}_t\}$ with just two output values such that $\mathcal{Y}_t$ is a function of $\mathcal{X}_t$. The process $\{\mathcal{Y}_t\}$ satisfies the finite Hankel rank condition; in fact $\text{Rank}(H) \leq 5$. And yet no Markov process with a finite state space can be found such that $\mathcal{Y}_t$ is a function of that Markov process. Since we already know from Section 9.4 that the existence of all the three kinds of HMMs is equivalent, this is enough to show that the process $\{\mathcal{Y}_t\}$ does not have a joint Markov process type of HMM.

The process $\{\mathcal{X}_t\}$ is Markovian with a countable state space $\{0, 1, 2, \ldots\}$. The transition probabilities of the Markov chain are defined as follows:

$$\Pr\{\mathcal{X}_{t+1} = 0 | \mathcal{X}_t = 0\} = 1 - \delta = 1 - \sum_{i=1}^{\infty} h_i,$$

$$\Pr\{\mathcal{X}_{t+1} = i | \mathcal{X}_t = 0\} = h_i \text{ for } i = 1, 2, \ldots,$$

$$\Pr\{\mathcal{X}_{t+1} = i | \mathcal{X}_t = i+1\} = 1 \text{ for } i = 1, 2, \ldots,$$

and all other probabilities are zero. Thus the dynamics of the Markov chain are as follows: If the chain starts in the initial state 0, then it makes a transition to state i with probability h_i, or remains in 0 with the probability $1 - \sum_i h_i = 1 - \delta$. Once the chain moves to the state i, it then successively goes through the states $i-1, i-2, \ldots, 1, 0$. Then the process begins again. Thus the dynamics of the Markov chain consist of a series of cycles beginning and ending at state 0, but where the lengths of the cycles are random, depending on the transition out of the state 0.

Clearly $\{\mathcal{X}_t\}$ is a Markov process. Now we define $\{\mathcal{Y}_t\}$ to be a function of this Markov process. Let $\mathcal{Y}_t = a$ if $\mathcal{X}_t = 0$, and let $\mathcal{Y}_t = b$ otherwise, i.e., if $\mathcal{X}_t = i$ for some $i \geq 1$. Thus the output process $\{\mathcal{Y}_t\}$ assumes just two values a and b. Note that in the interests of clarity we have chosen to denote the two output states as a and b instead of 1 and 2. For this process $\{\mathcal{Y}_t\}$ we shall show that (i)

$$\text{Rank}(H) \leq \text{Rank}(\bar{H}) + 2 = 5,$$

where H is the Hankel matrix associated with the process $\{\mathcal{Y}_t\}$, and (ii) there is no Markov process $\{\mathcal{Z}_t\}$ with a finite state space such that $\mathcal{Y}_t$ is a (deterministic) function of $\mathcal{Z}_t$.

The stationary distribution of the Markov chain is as follows:

$$\pi_0 = g := \left[1 + \sum_{i=1}^{\infty} i h_i \right]^{-1},$$

$$\pi_i = g \sum_{j=i}^{\infty} h_j, i \geq 1.$$

To verify this, note the structure of the state transition matrix A of the Markov chain: State 0 can be reached only from states 0 and 1. Thus column 0 of A has $1-\delta$ in row 0, 1 in row 1, and zeros in all other rows. For $i \geq 1$, state i can be reached only from states 0 and $i+1$. Hence column i has h_i in row 0, 1 in row $i+1$, and zeros elsewhere. As a result

$$(\pi A)_0 = g\left(1-\delta+\sum_{j=1}^{\infty} h_j\right) = g(1-\delta+\delta) = g = \pi_0,$$

while for $i \geq 1$,

$$(\pi A)_i = h_i \pi_0 + \pi_{i+1} = g\left[h_i + \sum_{j=i+1}^{\infty} h_j\right] = g\sum_{j=i}^{\infty} h_j = \pi_i.$$

To verify that this is indeed a probability vector, note that

$$\sum_{i=0}^{\infty} \pi_i = g\left[1+\sum_{i=1}^{\infty}\sum_{j=i}^{\infty} h_j\right] = g\left[1+\sum_{j=1}^{\infty}\sum_{i=1}^{j} h_j\right] = g\left[1+\sum_{j=1}^{\infty} j h_j\right] = 1$$

in view of the definition of g.

Next, let us compute the frequencies of various output strings. Note that if $\mathcal{Y}_t = a$, then certainly $\mathcal{X}_t = 0$. Hence, if $\mathcal{Y}_t = a$, then the conditional probability of $\mathcal{Y}_{t+1}$ does not depend on the values of $\mathcal{Y}_i, i < t$. Therefore, for arbitrary strings $\mathbf{u}, \mathbf{v} \in \{a, b\}^*$, we have

$$f_{\mathbf{u}a\mathbf{v}} = f_{\mathbf{u}a} \cdot f_{\mathbf{v}|\mathbf{u}a} = f_{\mathbf{u}a} \cdot f_{\mathbf{v}|a}.$$

Hence the infinite matrix $H^{(a)}$ defined by

$$H^{(a)} := [f_{\mathbf{u}a\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*]$$

has rank one. In such a case, it is customary to refer to a as a 'Markovian state.'

Next, let us compute the frequencies of strings of the form $ab^l a, ab^l, b^l a$, and b^l. A string of the form $ab^l a$ can occur only of $\mathcal{X}_t = 0, \mathcal{X}_{t+1} = l, \ldots, \mathcal{X}_{t+l} = 1, \mathcal{X}_{t+l+1} = 0$. All transitions except the first one have probability one, while the first transition has probability h_l. Finally, the probability that $\mathcal{X}_t = 0$ is π_0. Hence

$$f_{ab^l a} = \pi_0 h_l, \ \forall l.$$

Next, note that

$$f_{ab^l} = f_{ab^{l+1}} + f_{ab^l a}.$$

Hence, if we define

$$\pi_0 \gamma_l := f_{ab^l},$$

then γ_l satisfies the recursion

$$\pi_0 \gamma_l = \pi_0 \gamma_{l+1} + \pi_0 h_l.$$

To start the recursion, note that

$$\pi_0 \gamma_1 = f_{ab} = f_a - f_{aa} = \pi_0 - \pi_0(1-\delta) = \pi_0 \delta = \pi_0 \sum_{i=1}^{\infty} h_i.$$

Therefore

$$\pi_0 \gamma_l = \pi_0 \sum_{i=l}^{\infty} h_i, \text{ or } \gamma_l = \sum_{i=l}^{\infty} h_i.$$

Now we compute the frequencies f_{b^l} for all l. Note that

$$f_{b^l} = f_{b^{l+1}} + f_{ab^l} = f_{b^{l+1}} + \pi_0 \gamma_l.$$

Hence if we define $\pi_0 \eta_l := f_{b^l}$, then η_l satisfies the recursion

$$\eta_l = \eta_{l+1} + \gamma_l.$$

To start the recursion, note that

$$f_b = 1 - f_a = 1 - \pi_0.$$

Now observe that

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty} i h_i \right]^{-1}$$

and as a result

$$1 - \pi_0 = \pi_0 \sum_{i=1}^{\infty} i h_i = \pi_0 \sum_{i=1}^{\infty} \sum_{j=1}^{i} h_i = \pi_0 \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} h_i = \pi_0 \sum_{j=1}^{\infty} \gamma_j.$$

Hence

$$f_{b^l} = \pi_0 \eta_l, \text{ where } \eta_l = \sum_{i=l}^{\infty} \gamma_i.$$

Finally, to compute $f_{b^l a}$, note that

$$f_{b^l a} + f_{b^{l+1}} = f_{b^l}.$$

Hence

$$f_{b^l a} = f_{b^l} - f_{b^{l+1}} = \pi_0 (\eta_l - \eta_{l+1}) = \pi_0 \gamma_l.$$

Now let us look at the Hankel matrix H corresponding to the process $\{\mathcal{Y}_t\}$. We can think of H as the interleaving of two infinite matrices $H^{(a)}$ and $H^{(b)}$, where

$$H^{(a)} = [f_{\mathbf{u}a\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*],$$

$$H^{(b)} = [f_{\mathbf{u}b\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*].$$

We have already seen that $H^{(a)}$ has rank one, since a is a Markovian state. Hence it follows that

$$\text{Rank}(H) \leq \text{Rank}(H^{(a)}) + \text{Rank}(H^{(b)}) = \text{Rank}(H^{(b)}) + 1.$$

To bound $\text{Rank}(H^{(b)})$, fix integers l, n, and define

$$H^{(b)}_{l,n} := [f_{\mathbf{u}b\mathbf{v}}, \mathbf{u} \in \{a, b\}^l, \mathbf{v} \in \{a, b\}^n].$$

Note that $H^{(b)}_{l,n} \in [0,1]^{2^l \times 2^n}$. It is now shown that

$$\text{Rank}(H^{(b)}_{l,n}) \leq \text{Rank}(\bar{H}) + 1 = 4. \tag{10.7}$$

Since the right side is independent of l, n, it follows that

$$\text{Rank}(H^{(b)}) \leq 4,$$

whence

$$\text{Rank}(H) \leq 5.$$

To prove (10.7), suppose $\mathbf{u} \in \{a,b\}^{l-1}$ is arbitrary. Then

$$f_{\mathbf{u}ab\mathbf{v}} = f_{\mathbf{u}a} \cdot f_{b\mathbf{v}|\mathbf{u}a} = f_{\mathbf{u}a} \cdot f_{b\mathbf{v}|a},$$

because a is a Markovian state. Hence each of the 2^{l-1} rows $[f_{\mathbf{u}ab\mathbf{v}}, \mathbf{u} \in \{a,b\}^{l-1}]$ is a multiple of the row $[f_{b\mathbf{v}|a}]$, or equivalently, of the row $[f_{a^l b\mathbf{v}}]$. Hence Rank$(H^{(b)})$ is unaffected if we keep only this one row and jettison the remaining $2^{l-1} - 1$ rows. Similarly, as $\mathbf{u}$ varies over $\{a,b\}^{l-2}$, each of the rows $[f_{\mathbf{u}abb\mathbf{v}}]$ is proportional to $[f_{a^{l-2}abb\mathbf{v}}] = [f_{a^{l-1}b^2\mathbf{v}}]$. So we can again retain just the row $[f_{a^{l-1}b^2\mathbf{v}}]$ and discard the rest. Repeating this argument l times shows that $H^{(b)}$ has the same rank as the $(l+1) \times 2^n$ matrix

$$\begin{bmatrix} f_{a^l b\mathbf{v}} \\ f_{a^{l-1}b^2\mathbf{v}} \\ \vdots \\ f_{ab^l\mathbf{v}} \\ f_{b^{l+1}\mathbf{v}} \end{bmatrix}, \mathbf{v} \in \{a,b\}^n.$$

A similar exercise can now be repeated with $\mathbf{v}$. If $\mathbf{v}$ has the form $\mathbf{v} = a\mathbf{w}$, then

$$f_{a^i b^{l+1-i} a\mathbf{w}} = f_{a^i b^{l+1-i} a} \cdot f_{\mathbf{w}|a}.$$

So all 2^{n-1} columns $[f_{a^i b^{l+1-i} a\mathbf{w}}, \mathbf{w}\{a,b\}^{n-1}]$ are proportional to the single column $[f_{a^i b^{l+1-i} a^n}]$. So we can keep just this one column and throw away the rest. Repeating this argument shows that $H^{(b)}$ has the same rank as the $(l+1) \times (n+1)$ matrix

$$\begin{array}{cc} & \begin{array}{ccccc} ba^n & b^2a^{n-1} & \ldots & b^na & b^{n+1} \end{array} \\ \begin{array}{c} a^l \\ a^{l-1}b \\ \vdots \\ ab^{l-1} \\ b^l \end{array} & \left[\begin{array}{ccccc} f_{a^l ba^n} & f_{a^l b^2 a^{n-1}} & \ldots & f_{a^l b^n a} & f_{a^l b^{n+1}} \\ f_{a^{l-1} b^2 a^n} & f_{a^{l-1} b^3 a^{n-1}} & \ldots & f_{a^{l-1} b^{n+1} a} & f_{a^{l-1} b^{n+2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{ab^l a^n} & f_{ab^{l+1} a^{n-1}} & \ldots & f_{ab^{l+n-1} a} & f_{ab^{l+n}} \\ f_{b^{l+1} a^n} & f_{b^{l+2} a^{n-1}} & \ldots & f_{b^{l+n} a} & f_{b^{l+n+1}} \end{array} \right] \end{array}.$$

The structure of this matrix becomes clear if we note that

$$\begin{aligned} f_{a^i b^j a^t} &= f_{a^i} \cdot f_{b^j a^t|a} \\ &= f_{a^i} \cdot f_{b^j a|a} \cdot f_{a^{t-1}|a} \\ &= \pi_0 (1-\delta)^{i-1} \cdot h_j \cdot (1-\delta)^{t-1}. \end{aligned} \tag{10.8}$$

The strings in the last row and column either do not begin with a, or end with a, or both. So let us divide the first row by $\pi_0(1-\delta)^{l-1}$, the second row by $\pi_0(1-\delta)^{l-2}$, etc., the l-th row by π_0, and do nothing to the last row. Similarly, let us divide the first column by $(1-\delta)^{n-1}$, the second column by $(1-\delta)^{n-2}$, etc., the n-th column by $(1-\delta)^0 = 1$, and leave the last column as is. The resulting matrix has the same rank as $H^{(b)}_{l,n}$, and the matrix is

$$\begin{bmatrix} h_1 & h_2 & \ldots & h_n & \times \\ h_2 & h_3 & \ldots & h_{n+1} & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_l & h_{l+1} & \ldots & h_{l+n} & \times \\ \times & \times & \ldots & \times & \times \end{bmatrix},$$

where × denotes a number whose value does not matter. Now the upper left $l \times n$ submatrix is a submatrix of $\bar{H}$; as a result its rank is bounded by 3. This proves(10.7).[3]

To carry on our analysis of this example, we make use of z-transforms. This is not done in [41], but it simplifies the arguments to follow. As shown earlier, the z-transform of the sequence $\{h_i\}$ is given by

$$\tilde{h}(z) = \frac{1}{4}\left[\frac{\lambda\zeta}{1-\lambda\zeta z} + \frac{\lambda\zeta^{-1}}{1-\lambda\zeta^{-1}z} - 2\frac{\lambda}{1-\lambda z}\right] = \frac{\psi_h(z)}{\phi(z)},$$

where

$$\phi(z) := (1-\lambda\zeta)(1-\lambda\zeta^{-1})(1-\lambda), \tag{10.9}$$

and $\psi_h(z)$ is some polynomial of degree no larger than two; its exact form does not matter. Next, recall that

$$\gamma_i = \sum_{j=i}^{\infty} h_j.$$

Now it is an easy exercise to show that

$$\tilde{\gamma}(z) = \frac{\delta - \tilde{h}(z)}{1-z},$$

where, as defined earlier, $\delta = \sum_{i=1}^{\infty} h_i$. Even though we are dividing by $1-z$ in the above expression, in reality $\tilde{\gamma}$ does not have a pole at $z=1$, because $\tilde{h}(1) = \delta$. Hence we can write

$$\tilde{\gamma}(z) = \frac{\psi_\gamma(z)}{\phi(z)},$$

where again ψ_γ is some polynomial of degree no larger than two, and $\phi(z)$ is defined in (10.9). By entirely similar reasoning, it follows from the expression

$$\eta_i = \sum_{j=i}^{\infty} \gamma_j$$

[3] Through better book-keeping, Dharmadhikari and Nadkarni [41] show that the rank is bounded by 3, not 4. This slight improvement is not worthwhile since all that matters is that the rank is finite.

that

$$\tilde{\eta}(z) = \frac{s - \tilde{\gamma}(z)}{1 - z},$$

where

$$s := \sum_{i=1}^{\infty} \gamma_i = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} h_j = \sum_{j=1}^{\infty} \sum_{i=1}^{j} h_j = \sum_{j=1}^{\infty} j h_j.$$

Here again, $\tilde{\eta}(\cdot)$ does not have a pole at $z = 1$, and in fact

$$\tilde{\gamma}(z) = \frac{\psi_\eta(z)}{\phi(z)},$$

where ψ_η is also a polynomial of degree no larger than two. The point of all these calculations is to show that each of the quantities γ_l, η_l has the form

$$\gamma_l = c_{0,\gamma} \lambda^l + c_{1,\gamma} \lambda^l \zeta^l + c_{2,\gamma} \lambda^l \zeta^{-l}, \tag{10.10}$$

$$\eta_l = c_{0,\eta} \lambda^l + c_{1,\eta} \lambda^l \zeta^l + c_{2,\eta} \lambda^l \zeta^{-l}, \tag{10.11}$$

for appropriate constants. Note that, even though ζ is a complex number, the constants occur in conjugate pairs so that γ_l, η_l are always real. And as we have already seen from (10.5), we have

$$h_l = -\frac{1}{2} \lambda^l + \frac{1}{4} \lambda^l \zeta^l + \frac{1}{4} \lambda^l \zeta^{-l}.$$

Now the expression (10.11) leads at once to two very important observations.

Observation 1: Fix some positive number ρ, and compute the weighted average

$$\frac{1}{T} \sum_{l=1}^{T} \rho^{-l} \eta_l =: \theta(\rho, T).$$

Then it follows that

1. If $\rho < \lambda$, then $\theta(\rho, T) \to \infty$ as $T \to \infty$.
2. If $\rho > \lambda$, then $\theta(\rho, T) \to 0$ as $T \to \infty$.
3. If $\rho = \lambda$, then $\theta(\rho, T) \to c_{0,\eta}$ as $T \to \infty$, where $c_{0,\eta}$ is the constant in (10.11).

If $\rho \neq \lambda$, then the behavior of $\theta(\rho, T)$ is determined by that of $(\lambda/\rho)^l$. If $\rho = \lambda$, then the averages of the oscillatory terms $(\lambda\zeta/\rho)^l$ and $(\lambda/\rho\zeta)^l$ will both approach zero, and only the first term in (10.11) contributes to a nonzero average.

Observation 2: Let T be any fixed integer, and consider the moving average

$$\frac{1}{T} \sum_{j=l+1}^{l+T} \lambda^{-j} \eta_j =: \theta_l^T.$$

This quantity does not have a limit as $l \to \infty$ if α is not commensurate with π. To see this, take the z-transform of $\{\theta_l^T\}$. This leads to

$$\tilde{\theta}^T(z) = \frac{\beta^T(z)}{\phi(z)},$$

where $\beta^T(z)$ is some high degree polynomial. After dividing through by $\phi(z)$, we get

$$\tilde{\theta}^T(z) = \beta_q^T(z) + \frac{\beta_r^T(z)}{\phi(z)},$$

where β_q^T is the quotient and β_r^T is the remainder (and thus has degree no more than two). By taking the inverse z-transform, we see that the sequence $\{\theta_l^T\}$ is the sum of two parts: The first part is a sequence having finite support (which we can think of as the 'transient'), and the second is a sequence of the form

$$c_{0,\theta}\lambda^l + c_{1,\theta}\lambda^l\zeta^l + c_{2,\theta}\lambda^l\zeta^{-l}.$$

From this expression it is clear that if α is noncommensurate with π, then θ_l^T does not have a limit as $l \to \infty$.

These two observations are the key to the concluding part of this very long line of reasoning. Suppose by way of contradiction that the output process $\{\mathcal{Y}_t\}$ can be expressed as a function of a Markov process $\{\mathcal{Z}_t\}$ with a finite state space. Let $\mathcal{N} = \{1, \ldots, n\}$ denote the state space, and let π, A denote the stationary distribution and state transition matrix of the Markov chain $\{\mathcal{Z}_t\}$. Earlier we had used these symbols for the Markov chain $\{\mathcal{X}_t\}$, but no confusion should result from this recycling of notation. From Item 2 of Theorem 5.6, it follows that by a symmetric permutation of rows and columns (which corresponds to permuting the labels of the states), A can be arranged in the form

$$A = \begin{bmatrix} P & \mathbf{0} \\ R & Q \end{bmatrix},$$

where the rows of P correspond to the recurring states and those of R to transient states. Similarly, it follows from Item 7 of Theorem 5.6 that the components of π corresponding to transient states are all zero. Hence the corresponding states can be dropped from the set $\mathbb{N}$ without affecting anything. So let us assume that all states are recurrent.

Next, we can partition the state space $\mathbb{N}$ into those states that map into a, and those states that map into b. With the obvious notation, we can partition π as $[\pi_a \;\; \pi_b]$ and the state transition matrix as

$$A = \begin{bmatrix} A_{aa} & A_{ab} \\ A_{ba} & A_{bb} \end{bmatrix}.$$

Moreover, from Theorem 4.8, it follows that we can arrange A_{bb} in the form

$$A_{bb} = \begin{bmatrix} A_{11} & 0 & \ldots & 0 \\ A_{21} & A_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ A_{s1} & A_{s1} & \ldots & A_{ss} \end{bmatrix},$$

where s is the number of communicating classes within those states that map into the output b, and each of the diagonal matrices A_{ii} is irreducible. Of course, the fact that each of the diagonal blocks is irreducible does still not suffice to determine π uniquely, but as before we can assume that no component of π is zero, because if some component of π is zero, then we can simply drop that component from the state space.

Now it is claimed that $\rho(A_{bb}) = \lambda$, where $\rho(\cdot)$ denotes the spectral radius. To show this, recall that if B is an irreducible matrix with spectral radius $\rho(B)$, and $\boldsymbol{\theta}, \boldsymbol{\phi}$ are respectively the (unique strictly positive) row eigenvector and column eigenvector corresponding to the eigenvalue $\rho(B)$, then the 'ergodic average'

$$\frac{1}{T}\sum_{l=1}^{T}[\rho(B)]^{-l}B^{l}$$

converges to the rank one matrix $\phi\theta$ as $T \to \infty$. Now from the triangular structure of A_{bb}, it is easy to see that $\rho(A_{bb})$ is the maximum amongst the numbers $\rho(A_{ii}), i = 1, \ldots, s$. If we let θ_i, ϕ_i denote the unique row and column eigenvectors of A_{ii} corresponding to $\rho(A_{ii})$, it is obvious that

$$\frac{1}{T}\sum_{l=1}^{T}[\rho(A_{bb}]^{-l}A_{bb}^{l} \to \text{ Block Diag } \{\phi_i\theta_i I_{\{\rho(A_{ii})=\rho(A_{bb})\}}\}. \tag{10.12}$$

In other words, if $\rho(A_{ii}) = \rho(A_{bb})$, then the corresponding term $\phi_i\theta_i$ is present in the block diagonal matrix; if $\rho(A_{ii}) < \rho(A_{bb})$, then the corresponding entry in the block diagonal matrix is the zero matrix. Let D denote the block diagonal in (10.12), and note that at least one of the $\rho(A_{ii})$ equals $\rho(A_{bb})$. Hence at least one of the products $\phi_i\theta_i$ is present in the block diagonal matrix D.

From the manner in which the HMM has been set up, it follows that

$$\eta_l = f_{b^l} = \pi_b A_{bb}^{l}\mathbf{e}.$$

In other words, the only way in which we can observe a sequence of l symbols b in succession is for all states to belong to the subset of $\mathbb{N}$ that map into the output b. Next, let us examine the behavior of the quantity

$$\frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\eta_l = \frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\pi_b A_{bb}^{l}\mathbf{e},$$

where $\rho = \rho(A_{bb})$. Now appealing to (10.12) shows that the above quantity has a definite limit as $T \to \infty$. Moreover, since π_b and $\mathbf{e}$ are strictly positive, and the block diagonal matrix D has at least one positive block $\phi_i\theta_i$, it follows that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\eta_l = \pi D\mathbf{e} \in (0, \infty).$$

By Observation 1, this implies that $\rho(A_{bb}) = \lambda$.

Finally (and at long last), let us examine those blocks A_{ii} which have the property that $\rho(A_{ii}) = \rho(A_{bb}) = \rho$. Since each of these is an irreducible matrix, it follows from Theorem 4.11 that each such matrix has a unique 'period' n_i, which is an integer. Moreover, it follows from Theorem 4.23 that A_{ii} has eigenvalues at $\rho \exp(\mathbf{i}2\pi j/n_i)$, $j = 1, \ldots, n_i - 1$, and all other eigenvalues of A_{ii} have magnitude strictly less than ρ. This statement applies *only* to those indices i such that $\rho(A_{ii}) = \rho(A_{bb}) = \rho$. Now let N denote the least common multiple of all these integers n_i. Then it is clear that the matrix A_{bb} has a whole lot of eigenvalues of the form $\rho \exp(\mathbf{i}2\pi j/N)$ for some (though not necessarily all) values of j ranging from 0 to $N-1$; all other eigenvalues of A have magnitude strictly less than ρ. As a result, the quantity

$$\frac{1}{N} \sum_{l=t+1}^{t+N} A_{bb}^l$$

has a definite limit at $t \to \infty$. In turn this implies that the quantity

$$\frac{1}{N} \sum_{l=t+1}^{t+N} \pi_b A_{bb}^l \mathbf{e} = \frac{1}{N} \sum_{l=t+1}^{t+N} \eta_l$$

has a definite limit at $t \to \infty$. However, this contradicts Observation 2, since α is noncommensurate with π. This contradiction shows that the stochastic process $\{\mathcal{Y}_t\}$ cannot be realized as a function of a finite state Markov chain.

10.3 AN ABSTRACT NECESSARY AND SUFFICIENT CONDITION

In this section we reproduce an abstract necessary and sufficient condition for a given probability law to have an HMM realization, as first presented by Heller [55], with a significantly simplified proof due to Picci [89].

Recall that $\mathbb{M}^*$, the set of all finite strings over $\mathbb{M} = \{1, \ldots, m\}$, is a countable set. We let $\mu(\mathbb{M}^*)$ denote the set of all maps $p : \mathbb{M}^* \to [0, 1]$ satisfying the following two conditions:

$$\sum_{u \in \mathbb{M}} p_u = 1, \tag{10.13}$$

$$\sum_{v \in \mathbb{M}} p_{\mathbf{u}v} = p_{\mathbf{u}}, \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.14}$$

Note that by repeated application of (10.14), we can show that

$$\sum_{\mathbf{v} \in \mathbb{M}^l} p_{\mathbf{uv}} = p_{\mathbf{u}}, \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.15}$$

By taking $\mathbf{u}$ to be the empty string, so that $p_{\mathbf{u}} = 1$, we get from the above that

$$\sum_{\mathbf{v} \in \mathbb{M}^l} p_{\mathbf{v}} = 1, \ \forall l. \tag{10.16}$$

We can think of $\mu(\mathbb{M}^*)$ as the set of all frequency assignments to strings in $\mathbb{M}^*$ that are **right-consistent** by virtue of satisfying (10.14).

Definition 10.2 *Given a frequency assignment $p \in \mu(\mathbb{M}^*)$, we say that $\{\pi, M^{(1)}, \ldots, M^{(m)}\}$ is a* **HMM realization** *of p if*

$$\pi \in \mathbb{R}^n_+, \sum_{i=1}^{n} \pi_i = 1, \tag{10.17}$$

$$M^{(u)} \in [0,1]^{n \times n} \ \forall u \in \mathbb{M}, \tag{10.18}$$

$$\left[\sum_{u \in \mathbb{M}} M^{(u)} \right] \mathbf{e}_n = \mathbf{e}_n, \tag{10.19}$$

and finally

$$p_{\mathbf{u}} = \pi M^{(u_1)} \ldots M^{(u_l)} \mathbf{e}_n \ \forall \mathbf{u} \in \mathbb{M}^l. \tag{10.20}$$

Given a frequency distribution $p \in \mu(\mathbb{M}^*)$, for each $u \in \mathbb{M}$ we define the conditional distribution

$$p(\cdot|u) := \mathbf{v} \in \mathbb{M}^* \mapsto \frac{p_{u\mathbf{v}}}{p_u}. \tag{10.21}$$

If by chance $p_u = 0$, we define $p(\cdot|u)$ to equal p. Note that $p(\cdot|u) \in \mu(\mathbb{M}^*)$; that is, $p(\cdot|u)$ is also a frequency assignment map. By applying (10.21) repeatedly, for each $\mathbf{u} \in \mathbb{M}^*$ we can define the conditional distribution

$$p(\cdot|\mathbf{u}) := \mathbf{v} \in \mathbb{M}^* \mapsto \frac{p_{\mathbf{uv}}}{p_{\mathbf{u}}}. \tag{10.22}$$

Again, for each $\mathbf{u} \in \mathbb{M}^*$, the conditional distribution $p(\cdot|\mathbf{u})$ is also a frequency assignment. Clearly conditioning can be applied recursively and the results are consistent. Thus

$$p((\cdot|\mathbf{u})|\mathbf{v}) = p(\cdot|\mathbf{uv}), \ \forall \mathbf{u}, \mathbf{v} \in \mathbb{M}^*. \tag{10.23}$$

It is easy to verify that if p satisfies the right-consistency condition (10.14), then so do all the conditional distributions $p(\cdot|\mathbf{u})$ for all $\mathbf{u} \in \mathbb{M}^*$. Thus, if $p \in \mu(\mathbb{M}^*)$, then $p(\cdot|\mathbf{u}) \in \mu(\mathbb{M}^*)$ for all $\mathbf{u} \in \mathbb{M}^*$.

A set $\mathcal{C} \subseteq \mu(\mathbb{M}^*)$ is said to be **polyhedral** if there exist an integer n and distributions $q^{(1)}, \ldots, q^{(n)} \in \mu(\mathbb{M}^*)$ such that $\mathcal{C}$ is the convex hull of these $q^{(i)}$, that is, every $q \in \mathcal{C}$ is a convex combination of these $q^{(i)}$. A set $\mathcal{C} \subseteq \mu(\mathbb{M}^*)$ is said to be **stable** if

$$q \in \mathcal{C} \ \Rightarrow \ q(\cdot|\mathbf{u}) \in \mathcal{C} \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.24}$$

In view of (10.23), (10.24) can be replaced by weaker-looking condition

$$q \in \mathcal{C} \ \Rightarrow \ q(\cdot|u) \in \mathcal{C} \ \forall u \in \mathbb{M}. \tag{10.25}$$

Now we are ready to state the main result of this section, first proved in [55]. However, the proof below follows [89] with some slight changes in notation.

Theorem 10.3 *A frequency distribution $p \in \mu(\mathbb{M}^*)$ has a HMM realization if and only if there exists a stable polyhedral set $\mathcal{C} \subseteq \mu(\mathbb{M}^*)$ containing p.*

Proof: "If" Suppose $q^{(1)}, \ldots, q^{(n)} \in \mu(\mathbb{M}^*)$ are the generators of the polyhedral set $\mathcal{C}$. Thus every $q \in \mathcal{C}$ is of the form

$$q = \sum_{i=1}^{n} a_i q^{(i)}, a_i \geq 0, \sum_{i=1}^{n} a_i = 1.$$

In general neither the integer n nor the individual distributions $q^{(i)}$ are unique, but this does not matter. Now, since $\mathcal{C}$ is stable, $q(\cdot|u) \in \mathcal{C}$ for all $a \in \mathcal{C}, u \in \mathbb{M}$. In particular, for each i, u, there exist constants $\alpha_{ij}^{(u)}$ such that

$$q^{(i)}(\cdot|u) = \sum_{j=1}^{n} \alpha_{ij}^{(u)} q^{(j)}(\cdot), \alpha_{ij}^{(u)} \geq 0, \sum_{j=1}^{n} \alpha_{ij}^{(u)} = 1.$$

Thus from (10.21) it follows that

$$\begin{aligned} q_{u\mathbf{v}}^{(i)} &= \sum_{j=1}^{n} q_u^{(i)} \alpha_{ij}^{(u)} q_{\mathbf{v}}^{(j)} \\ &= \sum_{j=1}^{n} m_{ij}^{(u)} q_{\mathbf{v}}^{(j)}, \end{aligned} \tag{10.26}$$

where

$$m_{ij}^{(u)} := q_u^{(i)} \alpha_{ij}^{(u)}, \ \forall i, j, u. \tag{10.27}$$

We can express (10.26) more compactly by using matrix notation. For $\mathbf{u} \in \mathbb{M}^*$, define

$$\mathbf{q_u} := [q_{\mathbf{u}}^{(1)} \ldots q_{\mathbf{u}}^{(n)}]^t \in [0,1]^{n \times 1}.$$

Then (10.26) states that

$$\mathbf{q}_{u\mathbf{v}} = M^{(u)} \mathbf{q_v} \ \forall u \in M, \mathbf{v} \in \mathbb{M}^*,$$

where $M^{(u)} = [m_{ij}^{(u)}] \in [0,1]^{n \times n}$. Moreover, it follows from (10.23) that

$$\mathbf{q_{uv}} = M^{(u_1)} \ldots M^{(u_l)} \mathbf{q_v} \ \forall \mathbf{u} \mathbb{M}^l, \mathbf{v} \in \mathbb{M}^*.$$

If we define

$$M^{(\mathbf{u})} := M^{(u_1)} \ldots M^{(u_l)} \ \forall \mathbf{u} \in \mathbb{M}^l,$$

then the above equation can be written compactly as

$$\mathbf{q_{uv}} = M^{(\mathbf{u})} M^{(\mathbf{v})} \mathbf{q_v} \ \forall \mathbf{u}, \mathbf{v} \in \mathbb{M}^*. \tag{10.28}$$

By assumption, $p \in \mathcal{C}$. Hence there exist numbers $\pi_1, \ldots, \pi_n$, not necessarily unique, such that

$$p(\cdot) = \sum_{i=1}^{n} \pi_i q^{(i)}(\cdot), \pi_i \geq 0 \ \forall i, \sum_{i=1}^{n} \pi_i = 1. \tag{10.29}$$

We can express (10.29) as

$$p(\cdot) = \pi \mathbf{q}(\cdot).$$

Hence, for all $\mathbf{u}, \mathbf{v} \in \mathbb{M}^*$, it follows from (10.28) that

$$p_{\mathbf{uv}} = \pi \mathbf{q}_{\mathbf{uv}} = \pi M^{(\mathbf{u})} \mathbf{q}_{\mathbf{v}},\ \forall \mathbf{u}, \mathbf{v} \in \mathbb{M}^*. \tag{10.30}$$

In particular, if we let $\mathbf{v}$ equal the empty string, then $\mathbf{q}_{\mathbf{v}} = \mathbf{e}_n$, and $p_{\mathbf{uv}} = p_{\mathbf{u}}$. Thus (10.30) becomes

$$p_{\mathbf{u}} = \pi M^{(\mathbf{u})} \mathbf{e}_n,$$

which is the same as (10.20).

Next, we verify (10.19) by writing it out in component form. We have

$$\begin{aligned} \sum_{j=1}^{n} \sum_{u \in \mathbb{M}} m_{ij}^{(u)} &= \sum_{u \in \mathbb{M}} \sum_{j=1}^{n} m_{ij}^{(u)} \\ &= \sum_{u \in \mathbb{M}} q_u^{(i)} \left[\sum_{j=1}^{n} \alpha_{ij}^{(u)} \right] \\ &= \sum_{u \in \mathbb{M}} q_u^{(i)} \text{ because } \sum_{j=1}^{n} \alpha_{ij}^{(u)} = 1 \\ &= 1,\ \forall i \text{ because } q^{(i)} \in \mu(\mathbb{M}^*) \text{ and (10.13).} \end{aligned}$$

Before leaving the "If" part of the proof, we observe that if the probability distribution $p \in \mu(\mathbb{M}^*)$ is also *left-consistent* by satisfying

$$\sum_{u \in \mathbb{M}} p_{u\mathbf{v}} = p_{\mathbf{v}}\ \forall u \in \mathbb{M}, \mathbf{v} \in \mathbb{M}^*,$$

then *it is possible* to choose the vector π such that

$$\pi \left[\sum_{u \in \mathbb{M}} M^{(u)} \right] = \pi. \tag{10.31}$$

To see this, we substitute into (10.20) which has already been established. This gives

$$\pi M^{(\mathbf{v})} \mathbf{e}_n = p_{\mathbf{v}} = \sum_{u \in \mathbb{M}} p_{u\mathbf{v}} = \pi \left[\sum_{u \in \mathbb{M}} M^{(u)} \right] M^{(\mathbf{v})} \mathbf{e}_n,\ \forall \mathbf{v} \in \mathbb{M}^*.$$

Now it is not possible to "cancel" $M^{(\mathbf{v})} \mathbf{e}_n$ from both sides of the above equation. However, it is always possible to choose the coefficient vector π so as to satisfy (10.31).

"Only if" Suppose p has a HMM realization $\{\pi, M^{(1)}, \ldots, M^{(m)}\}$. Let n denote the dimension of the matrices $M(u)$ and the vector π. Define the distributions $q^{(1)}, \ldots, q^{(n)}$ by

$$\mathbf{q}_{\mathbf{u}} = [q_{\mathbf{u}}^{(1)} \ldots q_{\mathbf{u}}^{(n)}]^t := M^{(\mathbf{u})} \mathbf{e}_n,\ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.32}$$

Thus $q_{\mathbf{u}}^{(i)}$ is the i-th component of the column vector $M^{(\mathbf{u})}\mathbf{e}_n$. First it is shown that each $q^{(i)}$ is indeed a frequency distribution. From (10.32), it follows that

$$\sum_{u \in \mathbb{M}} \mathbf{q}_u = \left[\sum_{u \in \mathbb{M}} M^{(u)} \right] \mathbf{e}_n = \mathbf{e}_n,$$

where we make use of (10.19). Thus each $q^{(i)}$ satisfies (10.13) (with p replaced by $q^{(i)}$). Next, to show that each $q^{(i)}$ is right-consistent, observe that for each $\mathbf{u} \in \mathbb{M}^*, v \in \mathbb{M}$ we have

$$\sum_{v \in \mathbb{M}} \mathbf{q}_{\mathbf{u}v} = M^{(\mathbf{u})} \left[\sum_{v \in \mathbb{M}} M^{(v)} \right] \mathbf{e}_n = M^{(\mathbf{u})}\mathbf{e}_n = \mathbf{q}_{\mathbf{u}}.$$

Thus each $q^{(i)}$ is right-consistent. Finally, to show that the polyhedral set consisting of all convex combinations of $q^{(1)}, \ldots, q^{(n)}$ is stable, observe that

$$q^{(i)}(\mathbf{v}|u) = \frac{q^{(i)}(u\mathbf{v})}{q^{(i)}(u)}.$$

Substituting from (10.31) gives

$$\begin{aligned} q^{(i)}(\mathbf{v}|u) &= \frac{1}{q^{(i)}(u)} M^{(u)} M^{(\mathbf{v})} \mathbf{e}_n \\ &= \mathbf{a}_u^{(i)} M^{(\mathbf{v})} \mathbf{e}_n, \\ &= \mathbf{a}_u^{(i)} \mathbf{q}_{\mathbf{v}}, \end{aligned} \tag{10.33}$$

where

$$\mathbf{a}_u^{(i)} := \left[\frac{m_{ij}^{(u)}}{q^{(i)}(u)}, j = 1, \ldots, n \right] \in [0,1]^{1 \times n}.$$

Thus each conditional distribution $q^{(i)}(\cdot|u)$ is a linear combination of $q^{(1)}(\cdot), \ldots, q^{(n)}(\cdot)$. It remains only to show that $q^{(i)}(\cdot|u)$ is a *convex* combination, that is, that each $\mathbf{a}_u^{(i)} \in \mathbb{R}_+^n$ and that $\mathbf{a}_u^{(i)}\mathbf{e}_n = 1$. The first is obvious from the definition of the vector $\mathbf{a}^{(i)}$. To establish the second, substitute $\mathbf{v}$ equal to the empty string in (10.33). Then $q^{(i)}(\mathbf{v}|u) = 1$ for all i, u, and $\mathbf{q}_{\mathbf{v}} = \mathbf{e}_n$. Substituting these into (10.32) shows that

$$1 = \mathbf{a}_u^{(i)} \mathbf{e}_n,$$

as desired. Thus the polyhedral set $\mathcal{C}$ consisting of all convex combinations of the $q^{(i)}$ is stable. Finally, it is obvious from (10.20) that p is a convex combination of the $q^{(i)}$ and thus belongs to $\mathcal{C}$.

10.4 EXISTENCE OF REGULAR QUASI-REALIZATIONS

In this section, we study processes whose Hankel rank is finite, and show that it is *always* possible to construct a 'quasi-realization' of such a process. Moreover, any two *regular* quasi-realizations of a finite Hankel rank process are related through a similarity transformation.

Definition 10.4 *Suppose a process $\{\mathcal{Y}_t\}$ has finite Hankel rank r. Suppose $n \geq r$, $\mathbf{x}$ is a row vector in $\mathbb{R}^n$, $\mathbf{y}$ is a column vector in $\mathbb{R}^n$, and $C^{(u)} \in \mathbb{R}^{n \times n}$ $\forall u \in \mathbb{M}$. Then we say that $\{n, \mathbf{x}, \mathbf{y}, C^{(u)}, \mathbf{u} \in \mathbb{M}\}$ is a* **quasi-realization** *of the process if three conditions hold. First,*

$$f_{\mathbf{u}} = \mathbf{x} C^{(u_1)} \ldots C^{(u_l)} \mathbf{y} \ \forall \mathbf{u} \in \mathbb{M}^*, \tag{10.34}$$

where $l = |\mathbf{u}|$. Second,

$$\mathbf{x} \left[\sum_{u \in \mathbb{M}} C^{(u)} \right] = \mathbf{x}. \tag{10.35}$$

Third,

$$\left[\sum_{u \in \mathbb{M}} C^{(u)} \right] \mathbf{y} = \mathbf{y}. \tag{10.36}$$

We say that $\{n, \mathbf{x}, \mathbf{y}, C^{(u)}, \mathbf{u} \in \mathbb{M}\}$ is a **regular quasi-realization** *of the process if $n = r$, the rank of the Hankel matrix.*

The formula (10.34) is completely analogous to (10.3). Similarly, (10.35) and (10.36) are analogous to (10.4). The only difference is that the various quantities are not required to be nonnegative. This is why we speak of a 'quasi-realization' instead of a true realization. With this notion, it is possible to prove the following powerful statements:

1. Suppose the process $\{\mathcal{Y}_t\}$ has finite Hankel rank, say r. Then the process always has a regular quasi-realization.

2. Suppose a process $\{\mathcal{Y}_t\}$ has finite Hankel rank r, and suppose $\{\theta_1, \phi_1, D_1^{(u)}, u \in \mathbb{M}\}$ and $\{\theta_2, \phi_2, D_2^{(u)}, u \in \mathbb{M}\}$ are two regular quasi-realizations of this process. Then there exists a nonsingular matrix T such that

$$\theta_2 = \theta_1 T^{-1}, D_2^{(u)} = T D_1^{(u)} T^{-1} \ \forall u \in \mathbb{M}, \phi_2 = T \phi_1.$$

These two statements are formally stated and proven as Theorem 10.8 and Theorem 10.9 respectively.

The results of this section are not altogether surprising. Given that the infinite matrix H has finite rank, it is clear that there *must exist* recursive relationships between its various elements. Earlier work, most notably [37, 25], contains some such recursive relationships. However, the present formulae are the cleanest, and also the closest to the conventional formula (10.3). Note that Theorem 10.8 is more or less contained in the work of Erickson [44]. In [61], the authors generalize the work of Erickson by studying the relationship between two quasi-realizations, *without* assuming that the underlying state spaces have the same dimension. In this case, in place of the similarity transformation above, they obtain 'intertwining' conditions of the form $D_2^{(u)} T = T D_1^{(u)}$, where the matrix T may now be rectangular. In the interests of simplicity, in the present case we do not study this more general case. Moreover, the above formulae are the basis for the construction of a 'true' (as opposed to quasi) HMM realization in subsequent sections.

Some notation is introduced to facilitate the subsequent proofs. Suppose k, l are integers, and $I \subseteq \mathbb{M}^k, J \subseteq \mathbb{M}^l$; thus every element of I is a string of length k, while every element of J is a string of length l. Specifically, suppose $I = \{\mathbf{i}_1, \ldots, \mathbf{i}_{|I|}\}$, and $J = \{\mathbf{j}_1, \ldots, \mathbf{j}_{|J|}\}$. Then we define

$$F_{I,J} := \begin{bmatrix} f_{\mathbf{i}_1\mathbf{j}_1} & f_{\mathbf{i}_1\mathbf{j}_2} & \cdots & f_{\mathbf{i}_1\mathbf{j}_{|J|}} \\ f_{\mathbf{i}_2\mathbf{j}_1} & f_{\mathbf{i}_2\mathbf{j}_2} & \cdots & f_{\mathbf{i}_2\mathbf{j}_{|J|}} \\ \vdots & \vdots & \vdots & \vdots \\ f_{\mathbf{i}_{|I|}\mathbf{j}_1} & f_{\mathbf{i}_{|I|}\mathbf{j}_2} & \cdots & f_{\mathbf{i}_{|I|}\mathbf{j}_{|J|}} \end{bmatrix}. \tag{10.37}$$

Thus $F_{I,J}$ is a submatrix of $F_{k,l}$ and has dimension $|I| \times |J|$. This notation is easily reconciled with the earlier notation. Suppose k, l are integers. Then we can think of $F_{k,l}$ as shorthand for $F_{\mathbb{M}^k,\mathbb{M}^l}$. In the same spirit, if I is a subset of $\mathbb{M}^k$ and l is an integer, we use the 'mixed' notation $F_{I,l}$ to denote $F_{I,\mathbb{M}^l}$. This notation can be extended in an obvious way to the case where either k or l equals zero. If $l = 0$, we have that $\mathbb{M}^0 := \{\emptyset\}$. In this case

$$F_{I,0} := [f_{\mathbf{i}} : \mathbf{i} \in I] \in \mathbb{R}^{|I| \times 1}.$$

Similarly if $J \subseteq \mathbb{M}^l$ for some integer l, then

$$F_{0,J} := [f_{\mathbf{j}} : \mathbf{j} \in J] \in \mathbb{R}^{1 \times |J|}.$$

Finally, given any string $\mathbf{u} \in \mathbb{M}^*$, we define

$$F_{k,l}^{(\mathbf{u})} := [f_{\mathbf{iuj}}, \mathbf{i} \in \mathbb{M}^k \text{ in flo}, \mathbf{j} \in \mathbb{M}^l \text{ in llo}], \tag{10.38}$$

$$F_{I,J}^{(\mathbf{u})} := [f_{\mathbf{iuj}}, \mathbf{i} \in I, \mathbf{j} \in J]. \tag{10.39}$$

Lemma 10.5 *Suppose H has finite rank. Then there exists a smallest integer k such that*

$$Rank(F_{k,k}) = Rank(H).$$

Moreover, for this k, we have

$$Rank(F_{k,k}) = Rank(H_{k+l,k+s}), \ \forall l, s \geq 0. \tag{10.40}$$

Proof: We begin by observing that, for every pair of integers k, l, we have

$$\text{Rank}(H_{k,l}) = \text{Rank}(F_{k,l}). \tag{10.41}$$

To see this, observe that the row indexed by $\mathbf{u} \in \mathbb{M}^{k-1}$ in $F_{k-1,s}$ is the sum of the rows indexed by $v\mathbf{u}$ in $F_{k,s}$, for each s. This follows from (10.1). Similarly each row in $F_{k-2,s}$ is the sum of m rows in $F_{k-1,s}$ and thus of m^2 rows of $F_{k,s}$, and so on. Thus it follows that every row of $F_{t,s}$ for $t < k$ is a sum of m^{k-t} rows of $F_{k,s}$. Therefore

$$\text{Rank}(H_{k,l}) = \text{Rank}([F_{k,0} \ \ F_{k,1} \ \ \ldots F_{k,l}]).$$

Now repeat the same argument for the columns of this matrix. Every column of $F_{k,t}$ is the sum of m^{k-t} columns of $F_{k,l}$. This leads to the desired conclusion (10.41).

To complete the proof, observe that, since $H_{l,l}$ is a submatrix of $H_{l+1,l+1}$, we have that

$$\mathrm{Rank}(H_{1,1}) \leq \mathrm{Rank}(H_{2,2}) \leq \ldots \leq \mathrm{Rank}(H).$$

Now at each step, there are only two possibilities: Either $\mathrm{Rank}(H_{l,l}) < \mathrm{Rank}(H_{l+1,l+1})$, or else $\mathrm{Rank}(H_{l,l}) = \mathrm{Rank}(H_{l+1,l+1})$. Since $\mathrm{Rank}(H)$ is finite, the first possibility can only occur finitely many times. Hence there exists a smallest integer k such that

$$\mathrm{Rank}(H_{k,k}) = \mathrm{Rank}(H).$$

We have already shown that $\mathrm{Rank}(H_{k,k}) = \mathrm{Rank}(F_{k,k})$. Finally, since $H_{k+l,k+s}$ is a submatrix of H and contains $H_{k.k}$ as a submatrix, the desired conclusion (10.40) follows.

Note: Hereafter, the symbol k is used *exclusively* for this integer and nothing else. Similarly, hereafter the symbol r is used exclusively for the (finite) rank of the Hankel matrix H and nothing else.

Now consider the matrix $F_{k,k}$, which is chosen so as to have rank r. Thus there exist sets $I, J \subseteq \mathbb{M}^k$, such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. (Recall the definition of the matrix $F_{I,J}$ from (10.37).) In other words, the index sets I, J are chosen such that $F_{I,J}$ is any full rank nonsingular submatrix of $F_{k,k}$. Of course the choice of I and J is not unique. However, once I, J are chosen, there exist *unique* matrices $U \in \mathbb{R}^{m^k \times r}, V \in \mathbb{R}^{r \times m^k}$ such that $F_{k,k} = UF_{I,J}V$. Hereafter, the symbols U, V are used only for these matrices and nothing else.

The next lemma shows that, once the index sets I, J are chosen (thus fixing the matrices U and V), the relationship $F_{k,k} = UF_{I,J}V$ can be extended to strings of *arbitrary* lengths.

Lemma 10.6 *With the various symbols defined as above, we have*

$$F_{k,k}^{(\mathbf{u})} = UF_{I,J}^{(\mathbf{u})}V, \ \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.42}$$

This result can be compared to [6], Lemma 1, p. 99.

Proof: For notational convenience only, let us suppose I, J consist of the first r elements of $\mathbb{M}^r$. The more general case can be handled through more messy notation. The matrix U can be partitioned as follows:

$$U = \begin{bmatrix} I_r \\ \bar{U} \end{bmatrix}.$$

This is because $F_{I,k}$ is a submatrix of $F_{k,k}$. (In general we would have to permute the indices so as to bring the elements of I to the first r positions.) Now, by the rank condition and the assumption that $F_{k,k} = UF_{I,J}V (= UF_{I,k})$, it follows that

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k-r} \end{bmatrix} H_{k,\cdot} = \begin{bmatrix} F_{I,k} & F_{I,\cdot} \\ \mathbf{0} & F_{\mathbb{M}^k \setminus I,\cdot} - \bar{U}F_{I,\cdot} \end{bmatrix},$$

where

$$F_{I,.} = [F_{I,k+1} \;\; F_{I,k+2} \ldots], \text{ and } F_{\mathbb{M}^k \setminus I,.} = [F_{\mathbb{M}^k \setminus I,k+1} \;\; F_{\mathbb{M}^k \setminus I,k+2} \ldots].$$

This expression allows us to conclude that

$$F_{\mathbb{M}^k \setminus I,.} = \bar{U} F_{I,.}. \tag{10.43}$$

Otherwise the $(2, 2)$-block of the above matrix would contain some nonzero element, which would in turn imply that $\text{Rank}(H_{k,.}) > r$, a contradiction. Now the above relationship implies that

$$F_{k,k}^{(\mathbf{u})} = U F_{I,K}^{(\mathbf{u})}, \; \forall \mathbf{u} \in \mathbb{M}^*.$$

Next, as with U, partition V as $V = [I_r \;\; \bar{V}]$. (In general, we would have to permute the columns to bring the elements of J to the first positions.) Suppose $N > k$ is some integer. Observe that $F_{N,k}$ is just $[F_{k,k}^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^{N-k}$ in flo$]$. Hence

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k-r} \end{bmatrix} H_{.,k} = \begin{bmatrix} F_{I,k}^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^* \text{ in flo} \\ \mathbf{0} \end{bmatrix} = \left[\begin{array}{l} F_{I,k} \\ \mathbf{0} \\ \hline \\ F_{I,k}^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^* \setminus \emptyset \text{ in flo} \\ \mathbf{0} \end{array} \right].$$

Now post-multiply this matrix as shown below:

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k-r} \end{bmatrix} H_{.,k} \begin{bmatrix} I_r & -\bar{V} \\ \mathbf{0} & I_{m^k-r} \end{bmatrix} = \begin{bmatrix} F_{I,J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ F_{I,J}^{(\mathbf{u})} & F_{I,M^k \setminus J}^{(\mathbf{u})} - F_{I,J}^{(\mathbf{u})} \bar{V}, \mathbf{u} \in \mathbb{M}^* \text{in flo} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

So if $F_{I,M^k \setminus J}^{(\mathbf{u})} \neq F_{I,J}^{(\mathbf{u})} \bar{V}$ for some $\mathbf{u} \in \mathbb{M}^*$, then $\text{Rank}(H_{.,k})$ would exceed $\text{Rank}(F_{I,J})$, which is a contradiction. Thus it follows that

$$F_{I,M^k \setminus J}^{(\mathbf{u})} = F_{I,J}^{(\mathbf{u})} \bar{V}, \; \forall \mathbf{u} \in \mathbb{M}^*. \tag{10.44}$$

The two relationships (10.43) and (10.44) can together be compactly expressed as (10.42), which is the desired conclusion.

Lemma 10.7 *Choose unique matrices $\bar{D}^{(u)}, u \in \mathbb{M}$, such that*

$$F_{I,J}^{(u)} = F_{I,J} \bar{D}^{(u)}, \; \forall u \in \mathbb{M}. \tag{10.45}$$

Then for all $\mathbf{u} \in \mathbb{M}^*$, we have

$$F_{I,J}^{(\mathbf{u})} = F_{I,J} \bar{D}^{(u_1)} \ldots \bar{D}^{(u_l)}, \text{ where } l = |\mathbf{u}|. \tag{10.46}$$

Choose unique matrices $D^{(u)}, u \in \mathbb{M}$, such that

$$F_{I,J}^{(u)} = D^{(u)} F_{I,J}, \; \forall u \in \mathbb{M}. \tag{10.47}$$

Then for all $\mathbf{u} \in \mathbb{M}^*$, we have

$$F_{I,J}^{(\mathbf{u})} = D^{(u_1)} \ldots D^{(u_l)} F_{I,J}, \text{ where } l = |\mathbf{u}|. \tag{10.48}$$

This result can be compared to [6], Theorem 1, p. 90.

Proof: We prove only (10.46), since the proof of (10.48) is entirely similar. By the manner in which the index sets I, J are chosen, we have

$$\text{Rank}[F_{I,J} \;\; F_{I,J}^{(u)}] = \text{Rank}[F_{I,J}], \; \forall u \in \mathbb{M}.$$

Hence there exist unique matrices $\bar{D}^{(u)}, u \in \mathbb{M}$ such that (10.45) holds. Now suppose $\mathbf{v}$ is any *nonempty* string in $\mathbb{M}^*$. Then, since $F_{I,J}$ is a maximal rank submatrix of H, it follows that

$$\text{Rank} \begin{bmatrix} F_{I,J} & F_{I,J}^{(u)} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} \end{bmatrix} = \text{Rank} \begin{bmatrix} F_{I,J} \\ F_{I,J}^{(\mathbf{v})} \end{bmatrix}, \; \forall u \in \mathbb{M}.$$

Now post-multiply the matrix on the left side as shown below:

$$\begin{bmatrix} F_{I,J} & F_{I,J}^{(u)} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} \end{bmatrix} \begin{bmatrix} I & -\bar{D}^{(u)} \\ \mathbf{0} & I \end{bmatrix} = \begin{bmatrix} F_{I,J} & \mathbf{0} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} - F_{I,J}^{(\mathbf{v})} \bar{D}^{(u)} \end{bmatrix}.$$

This shows that

$$F_{I,J}^{(\mathbf{v}u)} = F_{I,J}^{(\mathbf{v})} \bar{D}^{(u)}, \; \forall \mathbf{v} \in \mathbb{M}^*, \; \forall u \in \mathbb{M}. \tag{10.49}$$

Otherwise, the (2,2)-block of the matrix on the right side would contain a nonzero element and would therefore have rank larger than that of $F_{I,J}$, which would be a contradiction. Note that if $\mathbf{v}$ is the empty string in (10.49), then we are back to the definition of the matrix $\bar{D}^{(u)}$. Now suppose $\mathbf{u} \in \mathbb{M}^*$ has length l and apply (10.49) recursively. This leads to the desired formula (10.46).

Suppose $\mathbf{u} \in \mathbb{M}^*$ has length l. Then it is natural to define

$$\bar{D}^{(\mathbf{u})} := \bar{D}^{(u_1)} \ldots \bar{D}^{(u_l)}, \; D^{(\mathbf{u})} := D^{(u_1)} \ldots D^{(u_l)}.$$

With this notation let us observe that the matrices $D^{(\mathbf{u})}$ and $\bar{D}^{(\mathbf{u})}$ 'intertwine' with the matrix $F_{I,J}$. That is,

$$F_{I,J} \bar{D}^{(\mathbf{u})} = D^{(\mathbf{u})} F_{I,J}, \text{ and } F_{I,J}^{-1} D^{(\mathbf{u})} = \bar{D}^{(\mathbf{u})} F_{I,J}^{-1}. \tag{10.50}$$

This follows readily from the original relationship

$$F_{I,J} \bar{D}^{(u)} = D^{(u)} F_{I,J} (= F_{I,J}^{(u)}) \; \forall u \in \mathbb{M}$$

applied recursively.

Finally we come to the main theorem about quasi-realizations. We begin by formalizing the notion.

Note that a regular quasi-realization in some sense completes the analogy with the formulas (10.3) and (10.4).

Theorem 10.8 *Suppose the process $\{\mathcal{Y}_t\}$ has finite Hankel rank, say r. Then the process always has a regular quasi-realization. In particular, choose the integer k as in Lemma 10.5, and choose index sets $I, J \subseteq \mathbb{M}^k$ such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. Define the matrices $U, V, D^{(u)}, \bar{D}^{(u)}$ as before. The following two choices are regular quasi-realizations. First, let*

$$\mathbf{x} = \theta := F_{0,J} F_{I,J}^{-1}, \; \mathbf{y} = \phi := F_{I,0}, \; C^{(u)} = D^{(u)} \; \forall u \in \mathbb{M}. \tag{10.51}$$

Second, let

$$\mathbf{x} = \bar{\theta} := F_{0,J}, \; \mathbf{y} = \bar{\phi} := F_{I,J}^{-1} F_{I,0}, \; C^{(u)} = \bar{D}^{(u)} \; \forall u \in \mathbb{M}. \tag{10.52}$$

This result can be compared to [6], Theorem 1, p. 90 and Theorem 2, p. 92.

Proof: With all the spade work done already, the proof is very simple. For any string $\mathbf{u} \in \mathbb{M}^*$, it follows from (10.47) that

$$F_{I,J}^{(\mathbf{u})} = D^{(u_1)} \ldots D^{(u_l)} F_{I,J}, \text{ where } l = |\mathbf{u}|.$$

Next, we have from (10.42) that

$$F_{k,k}^{(\mathbf{u})} = U F_{I,J}^{(\mathbf{u})} V, \; \forall \mathbf{u} \in \mathbb{M}^*.$$

Now observe that, by definition, we have

$$f_{\mathbf{u}} = \sum_{\mathbf{i} \in \mathbb{M}^k} \sum_{\mathbf{j} \in \mathbb{M}^k} f_{\mathbf{iuj}} = \mathbf{e}_{m^k}^t F_{k,k}^{(\mathbf{u})} \mathbf{e}_{m^k} = \mathbf{e}_{m^k}^t U D^{(u_1)} \ldots D^{(u_l)} F_{I,J} V \mathbf{e}_{m^k},$$

where $\mathbf{e}_{m^k}$ is the column vector with m^k one's. Hence (10.34) is satisfied with the choice

$$n = r, \theta := \mathbf{e}_{m^k}^t U, \phi := F_{I,J} V \mathbf{e}_{m^k}, C^{(u)} = D^{(u)} \; \forall u \in \mathbb{M},$$

and the matrices $D^{(u)}$ as defined in (10.47). Since $D^{(\mathbf{u})} F_{I,J} = F_{I,J} \bar{D}^{(\mathbf{u})}$, we can also write

$$f_{\mathbf{u}} = \mathbf{e}_{m^k}^t U F_{I,J} \bar{D}^{(u_1)} \ldots \bar{D}^{(u_l)} V \mathbf{e}_{m^k}.$$

Hence (10.34) is also satisfied with the choice

$$n = r, \bar{\theta} := \mathbf{e}_{m^k}^t U F_{I,J}, \bar{\phi} := V \mathbf{e}_{m^k}, C^{(u)} = \bar{D}^{(u)} \; \forall u \in \mathbb{M},$$

and the matrices $\bar{D}^{(u)}$ as defined in (10.46).

Next, we show that the vectors $\theta, \phi, \bar{\theta}, \bar{\phi}$ can also be written as in (10.51) and (10.52). For this purpose, we proceed as follows:

$$\theta = \mathbf{e}_{m^k}^t U = \mathbf{e}_{m^k}^t U F_{I,J} F_{I,J}^{-1} = \mathbf{e}_{m^k}^t F_{k,J} F_{I,J}^{-1} = F_{0,J} F_{I,J}^{-1}.$$

Therefore

$$\bar{\theta} = \theta F_{I,J} = F_{0,J}.$$

Similarly

$$\phi = F_{I,J} V \mathbf{e}_{m^k} = F_{I,k} \mathbf{e}_{m^k} = F_{I,0},$$

and

$$\bar{\phi} = F_{I,J}^{-1} F_{I,0}.$$

It remains only to prove the eigenvector properties. For this purpose, note that, for each $u \in \mathbb{M}$, we have

$$F_{0,J} \bar{D}^{(u)} = \mathbf{e}_{m^k}^t U F_{I,J} \bar{D}^{(u)} = \mathbf{e}_{m^k}^t U F_{I,J}^{(u)} = F_{0,J}^{(u)}.$$

Now

$$\theta D^{(u)} = F_{0,J} F_{I,J}^{-1} D^{(u)} = F_{0,J} \bar{D}^{(u)} F_{I,J}^{-1} = F_{0,J}^{(u)} F_{I,J}^{-1}.$$

Hence

$$\theta \left[\sum_{u \in \mathbb{M}} D^{(u)} \right] = \sum_{u \in \mathbb{M}} \theta D^{(u)} = \sum_{u \in \mathbb{M}} F_{0,J}^{(u)} F_{I,J}^{-1} = F_{0,J} F_{I,J}^{-1} = \theta,$$

since

$$\sum_{u \in \mathbb{M}} F_{0,J}^{(u)} = F_{0,J}.$$

As for ϕ, we have

$$D^{(u)} \phi = D^{(u)} F_{I,J} V \mathbf{e}_{m^k} = F_{I,J}^{(u)} V \mathbf{e}_{m^k} = F_{I,k}^{(u)} \mathbf{e}_{m^k} = F_{I,0}^{(u)}.$$

Hence

$$\left[\sum_{u \in \mathbb{M}} D^{(u)} \right] \phi = \sum_{u \in \mathbb{M}} D^{(u)} \phi = \sum_{u \in \mathbb{M}} F_{I,0}^{(u)} = F_{I,0} = \phi.$$

This shows that $\{r, \theta, \phi, D^{(u)}\}$ is a quasi-realization. The proof in the case of the barred quantities is entirely similar. We have

$$\bar{\theta} \bar{D}^{(u)} = F_{0,J} \bar{D}^{(u)} = F_{0,J}^{(u)},$$

so

$$\bar{\theta} \left[\sum_{u \in \mathbb{M}} \bar{D}^{(u)} \right] = \sum_{u \in \mathbb{M}} F_{0,J}^{(u)} = F_{0,J} = \bar{\theta}.$$

It can be shown similarly that

$$\left[\sum_{u \in \mathbb{M}} \bar{D}^{(u)} \right] \bar{\phi} = \bar{\phi}.$$

Next, it is shown that any two 'regular' quasi-realizations of the process are related through a similarity transformation.

Theorem 10.9 *Suppose a process $\{\mathcal{Y}_t\}$ has finite Hankel rank r, and suppose $\{\theta_1, \phi_1, D_1^{(u)}, u \in \mathbb{M}\}$ and $\{\theta_2, \phi_2, D_2^{(u)}, u \in \mathbb{M}\}$ are two regular quasi-realizations of this process. Then there exists a nonsingular matrix T such that*

$$\theta_2 = \theta_1 T^{-1}, D_2^{(u)} = T D_1^{(u)} T^{-1} \; \forall u \in \mathbb{M}, \phi_2 = T \phi_1.$$

Proof: Suppose the process has finite Hankel rank, and let r denote the rank of H. Choose the integer k as before, namely, the smallest integer k such that $\text{Rank}(F_{k,k}) = \text{Rank}(H)$. Choose subsets $I, J \subseteq \mathbb{M}^k$ such that $|I| = |J| = r$ and $\text{Rank}(F_{I,J}) = r$. Up to this point, all entities depend only on the process and its Hankel matrix (which depends on the law of the process), and not on the specific quasi-realization. Moreover, the fact that I, J are not unique is not important.

Now look at the matrix $F_{I,J}$, and express it in terms of the two quasi-realizations. By definition,

$$F_{I,J} = \begin{bmatrix} f_{\mathbf{i}_1\mathbf{j}_1} & \dots & f_{\mathbf{i}_1\mathbf{j}_r} \\ \vdots & \vdots & \vdots \\ f_{\mathbf{i}_r\mathbf{j}_1} & \dots & f_{\mathbf{i}_r\mathbf{j}_r} \end{bmatrix}.$$

Now, since we are given two quasi-realizations, the relationship (10.34) holds for each quasi-realization. Hence

$$F_{I,J} = \begin{bmatrix} \theta_s D_s^{(\mathbf{i}_1)} \\ \vdots \\ \theta_s D_s^{(\mathbf{i}_r)} \end{bmatrix} [D_s^{(\mathbf{j}_1)}\phi_s \dots D_s^{(\mathbf{j}_r)}\phi_s], \text{ for } s = 1, 2.$$

Define

$$P_s := \begin{bmatrix} \theta_s D_s^{(\mathbf{i}_1)} \\ \vdots \\ \theta_s D_s^{(\mathbf{i}_r)} \end{bmatrix}, \; Q_s := [D_s^{(\mathbf{j}_1)}\phi_s \dots D_s^{(\mathbf{j}_r)}\phi_s], \text{ for } s = 1, 2.$$

Then $F_{I,J} = P_1Q_1 = P_2Q_2$. Since $F_{I,J}$ is nonsingular, so are P_1, Q_1, P_2, Q_2. Moreover,

$$P_2^{-1}P_1 = Q_2Q_1^{-1} =: T, \text{ say}.$$

Next, fix $u \in \mathbb{M}$ and consider the $r \times r$ matrix $F_{I,J}^{(u)}$. We have from (10.34) that

$$F_{I,J}^{(u)} = P_1 D_1^{(u)} Q_1 = P_2 D_2^{(u)} Q_2.$$

Hence

$$D_2^{(u)} = P_2^{-1}P_1 D_1^{(u)} Q_1 Q_2^{-1} = T D_1^{(u)} T^{-1}, \; \forall u \in \mathbb{M}.$$

Finally, we can factor the entire matrix H as

$$H = [\theta_s D_s^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^* \text{ in flo}][D_s^{(\mathbf{v})}\phi_s, \mathbf{v} \in \mathbb{M}^* \text{ in llo}], \; s = 1, 2,$$

where

$$D^{(\mathbf{u})} := D^{(u_1)} \dots D^{(u_l)}, \; l = |\mathbf{u}|,$$

and $D^{(\mathbf{v})}$ is defined similarly. Note that the first matrix in the factorization of H has r columns and infinitely many rows, while the second matrix has r rows and infinitely many columns. Thus there exists a nonsingular matrix, say S, such that

$$[\theta_2 D_2^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^* \text{ in flo}] = [\theta_1 D_1^{(\mathbf{u})}, \mathbf{u} \in \mathbb{M}^* \text{ in flo}]S^{-1},$$

and

$$[D_2^{(\mathbf{v})}\phi_2, \mathbf{v} \in \mathbb{M}^* \text{ in llo}] = S[D_1^{(\mathbf{v})}\phi_1, \mathbf{v} \in \mathbb{M}^* \text{ in llo}].$$

Choosing $\mathbf{u} = \mathbf{i}_1, \dots, \mathbf{i}_r$ and $\mathbf{v} = \mathbf{j}_1, \dots, \mathbf{j}_r$ shows that in fact $S = T$. Finally, choosing $\mathbf{u} = \mathbf{v} = \emptyset$ shows that

$$\theta_2 = \theta_1 T^{-1}, \; \phi_2 = T\phi_1.$$

We conclude this section with an example from [37] of a regular quasi-realization that does not correspond to a regular realization.

Let $n = 4$, and define the 4×4 'state transition matrix'

$$A = \begin{bmatrix} \lambda_1 & 0 & 0 & 1-\lambda_1 \\ 0 & -\lambda_2 & 0 & 1+\lambda_2 \\ 0 & 0 & -\lambda_3 & 1+\lambda_3 \\ 1-\lambda_1 & c(1+\lambda_2) & -c(1+\lambda_3) & \lambda_1 + c(\lambda_3 - \lambda_2) \end{bmatrix},$$

as well as the 'output matrix'

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

It is easy to see that $A\mathbf{e}_4 = \mathbf{e}_4$, that is, the matrix A is 'stochastic.' Similarly $B\mathbf{e}_2 = \mathbf{e}_2$ and so B is stochastic (without quotes). Let $\mathbf{b}_i$ denote the i-th column of B, and let $\text{Diag}(\mathbf{b}_i)$ denote the diagonal 4×4 matrix with the elements of $\mathbf{b}_i$ on the diagonal. Let us define

$$C^{(1)} = A\text{Diag}(\mathbf{b}_1) = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & -\lambda_2 & 0 & 0 \\ 0 & 0 & -\lambda_3 & 0 \\ 1-\lambda_1 & c(1+\lambda_2) & -c(1+\lambda_3) & 0 \end{bmatrix},$$

$$C^{(2)} = A\text{Diag}(\mathbf{b}_2) = \begin{bmatrix} 0 & 0 & 0 & 1-\lambda_1 \\ 0 & 0 & 0 & 1+\lambda_2 \\ 0 & 0 & 0 & 1+\lambda_3 \\ 0 & 0 & 0 & \lambda_1 + c(\lambda_3 - \lambda_2) \end{bmatrix}.$$

Then $C^{(1)} + C^{(2)} = A$. Note that

$$\mathbf{x} = [0.5 \ \ 0.5c \ \ -0.5c \ \ 0.5]$$

is a 'stationary distribution' of A; that is, $\mathbf{x}A = \mathbf{x}$. With these preliminaries, we can define the 'quasi-frequencies'

$$f_{\mathbf{u}} = \mathbf{x}C^{(u_1)} \ldots C^{(u_l)}\mathbf{e}_4,$$

where $\mathbf{u} = u_1 \ldots u_l$. Because $\mathbf{x}$ and $\mathbf{e}_4$ are respectively row and column eigenvectors of A corresponding to the eigenvalue one, these quasi-frequencies satisfy the consistency conditions (10.1) and (10.2). Thus, in order to qualify as a quasi-realization, the only thing missing is the property that $f_{\mathbf{u}} \geq 0$ for all strings $\mathbf{u}$.

This nonnegativity property is established in [37] using a Markov chain analogy, and is not reproduced here. All the frequencies will all be nonnegative provided the following inequalities are satisfied:

$$0 < \lambda_i < 1, i = 1, 2, 3; \lambda_1 > \lambda_i, i = 2, 3; 0 < c < 1,$$

$$\lambda_1 + c(\lambda_3 - \lambda_2) > 0; (1 - \lambda_1)^k > c(1 + \lambda_i)^k, i = 2, 3, k = 1, 2.$$

One possible choice (given in [37]) is

$$\lambda_1 = 0.5, \lambda_2 = 0.4, \lambda_3 = 0.3, c = 0.06.$$

Thus the above is a quasi-realization.

To test whether this quasi-realization can be made into a realization (with nonnegative elements), we can make use of Theorem 10.9. *All possible* quasi-realizations of this process can be obtained by performing a similarity transformation on the above quasi-realization. Thus there exists a regular realization (not quasi-realization) of this process if and only if there exists a nonsingular matrix T such that $\mathbf{x}T^{-1}, TC^{(i)}T^{-1}, T\mathbf{e}_4$ are all nonnegative. This can in turn be written as the feasibility of a linear program, namely:

$$\pi T = \mathbf{x}; TC^{(i)} = M^{(i)}T, i = 1, 2; T\mathbf{e}_4 = \mathbf{e}_4; M^{(i)} \geq \mathbf{0}, i = 1, 2; \pi \geq \mathbf{0}.$$

It can be readily verified that the above linear program is *not* feasible, so that there is no regular realization for this process, only regular quasi-realizations.

As pointed out above, it is possible to check in polynomial time whether a given regular quasi-realization can be converted into a regular realization of a stationary process. There is a related problem that one can examine, namely: Suppose one is given a triplet $\{\mathbf{x}, C^{(u)}, u \in \mathbb{M}, \mathbf{y}\}$ with compatible dimensions. The problem is to determine whether the triple product

$$f_{\mathbf{u}} := \mathbf{x}C^{(\mathbf{u})}\mathbf{y} = \mathbf{x}C^{(u_1)} \cdots C^{(u_l)}\mathbf{y} \geq 0 \ \forall \mathbf{u} \in \mathbb{M}^l, \ \forall l.$$

This problem can be viewed as one of deciding whether a given rational power series always has nonnegative coefficients. This problem is known to be undecidable; see [94], Theorem 3.13. Even if $m = 2$, the above problem is undecidable if $n \geq 50$, where n is the size of the vector $\mathbf{x}$. The arguments of [21] can be adapted to prove this claim.[4] Most likely the problem remains undecidable even if we add the additional requirements that

$$\mathbf{x}\left[\sum_{u \in \mathbb{M}} C^{(u)}\right] = \mathbf{x},$$

$$\left[\sum_{u \in \mathbb{M}} C^{(u)}\right]\mathbf{y} = \mathbf{y},$$

because the above two conditions play no role in determining the nonnegativity or otherwise of the 'quasi-frequencies' $f_{\mathbf{u}}$, but serve only to assure that these quasi-frequencies are consistent.

10.5 SPECTRAL PROPERTIES OF ALPHA-MIXING PROCESSES

In this section, we add the assumption that the finite Hankel rank process under study is also α-mixing, and show that the regular quasi-realizations have an additional property, namely: The matrix that plays the role of the

[4] Thanks to Vincent Blondel for these references.

state transition matrix in the HMM has a spectral radius of one, this eigenvalue is simple, and all other eigenvalues have magnitude strictly less than one. This property is referred to as the 'quasi strong Perron property.' As a corollary, it follows that if an α-mixing process has a regular realization (and not just a quasi-realization), then the underlying Markov chain is irreducible and aperiodic.

We begin by reminding the reader about the notion of α-mixing. Suppose the process $\{\mathcal{Y}_t\}$ is defined on the probability space (S, Ω), where Ω is a σ-algebra on the set S. For each pair of indices s, t with $s < t$, define Σ_s^t to be the σ-algebra (a subalgebra of Ω) generated by the random variables $\mathcal{Y}_s, \ldots, \mathcal{Y}_t$. Then the **$\alpha$-mixing coefficient** $\alpha(l)$ of the process $\{\mathcal{Y}_t\}$ is defined as

$$\alpha(l) := \sup_{A \in \Sigma_0^t, B \in \Sigma_{t+l}^\infty} |P(A \cap B) - P(A)P(B)|.$$

The process $\{\mathcal{Y}_t\}$ is said to be **α-mixing** if $\alpha(l) \to 0$ as $l \to \infty$. Note that in the definition above, A is an event that depends strictly on the 'past' random variables before time t, whereas B is an event that depends strictly on the 'future' random variables after time $t+l$. If the future were to be completely independent of the past, we would have $P(A \cap B) = P(A)P(B)$. Thus the α-mixing coefficient measures the extent to which the future is independent of the past.

Remark: As will be evident from the proofs below, actually we do *not* make use of the α-mixing property of the process $\{\mathcal{Y}_t\}$. Rather, what is needed is that

$$\sum_{\mathbf{w} \in \mathbb{M}^l} f_{\mathbf{uwv}} \to f_{\mathbf{u}} f_{\mathbf{v}} \text{ as } l \to \infty, \ \forall \mathbf{u}, \mathbf{v} \in \mathbb{M}^k, \tag{10.53}$$

where k is the *fixed* integer arising from the finite Hankel rank condition. Since the process assumes values in a finite alphabet, (10.53) is equivalent to the condition

$$\max_{A \in \Sigma_1^k, B \in \Sigma_{l+k+1}^{2k}} |P(A \cap B) - P(A)P(B)| \to 0 \text{ as } l \to \infty. \tag{10.54}$$

To see this, suppose that (10.54) holds, and choose A to be the event $(y_1, \ldots, y_k) = \mathbf{u}$, and similarly, choose B to be the event $(y_{l+k+1}, \ldots, y_{l+2k}) = \mathbf{v}$, for some $\mathbf{u}, \mathbf{v} \in \mathbb{M}^k$. Then it is clear that $A \cap B$ is the event that a string of length $l + 2k$ begins with $\mathbf{u}$ and ends with $\mathbf{v}$. Thus

$$P(A) = f_{\mathbf{u}}, \ P(B) = f_{\mathbf{v}}, \ P(A \cap B) = \sum_{\mathbf{w} \in \mathbb{M}^l} f_{\mathbf{uwv}}.$$

Hence (10.54) implies (10.53). To show the converse, suppose (10.53) holds. Then (10.54) also holds for elementary events A and B. Since k is a *fixed* number and the alphabet of the process is finite, both of the σ-algebras Σ_1^k, Σ_{l+k+1}^{2k} are *finite* unions of elementary events. Hence (10.53) is enough to imply (10.54). It is not known whether (10.54) is *strictly weaker* than α-mixing for processes assuming values over a finite alphabet.

Now we state the main result of this section.

Theorem 10.10 *Suppose the process $\{\mathcal{Y}_t\}$ is α-mixing and has finite Hankel rank r. Let $\{r, \mathbf{x}, \mathbf{y}, C^{(\mathbf{u})}, u \in \mathbb{M}\}$ be any regular quasi-realization of the process, and define*

$$S := \sum_{u \in \mathbb{M}} C^{(\mathbf{u})}.$$

Then $S^l \to \mathbf{yx}$ as $l \to \infty$, $\rho(S) = 1$, $\rho(S)$ is a simple eigenvalue of S, and all other eigenvalues of S have magnitude strictly less than one.

This theorem can be compared with [6], Theorem 4, p. 94.

Proof: It is enough to prove the theorem for the *particular* quasi-realization $\{r, \theta, \phi, D^{(u)}, u \in \mathbb{M}\}$ defined in (10.35). This is because there exists a nonsingular matrix T such that $C^{(u)} = T^{-1} D^{(u)} T$ for all u, and as a result the matrices $\sum_{u \in \mathbb{M}} C^{(u)}$ and $\sum_{u \in \mathbb{M}} D^{(u)}$ have the same spectrum. The α-mixing property implies that, for each $\mathbf{i} \in I, \mathbf{j} \in J$, we have

$$\sum_{\mathbf{w} \in \mathbb{M}^l} f_{\mathbf{iwj}} \to f_{\mathbf{i}} f_{\mathbf{j}} \text{ as } l \to \infty. \tag{10.55}$$

This is a consequence of (10.53) since both I and J are subsets of $\mathbb{M}^k$. Now note that, for each fixed $\mathbf{w} \in \mathbb{M}^l$, we have from (10.34) that

$$[f_{\mathbf{iwj}}, \mathbf{i} \in I, \mathbf{j} \in J] = F_{I,J}^{(\mathbf{w})} = D^{(\mathbf{w})} F_{I,J}, \tag{10.56}$$

where, as per earlier convention, we write

$$D^{(\mathbf{w})} := D^{(w_1)} \ldots D^{(w_l)}.$$

It is clear that

$$\sum_{\mathbf{w} \in \mathbb{M}^l} D^{(\mathbf{w})} = \left[\sum_{u \in \mathbb{M}} D^{(u)} \right]^l = S^l. \tag{10.57}$$

Now (10.55) implies that

$$\sum_{\mathbf{w} \in \mathbb{M}^l} [f_{\mathbf{iwj}}, \mathbf{i} \in I, \mathbf{j} \in J] \to [f_{\mathbf{i}}, i \in I][f_{\mathbf{j}}, j \in J] =: F_{I,0} F_{0,J},$$

where $F_{I,0}$ is an r-dimensional column vector and $F_{0,J}$ is an r-dimensional row vector. Moreover, combining (10.56) and (10.57) shows that

$$S^l F_{I,J} \to F_{I,0} F_{0,J},$$

and since $F_{I,J}$ is nonsingular, that

$$S^l \to F_{I,0} F_{0,J} F_{I,J}^{-1} = \phi\theta \text{ as } l \to \infty.$$

So the conclusion is that S^l approaches $\phi\theta$, which is a rank one matrix, as $l \to \infty$. Moreover, this rank one matrix has one eigenvalue at one and the rest at zero. To establish this, we show that

$$F_{0,J} F_{I,J}^{-1} F_{I,0} = 1.$$

This is fairly straight-forward. Note that $F_{0,J}F_{I,J}^{-1} = \theta$ and $F_{I,0} = \phi$ as defined in (10.35). Then taking $\mathbf{u}$ to be the empty string in (10.34) (and of course, substituting $\mathbf{x} = \theta, \mathbf{y} = \phi$) shows that $\theta\phi = 1$, which is the desired conclusion. Let A denote the rank one matrix

$$A := F_{I,0}F_{0,J}F_{I,J}^{-1}.$$

Then $S^l \to A$ as $l \to \infty$. Suppose the spectrum of the matrix S is $\{\lambda_1, \ldots, \lambda_n\}$, where $n = m^k$, and $|\lambda_1| = \rho(S)$. Then, since the spectrum of S^l is precisely $\{\lambda_1^l, \ldots, \lambda_n^l\}$, it follows that

$$\{\lambda_1^l, \ldots, \lambda_n^l\} \to \{1, 0, \ldots, 0\} \text{ as } l \to \infty.$$

Here we make use of the facts that A is a rank one matrix, and that its spectrum consists of $n-1$ zeros plus one. This shows that S has exactly one eigenvalue on the unit circle, namely at $\lambda = 1$, and the remaining eigenvalues are all inside the unit circle.

Corollary 10.11 *Suppose a stationary process $\{\mathcal{Y}_t\}$ is α-mixing and has a regular realization. Then the underlying Markov chain is aperiodic and irreducible.*

Proof: Suppose that the process under study has a regular realization (and not just a regular quasi-realization). Let A denote the state transition matrix of the corresponding Markov process $\{\mathcal{X}_t\}$. From Theorem 10.9, it follows that A is similar to the matrix S defined in Theorem 10.10. Moreover, if the process $\{\mathcal{Y}_t\}$ is α-mixing, then the matrix A (which is similar to S) satisfies the strong Perron property. In other words, it has only one eigenvalue on the unit circle, namely a simple eigenvalue at one. Hence the Markov chain $\{\mathcal{X}_t\}$ is irreducible and aperiodic.

10.6 ULTRA-MIXING PROCESSES AND THE EXISTENCE OF HMM'S

In the previous two sections, we studied the existence of quasi-realizations. In this section, we study the existence of 'true' (as opposed to quasi) realizations. We introduce a new property known as 'ultra-mixing' and show that if a process has finite Hankel rank, and is both α-mixing as well as ultra-mixing, then modulo a technical condition it has a HMM where the underlying Markov chain is itself α-mixing (and hence aperiodic and irreducible) or else satisfies a 'consistency condition.' The converse is also true, modulo another technical condition.

The material in this section is strongly influenced by [6]. In that paper, the author *begins* with the assumption that the stochastic process under study is generated by an irreducible HMM (together with a few other assumptions), and then gives a constructive procedure for constructing an irreducible HMM for the process. Thus the paper does not give a set of conditions for the existence of a HMM *in terms of the properties of the process* *under study*. Moreover, even with the assumptions in [6], the order of the HMM constructed using the given procedure can in general be much larger than the order of the HMM that generates the process in the first place. In contrast, in the present paper we give conditions explicitly in terms of the process under study, that are sufficient to guarantee the existence of an irreducible HMM. However, the proof techniques used here borrow heavily from [6].

10.6.1 Constructing a Hidden Markov Model

We begin with a rather 'obvious' result that sets the foundation for the material to follow.

Lemma 10.12 *Suppose $\{\mathcal{Y}_t\}$ is a stationary process over a finite alphabet $\mathcal{M}$. Then the process $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM if and only if there exist an integer n, a stochastic row vector $\mathbf{h}$, and $n \times n$ nonnegative matrices $G^{(1)}, \ldots, G^{(m)}$ such that the following statements are true.*

1. *The matrix $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is stochastic, in that each of its rows adds up to one. Equivalently, $\mathbf{e}_n$ is a column eigenvector of Q corresponding to the eigenvalue one.*

2. *$\mathbf{h}$ is a row eigenvector $\mathbf{h}$ of Q corresponding to the eigenvalue one, i.e., $\mathbf{h}Q = \mathbf{h}$.*

3. *For every $\mathbf{u} \in \mathcal{M}^*$, we have*
$$f_{\mathbf{u}} = \mathbf{h}G^{(u_1)} \cdots G^{(u_l)}\mathbf{e}_n,$$
where $l = |\mathbf{u}|$.

In this case there exists a Markov process $\{\mathcal{X}_t\}$ evolving over $\mathbb{N} := \{1, \ldots, n\}$ such that the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is a Type 3 HMM.

Proof: One half of this lemma has already been proven in the course of proving Theorem 10.1. Suppose $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM model. Let $\{\mathcal{X}_t\}$ denote the associated Markov process. Define the matrices $M^{(1)}, \ldots, M^{(m)}$ as in (9.6). and let π denote the stationary distribution of the process $\{\mathcal{X}_t\}$. Then it is clear that the conditions of the lemma are satisfied with $\mathbf{h} = \pi$ and $G^{(u)} = M^{(u)}$ for each $u \in \mathcal{M}$.

To prove the converse, suppose $\mathbf{h}, G^{(1)}, \ldots, G^{(m)}$ exist that satisfy the stated conditions. Let $\{\mathcal{Z}_t\}$ be a stationary Markov process with the state transition matrix
$$A_{\mathcal{Z}} := \left[\begin{array}{cccc} G^{(1)} & G^{(2)} & \ldots & G^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ G^{(1)} & G^{(2)} & \ldots & G^{(m)} \end{array} \right].$$
and the stationary distribution
$$\pi_{\mathcal{Z}} = [\mathbf{h}G^{(1)}| \ldots |\mathbf{h}G^{(m)}].$$

To show that $\pi_{\mathcal{Z}}$ is indeed a stationary distribution of $A_{\mathcal{Z}}$, partition $\pi_{\mathcal{Z}}$ in the obvious fashion as $[\pi_1 \ldots \pi_m]$, and observe that $\pi_v = \mathbf{h}G^{(v)}$. Then, because of the special structure of the matrix $A_{\mathcal{Z}}$, in order to be a stationary distribution of the Markov chain, the vector $\pi_{\mathcal{Z}}$ needs to satisfy the relationship

$$\left[\sum_{v \in \mathcal{M}} \pi_v \right] \cdot G^{(u)} = \pi_u. \tag{10.58}$$

Now observe that

$$\left[\sum_{v \in \mathcal{M}} \pi_v \right] = \mathbf{h} \sum_{v \in \mathcal{M}} G^{(v)} = \mathbf{h}Q = \mathbf{h}.$$

Hence the desired relationship (10.58) follows readily. Now the stationary distribution of the $\mathcal{X}_t$ process is clearly $\sum_{v \in \mathcal{M}} \mathbf{h}G^{(v)} = \mathbf{h}$. Hence, by the formula (10.3), it follows that the frequencies of the $\mathcal{Y}_t$ process are given by

$$f_{\mathbf{u}} = \mathbf{h}G^{(u_1)} \cdots G^{(u_l)} \mathbf{e}_n.$$

This is the desired conclusion.

10.6.2 The Consistency Condition

Before presenting the sufficient condition for the existence of a HMM, we recall a very important result from [6]. Consider a 'joint Markov process' HMM where the associated matrix A (the transition matrix of the $\{\mathcal{X}_t\}$ process) is irreducible. In this case, it is well known and anyway rather easy to show that the state process $\{\mathcal{X}_t\}$ is α-mixing if and only if the matrix A is aperiodic in addition to being irreducible. If A is aperiodic (so that the state process is α-mixing), then the output process $\{\mathcal{Y}_t\}$ is also α-mixing. However, the converse is not always true. It is possible for the output process to be α-mixing even if the state process is not. Theorem 5 of [6] gives necessary and sufficient conditions for this to happen. We reproduce this important result below.

Suppose a 'joint Markov process' HMM has n states and that the state transition matrix A is irreducible. Let π denote the unique positive stationary probability distribution of the $\mathcal{X}_t$ process. As in (9.6), define the matrices $M^{(u)}, u \in \mathcal{M}$ by

$$m_{ij}^{(u)} = \Pr\{\mathcal{X}_1 = j \& \mathcal{Y}_1 = u | \mathcal{X}_0 = i\}, 1 \leq i, j \leq n, u \in \mathcal{M}.$$

Let p denote the number of eigenvalues of A on the unit circle (i.e., the period of the Markov chain). By renumbering the states if necessary, rearrange A so that it has the following cyclic form:

$$A = \left[\begin{array}{ccccc} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & A_1 \\ A_p & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{p-1} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & A_2 & \mathbf{0} \end{array} \right], \tag{10.59}$$

where all blocks have the same size $(n/p) \times (n/p)$ (which clearly implies that p is a divisor of n). The matrices $M^{(u)}$ inherit the same zero block structure as A; so the notation $M_i^{(u)}$ is unambiguous. For a string $\mathbf{u} \in \mathcal{M}^l$, define

$$M_i^{(\mathbf{u})} := M_i^{(u_1)} M_{i+1}^{(u_2)} \ldots M_{i+l-1}^{(u_l)},$$

where the subscripts on M are taken modulo p. Partition π into p equal blocks, and label them as π_1 through π_p.

Theorem 10.13 *The output process $\{\mathcal{Y}_t\}$ is α-mixing if and only if, for every string $\mathbf{u} \in \mathcal{M}^*$, the following 'consistency conditions' hold:*

$$\pi_1 M_1^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \pi_2 M_p^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \pi_3 M_{p-1}^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \ldots = \pi_p M_2^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \frac{1}{p} \pi M^{(\mathbf{u})} \mathbf{e}_n. \tag{10.60}$$

For a proof, see [6], Theorem 5.

10.6.3 The Ultra-Mixing Property

In earlier sections, we studied the spectrum of various matrices under the assumption that the process under study is α-mixing. For present purposes, we introduce a different kind of mixing property.

Definition 10.14 *Given the process $\{\mathcal{Y}_t\}$, suppose it has finite Hankel rank, and let k denote the unique integer defined in Lemma 10.5. Then the process $\{\mathcal{Y}_t\}$ is said to be* **ultra-mixing** *if there exists a sequence $\{\delta_l\} \downarrow 0$ such that*

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuv}}}{f_{\mathbf{uv}}} \right| \leq \delta_l, \ \forall \mathbf{i} \in \mathcal{M}^k, \mathbf{u} \in \mathcal{M}^l, \mathbf{v} \in \mathcal{M}^*. \tag{10.61}$$

Note that, the way we have defined it here, the notion of ultra-mixing is defined only for processes with finite Hankel rank.

In [64], Kalikow defines a notion that he calls a 'uniform martingale,' which is the same as an ultra-mixing stochastic process. He shows that a stationary stochastic process over a finite alphabet is a uniform martingale if and only if it is also a 'random Markov process,' which is defined as follows: A process $\{(\mathcal{Y}_t, N_t)\}$ where $\mathcal{Y}_t \in \mathcal{M}$ and N_t is a positive integer (natural number) for each t is said to be a 'random Markov process' if (i) The process $\{N_t\}$ is independent of the $\{\mathcal{Y}_t\}$ process, and (ii) for each t, we have

$$\Pr\{\mathcal{Y}_t | \mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \ldots\} = \Pr\{\mathcal{Y}_t | \mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \ldots, \mathcal{Y}_{t-N_t}\}.$$

Observe that if N_t equals a fixed integer N for all t, then the above condition says that $\{\mathcal{Y}_t\}$ is an N-step Markov process. Hence a 'random Markov process' is an N_t-step Markov process where the length of the 'memory' N_t is itself random and independent of $\mathcal{Y}_t$. One of the main results of [64] is that the ultra-mixing property is equivalent to the process being random Markov. However, the random Markov property seems to be quite different in spirit from a process having a HMM.

The ultra-mixing property can be interpreted as a kind of long-term indepedence. It says that the conditional probability that a string begins with $\mathbf{i}$, given the next l entries, is just about the same whether we are given just the next l entries, or the next l entries as well as the still later entries. This property is also used in [6]. It does not appear straight-forward to relate ultra-mixing to other notions of mixing such as α-mixing. This can be seen from the treatment of [6], Section 11, where the author assumes (in effect) that the process under study is *both* ultra-mixing as well as α-mixing.

10.6.4 The Main Result

Starting with the original work of Dharmadhikari [38], 'cones' have played a central role in the construction of HMM's. The present paper continues that tradition. Moreover, cones also play an important role in the so-called positive realization problem. Hence it is not surprising that the conditions given here also borrow a little bit from positive realization theory. See [16] for a survey of the current status of this problem.

Recall that a set $\mathcal{S} \subseteq \mathbb{R}^r$ is said to be a 'cone' if $\mathbf{x}, \mathbf{y} \in \mathcal{S} \ \Rightarrow \ \alpha\mathbf{x} + \beta\mathbf{y} \in \mathcal{S} \ \forall \alpha, \beta \geq 0$. The term 'convex cone' is also used to describe such an object. Given a (possibly infinite) set $\mathcal{V} \subseteq \mathbb{R}^r$, the symbol Cone($\mathcal{V}$) denotes the smallest cone containing $\mathcal{V}$, or equivalently, the intersection of all cones containing $\mathcal{V}$. If $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a finite set, then it is clear that

$$\text{Cone}(\mathcal{V}) = \{\sum_{i=1}^{n} \alpha_i \mathbf{v}_i : \ \alpha_i \geq 0 \ \forall i\}.$$

In such a case, Cone($\mathcal{V}$) is said to be 'polyhedral' and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are said to be 'generators' of the cone. Note that, in the way we have defined the concept here, the generators of a polyhedral cone are not uniquely defined. It is possible to refine the definition; however, the above definition is sufficient for the present purposes. Finally, given a cone $\mathcal{C}$ (polyhedral or otherwise), the 'polar cone' $\mathcal{C}^p$ is defined by

$$\mathcal{C}^p := \{\mathbf{y} \in \mathbb{R}^r : \mathbf{y}^t \mathbf{x} \geq 0 \ \forall \mathbf{x} \in \mathcal{C}\}.$$

It is easy to see that $\mathcal{C}^p$ is also a cone, and that $\mathcal{C} \subseteq (\mathcal{C}^p)^p$.

Next, we introduce two cones that play a special role in the proof. Suppose as always that the process under study has finite Hankel rank, and define the integer k as in Lemma 10.5. Throughout, we use the quasi-realization $\{r, \theta, \phi, D^{(u)}\}$ defined in (10.35). Now define

$$\mathcal{C}_c := \text{Cone}\{D^{(\mathbf{u})}\phi : \mathbf{u} \in \mathcal{M}^*\},$$

$$\mathcal{C}_o := \{\mathbf{y} \in \mathbb{R}^r : \theta D^{(\mathbf{v})} \mathbf{y} \geq 0, \ \forall \mathbf{v} \in \mathcal{M}^*\}.$$

The subscripts o and c have their legacy from positive realization theory, where $\mathcal{C}_c$ is called the 'controllability cone' and $\mathcal{C}_o$ is called the 'observability cone.' See for example [16]. However, in the present context, we could have used any other symbols. Note that from (10.34) and (10.35) we have

$$\theta D^{(\mathbf{v})} D^{(\mathbf{u})} \phi = f_{\mathbf{uv}} \geq 0, \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^*.$$

Hence $D^{(\mathbf{u})}\phi \in \mathcal{C}_o \ \forall \mathbf{u} \in \mathcal{M}^*$, and as a result $\mathcal{C}_c \subseteq \mathcal{C}_o$. Moreover, both $\mathcal{C}_c$ and $\mathcal{C}_o$ are invariant under $D^{(w)}$ for each $w \in \mathcal{M}$. To see this, let $w \in \mathcal{M}$ be arbitrary. Then $D^{(w)}D^{(\mathbf{u})}\phi = D^{(w\mathbf{u})}\phi$, for all $\mathbf{u} \in \mathcal{M}^*$. Hence

$$D^{(w)}\mathcal{C}_c = \text{Cone}\{D^{(w\mathbf{u})}\phi : \mathbf{u} \in \mathcal{M}^*\} \subseteq \mathcal{C}_c.$$

Similarly, suppose $\mathbf{y} \in \mathcal{C}_o$. Then the definition of $\mathcal{C}_o$ implies that $\theta D^{(\mathbf{v})}\mathbf{y} \geq 0$ for all $\mathbf{v} \in \mathcal{M}^*$. Therefore

$$\theta D^{(\mathbf{v})} D^{(w)}\mathbf{y} = \theta D^{(\mathbf{v}w)}\mathbf{y} \geq 0 \ \forall \mathbf{v} \in \mathcal{M}^*.$$

Hence $D^{(w)}\mathbf{y} \in \mathcal{C}_c$. The key difference between $\mathcal{C}_c$ and $\mathcal{C}_o$ is that the former cone need not be closed, whereas the latter cone is always closed (this is easy to show).

In order to state the sufficient condition for the existence of a HMM, a few other bits of notation are introduced. Suppose the process under study has finite Hankel rank, and let k be the unique integer defined in Lemma 10.5. Let r denote the rank of the Hankel matrix, and choose subsets $I, J \subseteq \mathcal{M}^k$ such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. For each finite string $\mathbf{u} \in \mathcal{M}^*$, define the vectors

$$\mathbf{p_u} := \frac{1}{f_\mathbf{u}} F_{I,0}^{(\mathbf{u})} = [f_{\mathbf{iu}}/f_\mathbf{u}, i \in I] \in [0,1]^{r\times 1}, \ \mathbf{q_u} := \frac{1}{f_\mathbf{u}} F_{0,J}^{(\mathbf{u})} = [f_{\mathbf{uj}}/f_\mathbf{u}, j \in J] \in [0,1]^{1\times r}.$$

The interpretation of $\mathbf{p_u}$ is that the $\mathbf{i}$-th component of this vector is the conditional probability, given that the last part of a sample path consists of the string $\mathbf{u}$, that the immediately preceding k symbols are $\mathbf{i}$. The vector $\mathbf{q_u}$ is interpreted similarly. The $\mathbf{j}$-th component of this vector is the conditional probability, given that the first part of a sample path consists of the string $\mathbf{u}$, that the next k symbols are $\mathbf{j}$.

Lemma 10.15 *Let $\|\cdot\|$ denote the ℓ_1-norm on $\mathbb{R}^r$. Then there exists a constant $\gamma > 0$ such that*

$$\gamma \leq \|\mathbf{p_u}\| \leq 1, \gamma \leq \|\mathbf{q_u}\| \leq 1, \ \forall \mathbf{u} \in \mathcal{M}^*.$$

Proof: Note that the vector $[f_{\mathbf{iu}}/f_\mathbf{u}, \mathbf{i} \in \mathcal{M}^k]$ is a probability vector, in the sense that its components are nonnegative and add up to one. Hence this vector has ℓ_1-norm of one. Since $\mathbf{p_u}$ is a subvector of it, it follows that $\|\mathbf{p_u}\| \leq 1$. On the other hand, we have

$$[f_{\mathbf{iu}}/f_\mathbf{u}, \mathbf{i} \in \mathcal{M}^k] = U\mathbf{p_u}, \ \forall \mathbf{u} \in \mathcal{M}^*,$$

and U has full column rank. Hence $\|\mathbf{p_u}\|$ is bounded away from zero independently of $\mathbf{u}$. Similar arguments apply to $\mathbf{q_u}$.

The vectors $\mathbf{p_u}$ and $\mathbf{q_u}$ satisfy some simple recurrence relationships.

Lemma 10.16 *Suppose $\mathbf{u}, \mathbf{v} \in \mathcal{M}^*$. Then*

$$D^{(\mathbf{u})}\mathbf{p_v} = \frac{f_{\mathbf{uv}}}{f_\mathbf{v}}\mathbf{p_{uv}}, \ \mathbf{q_u}C^{(\mathbf{v})} = \frac{f_{\mathbf{uv}}}{f_\mathbf{u}}\mathbf{q_{uv}}.$$

Proof: From Lemmas 10.6 and 10.7, it follows that

$$F_{I,0}^{(\mathbf{v})} = F_{I,k}^{(\mathbf{v})} \mathbf{e}_{m^k} = D^{(\mathbf{v})} F_{I,J} V \mathbf{e}_{m^k} = D^{(\mathbf{v})} \phi, \ \forall \mathbf{v} \in \mathcal{M}^*.$$

This shows that

$$\mathbf{p_v} = \frac{1}{f_\mathbf{v}} F_{I,0}^{(\mathbf{v})} = \frac{1}{f_\mathbf{v}} D^{(\mathbf{v})} \phi.$$

Hence, for arbitrary $\mathbf{u}, \mathbf{v} \in \mathcal{M}^*$, we have

$$D^{(\mathbf{u})} \mathbf{p_v} = \frac{1}{f_\mathbf{v}} D^{(\mathbf{u})} D^{(\mathbf{v})} \phi = \frac{1}{f_\mathbf{v}} D^{(\mathbf{uv})} \phi = \frac{f_\mathbf{uv}}{f_\mathbf{v}} \mathbf{p_{uv}}.$$

The proof in the case of $\mathbf{q_v}$ is entirely similar.

Now let us consider the countable collection of probability vectors $\mathcal{A} := \{\mathbf{p_u} : \mathbf{u} \in \mathcal{M}^*\}$. Since $\mathbf{p_u}$ equals $D^{(\mathbf{u})}\phi$ within a scale factor, it follows that $\mathcal{C}_c = \text{Cone}(\mathcal{A})$. Moreover, since $\mathcal{A} \subseteq \mathcal{C}_c \subseteq \mathcal{C}_o$ and $\mathcal{C}_o$ is a closed set, it follows that the set of cluster points of $\mathcal{A}$ is also a subset of $\mathcal{C}_o$.[5] Finally, it follows from Lemma 10.15 that every cluster point of $\mathcal{A}$ has norm no smaller than γ.

Now we state the main result of this section.

Theorem 10.17 *Suppose the process $\{\mathcal{Y}_t\}$ satisfies the following conditions:*

1. *It has finite Hankel rank.*
2. *It is ultra-mixing.*
3. *It is α-mixing.*
4. *The cluster points of the set $\mathcal{A}$ of probability vectors are finite in number and lie in the interior of the cone $\mathcal{C}_o$.*

Under these conditions, the process has an irreducible 'joint Markov process' hidden Markov model. Moreover the HMM satisfies the consistency conditions (10.60).

Remark: Among the hypotheses of Theorem 10.17, Conditions 1 through 3 are 'real' conditions, whereas Condition 4 is a 'technical' condition.

The proof proceeds via two lemmas. The first lemma gives insight into the behaviour of the matrix $D^{(\mathbf{u})}$ as $|\mathbf{u}| \to \infty$. To put these lemmas in context, define the matrix $S = \sum_{u \in \mathcal{M}} D^{(u)}$. Then by Theorem 10.10, we know that if the process $\{\mathcal{Y}_t\}$ is α-mixing, then S^l approaches a rank one matrix as $l \to \infty$. In the present case it is shown that, if the process is ultra-mixing, then *each individual* matrix $D^{(\mathbf{u})}$ approaches a rank one matrix as $|\mathbf{u}| \to \infty$. This result has no counterpart in earlier literature and may be of independent interest.

[5] Recall that a vector $\mathbf{y}$ is said to be a 'cluster point' of $\mathcal{A}$ if there exists a sequence in $\mathcal{A}$, no entry of which equals $\mathbf{y}$, converging to $\mathbf{y}$. Equivalently, $\mathbf{y}$ is a cluster point if $\mathcal{A}$ if every neighbourhood of $\mathbf{y}$ contains a point of $\mathcal{A}$ not equal to $\mathbf{y}$.

Lemma 10.18 *Let $\| \cdot \|$ denote both the ℓ_1-norm of a vector in $\mathbb{R}^{m^k}$ as well as the corresponding induced norm on the set of $m^k \times m^k$ matrices. Suppose the process $\{\mathcal{Y}_t\}$ is ultra-mixing. Define*

$$\mathbf{b}_U := \mathbf{e}^t_{m^k} U \in \mathbb{R}^{1 \times r}.$$

Then

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} - \frac{1}{f_{\mathbf{u}}} \mathbf{p}_{\mathbf{u}} \mathbf{b}_U D^{(\mathbf{u})} \| \leq r \delta_{|\mathbf{u}|} \| F^{-1}_{I,J} \|, \tag{10.62}$$

where $\{\delta_l\}$ is the sequence in the definition of the ultra-mixing property, and $|\mathbf{u}|$ denotes the length of the string u.

Proof: If we substitute $\mathbf{j}$ for $\mathbf{v}$ in (10.61), we get

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{uj}}} \right| \leq \delta_{|\mathbf{u}|}.$$

For each $\mathbf{j} \in J$, we have that $f_{\mathbf{uj}}/f_{\mathbf{u}} \leq 1$. Hence we can multiply both sides of the above equation by $f_{\mathbf{uj}}/f_{\mathbf{u}} \leq 1$, which gives

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{uj}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} \right| = \left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{u}}} \right| \leq \delta_{|\mathbf{u}|} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} \leq \delta_{|\mathbf{u}|}.$$

Now define the $r \times r$ matrix $R^{(\mathbf{u})}$ by

$$(R^{(\mathbf{u})})_{\mathbf{ij}} := \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{u}}}.$$

Then (see for example [109])

$$\| R^{(\mathbf{u})} \| = \max_{\mathbf{j} \in \mathcal{M}^k} \sum_{\mathbf{i} \in \mathcal{M}^k} |(R^{(\mathbf{u})})_{\mathbf{ij}}| \leq r \delta_{|\mathbf{u}|}.$$

Next, note that

$$R^{(\mathbf{u})} = \mathbf{p}_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} - \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} F_{I,J}.$$

Hence we have established that

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} F_{I,J} - \mathbf{p}_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} \| \leq r \delta_{|\mathbf{u}|}. \tag{10.63}$$

Therefore

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} - \mathbf{p}_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} F^{-1}_{I,J} \| \leq r \delta_{|\mathbf{u}|} \| F^{-1}_{I,J} \|.$$

Thus the proof is complete once it is shown that

$$\mathbf{q}_{\mathbf{u}} F^{-1}_{I,J} = \frac{1}{f_{\mathbf{u}}} \mathbf{b}_U D^{(\mathbf{u})}.$$

But this last step is immediate, because

$$f_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} F^{-1}_{I,J} = F^{(\mathbf{u})}_{0,J} F^{-1}_{I,J} = \mathbf{e}^t_{m^k} U F^{(\mathbf{u})}_{I,J} F^{-1}_{I,J} == \mathbf{e}^t_{m^k} U D^{(\mathbf{u})} F_{I,J} F^{-1}_{I,J} = \mathbf{b}_U D^{(\mathbf{u})}.$$

This completes the proof.

The reader may wonder about the presence of the factor $1/f_{\mathbf{u}}$ in (10.62). Obviously, in any reasonable stochastic process, the probability $f_{\mathbf{u}}$ approaches zero as $|\mathbf{u}| \to \infty$. Hence, unless we divide by this quantity, we would get an inequality that is trivially true because both quantities individually approach zero. In contrast, (10.63) shows that the matrix $(1/f_{\mathbf{u}})D^{(\mathbf{u})}$ is both bounded and bounded away from zero for all $\mathbf{u} \in \mathcal{M}^*$.

Thus Lemma 10.18 serves to establish the behaviour of the matrix $D^{(\mathbf{u})}$ as $|\mathbf{u}| \to \infty$. Whatever be the vector $\mathbf{x} \in \mathbb{R}^r$, the vector $(1/f_{\mathbf{u}})D^{(\mathbf{u})}\mathbf{x}$ approaches $(1/f_{\mathbf{u}})\mathbf{p}_{\mathbf{u}}\mathbf{b}_U D^{(\mathbf{u})}\mathbf{x}$ and thus eventually gets 'aligned' with the vector $\mathbf{p}_{\mathbf{u}}$ as $|\mathbf{u}| \to \infty$.

Lemma 10.19 *Suppose the process under study is ultra-mixing, and that the cluster points of the probability vector set $\mathcal{A}$ are finite in number and belong to the interior of the cone $\mathcal{C}_c$. Then there exists a polyhedral cone $\mathcal{P}$ such that*

1. *$\mathcal{P}$ is invariant under each $D^{(u)}, \mathbf{u} \in \mathcal{M}$.*

2. *$\mathcal{C}_c \subseteq \mathcal{P} \subseteq \mathcal{C}_o$.*

3. *$\phi \in \mathcal{P}$.*

4. *$\theta^t \in \mathcal{P}^p$.*

Remark: In some sense this is the key lemma in the proof of the main theorem. It is noteworthy that the hypotheses do *not* include the assumption that the process under study is α-mixing.

Proof: First, note that, given any $\epsilon > 0$, there exists an $L = L(\epsilon)$ such that the following is true: For each $\mathbf{w} \in \mathcal{M}^*$ with $|\mathbf{w}| > L$, write $\mathbf{w} = \mathbf{uv}$ with $|\mathbf{u}| = L$. Then $\|\mathbf{p}_{\mathbf{w}} - \mathbf{p}_{\mathbf{u}}\| \leq \epsilon$. To see this, given $\epsilon > 0$, choose L such that $\delta_L \leq \epsilon/m^k$. Then (10.61) implies that $\|\mathbf{p}_{\mathbf{u}} - \mathbf{p}_{\mathbf{w}}\| \leq \epsilon$.

By assumption, the set of probability vectors $\mathcal{A} := \{\mathbf{p}_u : \mathbf{u} \in \mathcal{M}^k\}$ has only finitely many cluster points. Let us denote them as $\mathbf{x}_1, \ldots, \mathbf{x}_n$. By assumption again, each of these vectors lies in the interior of $\mathcal{C}_o$. Hence there exists an $\epsilon > 0$ such that the sphere (in the ℓ_1-norm) centered at each $\mathbf{x}_i$ of radius 2ϵ is also contained in $\mathcal{C}_o$.

Next, note that there exists an integer L such that *every* vector $\mathbf{p}_{\mathbf{u}}$ with $|\mathbf{u}| \geq L$ lies within a distance of ϵ (in the ℓ_1-norm) from at least one of the $\mathbf{x}_i$. In other words, there exists an integer L such that

$$\min_{1 \leq i \leq n} \|\mathbf{p}_{\mathbf{u}} - \mathbf{x}_i\| \leq \epsilon, \ \forall \mathbf{u} \in \mathcal{M}^l \text{ with } l > L.$$

To see why this must be so, assume the contrary. Thus there exists a sequence $\mathbf{p}_{\mathbf{u}_j}$ such that $\|\mathbf{p}_{\mathbf{u}_j} - \mathbf{x}_i\| > \epsilon$ for all i, j. Now the sequence $\{\mathbf{p}_{\mathbf{u}_j}\}$ is bounded and therefore has a convergent subsequence. The limit of this convergent subsequence cannot be any of the $\mathbf{x}_i$ by the assumption that $\|\mathbf{p}_{\mathbf{u}_j} - \mathbf{x}_i\| > \epsilon$ for all i, j. This violates the earlier assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *all* the cluster points of the set $\mathcal{A}$.

Now choose a set $\mathbf{z}_1, \ldots, \mathbf{z}_r$ of basis vectors for $\mathbb{R}^r$ such that each $\mathbf{z}_j$ has unit norm. For instance, we can take $\mathbf{z}_j$ to be the unit vector with a 1 in position j and zeros elsewhere. With ϵ already defined above, define the unit vectors

$$\mathbf{y}_{i,j}^{+} := \frac{\mathbf{x}_i + 2\epsilon \mathbf{z}_j}{\|\mathbf{x}_i + 2\epsilon \mathbf{z}_j\|},\ \mathbf{y}_{i,j}^{-} := \frac{\mathbf{x}_i - 2\epsilon \mathbf{z}_j}{\|\mathbf{x}_i - 2\epsilon \mathbf{z}_j\|},\ 1 \le i \le n,\ 1 \le j \le s.$$

With this definition, it is clear that *every* vector in the ball of radius 2ϵ centered at each $\mathbf{x}_i$ can be written as a nonnegative combination of the set of vectors $\{\mathbf{y}_{i,j}^{+}, \mathbf{y}_{i,j}^{-}\}$.

Now define the cone

$$\mathcal{B} := \text{Cone}\{\mathbf{y}_{i,j}^{+}, \mathbf{y}_{i,j}^{-}\}.$$

We begin by observing that $\mathbf{p}_{\mathbf{u}} \in \mathcal{B}$ whenever $|\mathbf{u}| \ge L$. This is because each such $\mathbf{p}_{\mathbf{u}}$ lies within a distance of ϵ from one of the $\mathbf{x}_i$ whenever $|\mathbf{u}| \ge L$. In particular, $\mathbf{p}_{\mathbf{u}} \in \mathcal{B}$ whenever $|\mathbf{u}| = L$. Moreover, by (10.1) and (10.2), every $\mathbf{p}_{\mathbf{v}}$ with $|\mathbf{v}| < L$ is a nonnegative combination of $\mathbf{p}_{\mathbf{u}}$ with $|\mathbf{u}| = L$. To see this, let $s := L - |\mathbf{v}|$, and note that

$$f_{\mathbf{v}} \mathbf{p}_{\mathbf{v}} = F_{I,0}^{(\mathbf{v})} = \sum_{\mathbf{w} \in \mathcal{M}^s} F_{I,0}^{(\mathbf{vw})},$$

and each vector $F_{I,0}^{(\mathbf{vw})}$ belongs to $\mathcal{B}$. Hence $\mathbf{p}_{\mathbf{u}} \in \mathcal{B}$ whenever $|\mathbf{u}| < L$. Combining all this shows that $\mathbf{p}_{\mathbf{u}} \in \mathcal{B}$ for all $\mathbf{u} \in \mathcal{M}^*$. As a result, it follows that $\mathcal{C}_c \subseteq \mathcal{B}$.

While the cone $\mathcal{B}$ is polyhedral, it is not necessarily invariant under each $D^{(u)}$. For the purpose of constructing such an invariant cone, it is now shown that $\mathcal{B}$ is invariant under each $D^{(\mathbf{u})}$ whenever $|\mathbf{u}|$ is sufficiently long. By Lemma 10.18, it follows that for every vector $\mathbf{y}$, the vector $(1/f_{\mathbf{u}})D^{(\mathbf{u})}\mathbf{y}$ gets 'aligned' with $p_{\mathbf{u}}$ as $|\mathbf{u}|$ becomes large. Therefore it is possible to choose an integer s such that

$$\|\frac{\|\mathbf{p}_{\mathbf{u}}\|}{\|D^{(\mathbf{u})}\mathbf{y}\|} D^{(\mathbf{u})}\mathbf{y} - \mathbf{p}_{\mathbf{u}}\| \le \epsilon \text{ whenever } |\mathbf{u}| \ge s,$$

whenever $\mathbf{y}$ equals one of the $2nr$ vectors $\mathbf{y}_{i,j}^{+}, \mathbf{y}_{i,j}^{-}$. Without loss of generality it may be assumed that $s \ge L$. In particular, the vectors $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{+}$ and $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{-}$, after normalization, are all within a distance of ϵ from $\mathbf{p}_{\mathbf{u}}$, which in turn is within a distance of ϵ from some $\mathbf{x}_t$. By the triangle inequality, this implies that the normalized vectors corresponding to $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{+}$ and $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{-}$ are all within a distance of 2ϵ from some $\mathbf{x}_t$, and hence belong to $\mathcal{B}$. In other words, we have shown that

$$D^{(\mathbf{u})}\mathcal{B} \subseteq \mathcal{B}\ \forall \mathbf{u} \text{ with } |\mathbf{u}| \ge s.$$

Now we are in a position to construct the desired polyhedral cone $\mathcal{P}$. Define

$$\mathcal{B}_i := \{D^{(\mathbf{u})}\mathcal{B} : |\mathbf{u}| = i\}, 1 \le i \le s - 1.$$

Thus $\mathcal{B}_i$ is the set obtained by multiplying each vector in $\mathcal{B}$ by a matrix of the form $D^{(\mathbf{u})}$ where $\mathbf{u}$ has length precisely i. It is easy to see that, since $\mathcal{B}$ is polyhedral, so is each $\mathcal{B}_i$. Now define

$$\mathcal{P} := \text{Cone}\{\mathcal{B}, \mathcal{B}_1, \ldots, \mathcal{B}_{s-1}\}.$$

For this cone, we establish in turn each of the four claimed properties.

Property 1: By definition we have that $D^{(u)}\mathcal{B}_i \subseteq \mathcal{B}_{i+1} \; \forall u \in \mathcal{M}$, whenever $0 \leq i \leq s-2$, and we take $\mathcal{B}_0 = \mathcal{B}$. On the other hand, $D^{(u)}\mathcal{B}_{s-1} \subseteq \mathcal{B}$ as has already been shown. Hence $\mathcal{P}$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$.

Property 2: We have already seen that $\mathbf{p_u} \in \mathcal{B}$ for all $\mathbf{u} \in \mathcal{M}^*$. Hence $\mathcal{C}_c = \text{Cone}\{\mathbf{p_u} : \mathbf{u} \in \mathcal{M}^*\} \subseteq \mathcal{B} \subseteq \mathcal{P}$. To prove the other containment, note that by assumption, the sphere of radius 2ϵ centered at each cluster point $\mathbf{x}_i$ is contained in $\mathcal{C}_o$. Hence $\mathcal{B} \subseteq \mathcal{C}_o$. Moreover, $\mathcal{C}_o$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$. Hence $\mathcal{B}_i \subseteq \mathcal{C}_o$ for each $i \in \{1, \ldots, s-1\}$. Finally $\mathcal{P} \in \mathcal{C}_o$.

Property 3: Note that each $\mathbf{p_u}$ belongs to $\mathcal{B}$, which is in turn a subset of $\mathcal{P}$. In particular, $\phi = \mathbf{p}_\emptyset \in \mathcal{P}$.

Property 4: Since $\mathcal{P} \subseteq \mathcal{C}_o$, it follows that $(\mathcal{P})^p \supseteq (\mathcal{C}_o)^p$. Hence it is enough to show that $\theta^t \in (\mathcal{C}_o)^p$. But this is easy to establish. Let $\mathbf{y} \in \mathcal{C}_o$ be arbitrary. Then by the definition of $\mathcal{C}_o$ we have that

$$\theta D^{(\mathbf{u})}\mathbf{y} \geq 0 \; \forall \mathbf{u} \in \mathcal{M}^* \; \forall \mathbf{y} \in \mathcal{C}_o.$$

In particular, by taking $\mathbf{u}$ to be the empty string (leading to $D^{(\mathbf{u})} = I$), it follows that $\theta\mathbf{y} \geq 0 \; \forall \mathbf{y} \in \mathcal{C}_o$. Since $\mathbf{y}$ is arbitrary, this shows that $\theta^t \in (\mathcal{C}_o)^p$.

Proof of Theorem 10.17: The proof of the main theorem closely follows the material in [6], pp. 117-119. Let us 'recycle' the notation and let $\mathbf{y}_1, \ldots, \mathbf{y}_s$ denote generators of the polyhedral cone $\mathcal{P}$. In other words, $\mathcal{P}$ consists of all nonnegative combinations of the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_s$. Note that neither the integer s nor the generators need be uniquely defined, but this does not matter. Define the matrix

$$Y := [\mathbf{y}_1 | \ldots | \mathbf{y}_s] \in \mathbb{R}^{m^k \times s}.$$

Then it is easy to see that

$$\mathcal{P} = \{Y\mathbf{x} : \mathbf{x} \in \mathbb{R}^s_+\}.$$

Now we can reinterpret the four properties of Lemma 10.19 in terms of this matrix. Actually we need not bother about Property 2.

Property 1: Since $\mathcal{P}$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$, it follows that each $D^{(u)}\mathbf{y}_i$ is a nonnegative combination of $\mathbf{y}_1, \ldots, \mathbf{y}_s$. Hence there exist nonnegative matrices $G^{(u)} \in \mathbb{R}^{s \times m^k}_+, u \in \mathcal{M}$ such that

$$D^{(u)}Y = YG^{(u)}, \; \forall u \in \mathcal{M}.$$

Property 3: Since $\phi \in \mathcal{P}$, there exists a nonnegative vector $\mathbf{z} \in \mathbb{R}^s_+$ such that

$$\phi = Y\mathbf{z}.$$

Property 4: Since $\theta \in \mathcal{P}^p$, we have in particular that $\theta\mathbf{y}_i \geq 0$ for all i. Hence

$$\mathbf{h} := \theta Y \in \mathbb{R}^s_+.$$

Moreover, $\mathbf{h} \neq \mathbf{0}$, because $\theta\phi = \mathbf{hz} = 1$, the frequency of the empty string.

With these observations, we can rewrite the expression for the frequency of an arbitrary string $\mathbf{u} \in \mathcal{M}^*$. We have

$$\begin{aligned} f_{\mathbf{u}} &= \theta D^{(u_1)} \cdots D^{(u_l)} \phi \\ &= \theta D^{(u_1)} \cdots D^{(u_l)} Y \mathbf{z} \\ &= \theta D^{(u_1)} \cdots D^{(u_{l-1})} Y G^{(u_l)} \mathbf{z} = \cdots \\ &= \theta Y G^{(u_1)} \cdots G^{(u_l)} \mathbf{z} \\ &= \mathbf{h} G^{(u_1)} \cdots G^{(u_l)} \mathbf{z} \end{aligned} \tag{10.64}$$

The formula (10.64) is similar in appearance to (10.34), but with one very important difference: *Every matrix and vector in (10.64) is nonnegative.* Therefore, in order to construct an irreducible HMM from the above formula, we need to ensure that the matrix $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is irreducible and row stochastic, that $\mathbf{h}$ satisfies $\mathbf{h} = \mathbf{h}Q$, and that $\mathbf{z} = \mathbf{e}_s$. This is achieved through a set of three reductions. Note that these reductions are the same as in [6], pp. 117-119.

Now for the first time we invoke the assumption that the process $\{\mathcal{Y}_t\}$ is α-mixing. From Theorem 10.10, this assumption implies that the matrix $S = \sum_{u \in \mathcal{M}} D^{(u)}$ has the ‘strong Perron property,’ namely: The spectral radius of S is one, and one is an eigenvalue of S; moreover, if λ is any eigenvalue of S besides one, then $|\lambda| < 1$. We also know that ϕ and θ are respectively a column eigenvector and a row eigenvector of S corresponding to the eigenvalue one.

Now let us return to the formula (10.64). Define $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ as before. Observe that Q is a nonnegative matrix; hence, by [17], Theorem 1.3.2, p. 6, it follows that the spectral radius $\rho(Q)$ is also an eigenvalue. Moreover, $\rho(Q)$ is at least equal to one, because

$$\mathbf{h}Q = \theta \sum_{u \in \mathcal{M}} Y G^{(u)} = \theta \left(\sum_{u \in \mathcal{M}} D^{(u)} \right) Y = \theta Y = \mathbf{h}.$$

Here we make use of the fact that θ is a row eigenvector of $\sum_{u \in \mathcal{M}} D^{(u)}$ corresponding to the eigenvalue one.

In what follows, we cycle through three steps in order to arrive at a situation where Q is irreducible and row stochastic. In each step we will be replacing the various matrices by other, smaller matrices that play the same role. To avoid notational clutter, the old and new matrices are denoted by the same symbols.

Step 1: If Q is irreducible, go to Step 3. If Q is reducible, permute rows and columns if necessary and partition Q as

$$Q = \left[\begin{array}{cc} Q_{11} & Q_{12} \\ \mathbf{0} & Q_{22} \end{array} \right],$$

where Q_{11} is irreducible and has dimension $(s-l) \times (s-l)$, and Q_{22} has dimension $l \times l$ for some $l < s$. (It is not assumed that Q_{22} is irreducible, since an irreducible partition of Q may have more than two ‘blocks.’) Since

$Q = \sum_{u \in \mathcal{M}} G^{(u)}$ and each $G^{(u)}$ is nonnegative, if we partition each $G^{(u)}$ commensurately, then the block zero structure of Q will be reflected in each $G^{(u)}$. Now there are two possibilities: Either $\rho(Q_{11}) = 1$, or it is not. If $\rho(Q_{11}) = 1$, go to Step 2. If $\rho(Q_{11}) \neq 1$, proceed as follows: Let $\lambda_1 = \rho(Q_{11}) \neq 1$. Choose a positive vector $\mathbf{x}_1 \in \mathbb{R}_+^{s-l}$ such that $Q_{11}\mathbf{x}_1 = \lambda_1 \mathbf{x}_1$. (Note that, by [17], Theorem 2.2.10, p. 30, it is possible to choose a strictly positive eigenvector of Q_{11} corresponding to the eigenvalue $\rho(Q_{11})$, since Q_{11} is irreducible.) Then clearly $Q\mathbf{x} = \lambda_1 \mathbf{x}$, where $\mathbf{x} = [\mathbf{x}_1^t \;\; \mathbf{0}^t]^t$. Since $\lambda_1 \neq 1$, it follows that $\mathbf{hx} = 0$. (Recall that a row eigenvector and a column eigenvector corresponding to different eigenvalues are orthogonal.) So if we partition $\mathbf{h}$ as $[\mathbf{h}_1 \;\; \mathbf{h}_2]$, then $\mathbf{h}_1 = \mathbf{0}$ since $\mathbf{x}_1$ is a positive vector. Now observe that each $G^{(u)}$ has the same block-triangular structure as Q. Hence, by a slight abuse of notation, let us define, for every string $\mathbf{u} \in \mathcal{M}^*$,

$$G^{(\mathbf{u})} = \begin{bmatrix} G_{11}^{(\mathbf{u})} & G_{12}^{(\mathbf{u})} \\ \mathbf{0} & G_{22}^{(\mathbf{u})} \end{bmatrix}.$$

Let us partition $\mathbf{z}$ commensurately. Because the first block of $\mathbf{h}$ is zero, it is easy to verify that, for every $\mathbf{u} \in \mathcal{M}^*$, we have

$$f_{\mathbf{u}} = \mathbf{h}G^{(\mathbf{u})}\mathbf{z} = \mathbf{h}_2 G_{22}^{(\mathbf{u})} \mathbf{z}_2,$$

where $\mathbf{z}_2$ consists of the last l components of $\mathbf{z}$. Hence we can partition Y as $[Y_1|Y_2]$ where $Y_2 \in \mathbb{R}^{r \times l}$ and make the following substitutions:

$$s \leftarrow l, Y \leftarrow Y_2, G^{(u)} \leftarrow G_{22}^{(u)} \; \forall u \in \mathcal{M}, \mathbf{h} \leftarrow \mathbf{h}_2, \mathbf{z} \leftarrow \mathbf{z}_2.$$

In this way, we have reduced the number of columns of Y from s to r, and (10.64) continues to hold. Now go back to Step 1.

Step 2: If we have reached this point, then Q is reducible, and if it is partitioned as above, we have $\rho(Q_{11}) = 1$. Choose a positive vector $\mathbf{x}_1$ such that $Q_{11} = \mathbf{x}_1$. Then $Q\mathbf{x} = \mathbf{x}$, where as before $\mathbf{x} = [\mathbf{x}_1^t \;\; \mathbf{0}^t]^t$. Next, note that

$$SY\mathbf{x} = \left(\sum_{u \in \mathcal{M}} D^{(u)} \right) Y\mathbf{x} = Y \left(\sum_{u \in \mathcal{M}} G^{(u)} \right) \mathbf{x} = YQ\mathbf{x} = Y\mathbf{x}.$$

Hence $Y\mathbf{x}$ is a column eigenvector of S corresponding to the eigenvalue one. However, from Theorem 10.10, the α-mixing property implies that S has a simple eigenvalue at one, with corresponding column eigenvector $\phi = F_{I,0}$. Hence $F_{I,0}$ equals $Y\mathbf{x}$ times some scale factor, which can be taken as one without loss of generality (since both vectors are nonnegative). Partition Y as $[Y_1 \;\; Y_2]$ where $Y_1 \in \mathbb{R}^{r \times (s-l)}$. Then

$$F_{I,0} = [Y_1 \;\; Y_2] \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{0} \end{bmatrix} = Y_1 \mathbf{x}_1.$$

Moreover, since each $G^{(u)}$ inherits the zero structure of Q, we have that

$$D^{(u)}[Y_1 \;\; Y_2] = [Y_1 \;\; Y_2] \begin{bmatrix} G_{11}^{(u)} & G_{12}^{(u)} \\ \mathbf{0} & G_{22}^{(u)} \end{bmatrix}.$$

In particular, we have that $D^{(u)} Y_1 = Y_1 G_{11}^{(u)}$. This means that $F_{I,0}$ lies in the cone generated by the columns of Y_1, and that this cone is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$. So if we define $\mathbf{h}_1 := \theta Y_1$, then because of the zero block in $\mathbf{x}$ it follows that

$$f_{\mathbf{u}} = \theta D^{(\mathbf{u})} F_{I,0} = \mathbf{h}_1 G_{11}^{(\mathbf{u})} \mathbf{x}_1.$$

So now we can make the substitutions

$$s \leftarrow s - l, Y \leftarrow Y_1, G^{(u)} \leftarrow G_{11}^{(\mathbf{u})}, \mathbf{h} \leftarrow \mathbf{h}_1, \mathbf{z} \leftarrow \mathbf{x}_1.$$

With these substitutions we have the relationship (10.64) continues to hold. In the process, the number of columns of Y has been reduced from s to $s-l$. Moreover, the resulting matrix Q is the old Q_{11}, which is irreducible. Now go to Step 3.

Step 3: When we reach this stage, (10.64) continues to hold, but with two crucial additional features: Q is irreducible and $\rho(Q) = 1$. As before, let s denote the size of the matrix Q, and write $\mathbf{z} = [z_1 \ldots z_s]^t$, where each z_i is positive. Define $Z = \text{Diag}\{z_1, \ldots, z_s\}$. Now (10.64) can be rewritten as

$$f_{\mathbf{u}} = \mathbf{h} Z Z^{-1} G^{(u_1)} Z \cdot Z^{-1} G^{(u_2)} Z \ldots Z^{-1} G^{(u_l)} Z \cdot Z^{-1} \mathbf{z}.$$

Thus (10.64) holds with the substitutions

$$G^{(u)} \leftarrow Z^{-1} G^{(u)} Z, \mathbf{h} \leftarrow \mathbf{h} Z, \mathbf{z} \leftarrow Z^{-1} \mathbf{z}.$$

In this process, Q gets replaced by $Z^{-1} Q Z$. Now observe that

$$Z^{-1} Q Z \mathbf{e}_s = Z^{-1} Q \mathbf{z} = Z^{-1} \mathbf{z} = \mathbf{e}_s.$$

In other words, the matrix $Z^{-1} Q Z$ is row stochastic. It is obviously nonnegative and irreducible. Moreover, we have that $\mathbf{hz} = 1$ since it is the frequency of the empty string, which by definition equals one. Hence the row vector $\mathbf{h} Z^{-1}$ is row stochastic in that its entries add up to one. Hence, after we make the substitutions, (10.64) holds with the additional properties that (i) $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is row-stochastic, (ii) $\mathbf{h}$ is row-stochastic and satisfies $\mathbf{h} = \mathbf{h} Q$, and (iii) $\mathbf{z} = \mathbf{e}_s$. Now it follows from Lemma 10.12 that the process $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM. Moreover, the matrix Q is irreducible.

Thus far it has been established that the stochastic process $\{\mathcal{Y}_t\}$ has an irreducible HMM. Moreover, this process is assumed to be α-mixing. So from Theorem 10.13, it finally follows that either the corresponding state transition matrix is aperiodic, or else the consistency conditions (10.60) hold.

Theorem 10.17 gives *sufficient* conditions for the existence of an irreducible HMM that satisfies some consistency conditions in addition. It is therefore natural to ask how close these sufficient conditions are to being necessary. The paper [6] also answers this question.

Theorem 10.20 *Given an irreducible HMM with n states and m outputs, define its period p. Rearrange the state transition matrix A as in Theorem 10.13, permute the matrices $M^{(u)}, u \in \mathcal{M}$ correspondingly, and define the* *blocks $M_i^{(u)}$ in analogy with the partition of A. Suppose in addition that there exists an index $q \leq s$ such that the following property holds: For every string $\mathbf{u} \in \mathcal{M}^q$ and every integer r between 1 and p, every column of the product $M_r^{(u_1)} M_{r+1}^{(u_2)} \ldots M_{r+q-1}^{(u_q)}$ is either zero or else is strictly positive. In this computation, any subscript M_i is replaced by i mod p if $i > p$. With this property, the HMM is α-mixing and also ultra-mixing.*

For a proof, see [6], Lemma 2.

Thus we see that there is in fact a very small gap between the sufficiency condition presented in Theorem 10.17 and the necessary condition discovered earlier in [6]. If the sufficient conditions of Theorem 10.17 are satisfied, then there exists an irreducible HMM that also satisfies the consistency conditions (10.60). Conversely, if an irreducible HMM satisfies the consistency conditions (10.60) and one other technical condition, then it satisfies three out of the four hypotheses of Theorem 10.17, the only exception being the technical condition about the cluster points lying in the interior of the cone $\mathcal{C}_c$.

We conclude this section by discussing the nature of the ‘technical’ conditions in the hypotheses of Theorems 10.17 and 10.20. The idea is to show that, in a suitably defined topology, each of the conditions is satisfied by an ‘open dense subset’ of stochastic processes. Thus, if the given process satisfies the condition, so does any sufficiently small perturbation of it, whereas if a given process fails to satisfy the condition, an arbitrarily small perturbation will cause the condition to hold.

Let us begin with the fourth hypothesis of Theorem 10.13. We follow [85] and define a topology on the set of all stationary stochastic processes assuming values in $\mathcal{M}$. Suppose we are given two stochastic processes assuming values in a common finite alphabet $\mathcal{M}$. Let $f_{\mathbf{u}}, g_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*$ denote the frequency vectors of the two stochastic processes. This is equivalent to specifying the joint distribution of l-tuples of each stochastic process, for every integer l. If we arrange all strings $\mathbf{u} \in \mathcal{M}^*$ in some appropriate lexical ordering (say first lexical), then each of $[f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*]$, $[g_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*]$ is a vector with a countable number of components, and each component lies between 0 and 1.[6] Let the symbols $\mathbf{f}, \mathbf{g}$, without any subscript, denote these vectors belonging to ℓ_∞. We might be tempted to compare the two stochastic processes by computing the norm $\|\mathbf{f} - \mathbf{g}\|_\infty$. The difficulty with this approach is that, as the length of the string $\mathbf{u}$ approaches infinity, the likelihood of that sequence will in general approach zero. Thus, in any ‘reasonable’ stochastic process, the difference $f_{\mathbf{u}} - g_{\mathbf{u}}$ will approach zero as $|\mathbf{u}| \to \infty$, but this tells us nothing about how close the two probability laws are. To get around this difficulty, for each $\mathbf{u} \in \mathcal{M}^*$, we define the vector $\mathbf{p}_{|\mathbf{u}} \in [0,1]^m$ as follows:

$$\mathbf{p}_{|\mathbf{u}} = \frac{1}{f_{\mathbf{u}}} \mathbf{f}_{\mathbf{u}v, v \in \mathcal{M}} = \left[\frac{f_{\mathbf{u}v}}{f_{\mathbf{u}}}, v \in \mathcal{M} \right].$$

[6] Note that there is a lot of redundancy in this description of a stochastic process because, as we have already seen, the joint distribution of l-tuples can be uniquely determined from the joint distribution of s-tuples if $s > l$.

Thus $\mathbf{p}_{|\mathbf{u}}$ is just the conditional distribution of the *next* symbol, given the past history $\mathbf{u}$. The advantage of $\mathbf{p}_{|\mathbf{u}}$ is that, even as $|\mathbf{u}|$ becomes large, the elements of this vector must still add up to one, and as a result they cannot all go to zero. With this convention, let us list all strings $\mathbf{u} \in \mathcal{M}^*$ in some appropriate lexical ordering (say first lexical), and for each $\mathbf{u}$ let us define the conditional distribution vectors $\mathbf{p}_{|\mathbf{u}}$ corresponding to $\{f_{\mathbf{u}}\}$, and the conditional distribution vectors $\mathbf{q}_{|\mathbf{u}}$ corresponding to the vector $\{\mathbf{g}_{\mathbf{u}}\}$. Finally, let us define the vectors

$$\tilde{\mathbf{p}} := [\mathbf{p}_{|\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*], \tilde{\mathbf{q}} := [\mathbf{q}_{|\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*].$$

Thus both $\tilde{\mathbf{p}}, \tilde{\mathbf{q}}$ have a countable number of components, since $\mathcal{M}^*$ is a countable set. Thus the ℓ_∞ norm of the difference $\tilde{\mathbf{p}} - \tilde{\mathbf{q}}$ is a measure of the disparity between the two stochastic processes. This is essentially the distance measure introduced in [85]. With this measure, it is easy to see that the fourth hypothesis of Theorem 10.13 is truly technical: If a given stochastic process satisfies the condition about the cluster points, then so will any sufficiently small perturbation of it, while if a given stochastic process fails to satisfy this condition, any sufficiently small perturbation of it will cause the condition to be satisfied.

Now let us turn to the condition in Theorem 10.20. Given two HMMs over a common state space, a natural metric is

$$\sum_{u \in \mathcal{M}} \|M_1^{(u)} - M_2^{(u)}\|,$$

where $\| \cdot \|$ is any reasonable matrix norm. Again, it is easy to see that the condition in Theorem 10.20 about the various columns being either identically zero or strictly positive is 'technical.' In fact, if for a HMM some elements of the matrices $M_r^{(u_1)} M_{r+1}^{(u_2)} \ldots M_{r+q-1}^{(u_q)}$ are zero, then by simply making an arbitrarily small perturbation in the matrices we can ensure that every entry is strictly positive.

Chapter Eleven

Applications to Computational Biology Problems

11.1 GENE-FINDING

11.2 PROTEIN CLASSIFICATION

Bibliography

[1] N. Alon, M. Krivelevich, I. Newman and M. Szegedy, "Regular languages are testable with a constant number of queries," *Foundations of Computer Sciences*, IEEE, New York, 645-655, 1999.

[2] N. Alon, M. Krivelevich, I. Newman and M. Szegedy, "Regular languages are testable with a constant number of queries," *SIAM J. Computing*, 30, 1842-1862, 2001.

[3] S. F. Altschul, "The statistics of similarity sequence scores," found at http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html .

[4] S. F. Altschul, W. Gish, W. Q. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, 215, 403-410, 1990.

[5] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Q. Miller and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, 25, 3389-3401, 1997.

[6] B. D. O. Anderson, "The realization problem for hidden Markov models," *Mathematics of Control, Signals, and Systems*, 12(1), 80-120, 1999.

[7] B. D. O. Anderson, "From Wiener to hidden Markov models," *IEEE Control Systems Magazine*, 19(6), 41-51, June 1999.

[8] B. D. O. Anderson, M. Deistler, L. Farina and L. Benvenuti, "Nonnegative realization of a system with a nonnegative impulse response," *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 43, 134-142, 1996.

[9] L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Info. Thy.*, 20(2), 284-287, March 1974.

[10] L. R. Bahl and F. Jelinek, "Decoding for channels with insertions, deletions and substitutions with applications to speech recognition," *IEEE Trans. Info. Thy.*, 21(4), 404-411, July 1975.

[11] Y. Bakhtin and C. E. Heitsch, "Large deviations for random trees and the branching of RNA secondary structures," *Bull. Math. Biol.*, 71(1), 84-106, 2009.

[12] P. Baldi and S. Brunak, *Bioinformatics: A Machine Learning Approach*, (Second Edition), MIT Press, Cambridge, MA, 2001.

[13] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, 37, 1554-1563, 1966.

[14] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, 41(1), 164-171, 1970.

[15] R. Begleiter, R. El-Yaniv and G. Yona, "On prediction using variable order Markov models," *Journal of Artificial Intelligence Research*, 22, 385-421, 2004.

[16] L. Benvenuti and L. Farina, "A tutorial on the positive realization problem," *IEEE Transactions on Automatic Control*, 49, 651-664, 2004.

[17] A. Berman and R. J. Plemmons, *Nonnegative Matrices*, Academic Press, New York, 1979.

[18] P. Billingsley, *Probability and Measure*, Wiley, New York, 1986.

[19] P. Billingsley, *Probability and Measure*, (Third Edition), Wiley, New York, 1995.

[20] D. Blackwell and L. Koopmans, "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Stat.*, 28, 1011-1015, 1957.

[21] V. Blondel and V. Catarini, "Undecidable problems for probabilistic automata of fixed dimension," *Theory of Computing Systems*, **36**, 231-245, 2003.

[22] L. Breiman, *Probability*, SIAM Classics in Applied Mathematics, No. 7, SIAM, Philadelphia, 1992.

[23] C. Burge and S. Karlin, "Prediction of complete gene structure in human genomic DNA," *J. Mol. Biol.*, 268, 78-94, 1997.

[24] J. W. Carlyle, “Identification of state-calculable functions of finite Markov chains,” *Annals of Mathematical Statistics*, 38, 201-205, 1967.

[25] J. W. Carlyle, “Stochastic finite-state system theory,” in *System Theory*, L. Zadeh and E. Polak (Eds.), Chapter 10, McGraw-Hill, New York, 1969.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Second Edition), Wiley Interscience, New York, 2006.

[27] F. Crick, *What Mad Pursuit*, Basic Books, 1988.

[28] I. Csiszár, “The method of types,” *IEEE Trans. Info. Thy.*, **44(6)**, 2505-2523, 1998.

[29] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[30] A. L. Delcher, D. Harmon, S. Kasif, O. White and S. L. Salzberg, “Improved microbial gene identification with GLIMMER,” *Nucleic Acids Research*, 27(23), 4636-4641, 1999.

[31] A. Dembo and S. Karlin, “Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables”, *Annals of Prob.*, 19(4), 1737-1755, 1991.

[32] A. Dembo and S. Karlin, “Strong limit theorems of empirical functionals for large exceedances of partial sums of Markov variables”, *Annals of Prob.*, 19(4), 1756-1767, 1991.

[33] A. Dembo, S. Karlin and O. Zeitouni, “Critical phenomena for sequence matching with scoring,” Annals of Prob., 22(4), 1993-2021, 1994.

[34] A. Dembo, S. Karlin and O. Zeitouni, “Limit distribution of maximal non-aligned two-sequence segmental score,” Annals of Prob., 22(4), 2022-2039, 1994.

[35] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Springer-Verlag, Berlin, 1998.

[36] L. Devroye, L. Gyorfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[37] S. W. Dharmadhikari, “Functions of finite Markov chains,” *Annals of Mathematical Statistics*, 34, 1022-1031, 1963.

[38] S. W. Dharmadhikari, "Sufficient conditions for a stationary process to be a function of a Markov chain," *Annals of Mathematical Statistics*, 34, 1033-1041, 1963.

[39] S. W. Dharmadhikari, "A characterization of a class of functions of finite Markov chains," *Annals of Mathematical Statistics*, 36, 524-528, 1965.

[40] S. W. Dharmadhikari, "A note on exchangeable processes with states of finite rank," *Annals of Mathematical Statistics*, 40(6), 2207-2208, 1969.

[41] S. W. Dharmadhikari and M. G. Nadkarni, "Some regular and non-regular functions of finite Markov chains," *Annals of Mathematical Statistics*, 41(1), 207-213, 1970.

[42] J. Dixmier, "Proof of a conjecture by Erdös and Graham concerning the problem of Frobenius," *J. Number Thy.*, 34, 198-209, 1990.

[43] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Info. Thy.*, 48(6), 1518-1569, June 2002.

[44] R. V. Erickson, "Functions of Markov chains," *Annals of Mathematical Statistics*, 41, 843-850, 1970.

[45] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics*, Springer-Verlag, New York, 2001.

[46] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics*, Second Edition, Springer-Verlag, New York, 2006.

[47] M. Fliess, "Series rationelles positives et processus stochastique," *Annales de l'Institut Henri Poincaré, Section B*, XI, 1-21, 1975.

[48] J. Feng and T. G. Kurtz, *Large Deviations for Stochastic Processes*, Amer. Math. Soc., Providence, RI, 2006.

[49] M. Fox and H. Rubin, "Functions of processes with Markovian states," *Annals of Mathematical Statistics*, 39, 938-946, 1968.

[50] G. Frobenius, "Über Matrizen aus positiven Elementen," *S.-B. Preuss. Akad. Wiss. (Berlin)*, 471-476, 1908.

[51] G. Frobenius, "Über Matrizen aus positiven Elementen," *S.-B. Preuss. Akad. Wiss. (Berlin)*, 514-518, 1909.

[52] G. Frobenius, "Über Matrizen aus nicht negativen Elementen," *S.-B. Preuss. Akad. Wiss. (Berlin)*, 456-477, 1912.

[53] E. J. Gilbert, "The identifiability problem for functions of Markov chains," *Annals of Mathematical Statistics*, 30, 688-697, 1959.

[54] D. Gusfield, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, Cambridge, UK, 1997.

[55] A. Heller, "On stochastic processes derived from Markov chains," *Annals of Mathematics*, 36, 1286-1291, 1965.

[56] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.* 58, 13-30, 1963.

[57] J. M. van den Hof, "Realization of continuous-time positive linear systems," *Systems and Control Letters*, 31, 243-253, 1997.

[58] J. M. van den Hof and J. H. van Schuppen, "Realization of positive linear systems using polyhedral cones," *Proceedings of the 33rd IEEE Conference on Decision and Control*, 3889-3893, 1994.

[59] F. Den Hollander, *Large Deviations*, Amer. Math. Soc., Providence, RI, 2000.

[60] International Human Genome Research Consortium, "Initial sequencing and analysis of the human genome," *Nature*, 409, 860-921, February 2001.

[61] H. Ito, S. Amari and K. Kobayashi, "Identifiability of hidden Markov information sources and their minimum degrees of freedom," *IEEE Transactions on Information Theory*, 38, 324-333, 1992.

[62] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.

[63] B-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.*, 64(2), 391-408, February 1985.

[64] S. Kalikow, "Random Markov processes and uniform martingales," *Israel Journal of Mathematics*, 71(1), 33-54, 1990.

[65] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc. Nat. Acad. Sci.*, 87, 2264-2268, 1990.

[66] S. Karlin and S. F. Altschul, "Applications and statistics for multiple high-scoring segments in molecular sequences," *Proc. Nat. Acad. Sci.*, 90, 5873-5877, 1993.

[67] S. Karlin and A. Dembo, "Limit distributions of maximal segmental score among Markov-dependent partial sums," *Adv. Appl. Prob.*, 24(1), 113-140, 1992.

[68] S. Karlin and F. Ost, "Maximal length of common words among random letter sequences," *Ann. Prob.*, 16, 535-563, 1988.

[69] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975.

[70] A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover, New York, 1957.

[71] A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea, New York, 1950.

[72] T. Koski, *Hidden Markov Models for Bioinformatics*, Kluwer, Dordrecht, The Netherlands, 2001.

[73] A. Krogh, M. Brown, I. S. Mian, K. Sjölander and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, 235, 1501-1531, 1994.

[74] A. Krogh, I. S. Mian and D. Haussler, "A hidden Markov model that finds genes in *E. coli* DNA," *Nucleic Acids Research*, 22(22), 4768-4778, 1994.

[75] L. Kronecker "Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen," *Monatsber. Königl. Preuss. Akad. Wiss. Berlin*, 535-600, 1881.

[76] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, 22, 79-86, 1951.

[77] G. Langdon, "A note on the Ziv-Lempel model for compressing individual data sequences," *IEEE Transactions on Information Theory*, 29, 284-287, 1983.

[78] M. Lewin, "A bound for a solution of a linear Diophantine problem," *J. London Math. Soc. (2)*, 6, 61-69, 1972.

[79] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Info. Thy.*, 52(10), 4394-4412, Oct. 2006.

[80] W. H. Majoros and S. L. Salzberg, "An empirical analysis of training protocols for probabilistic gene finders," *BMC Bioinformatics*, available at `http://www.biomedcentral.com/1471-2105/5/206`, 21 December 2004.

[81] P. Massart "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Ann. Probab.*, Vol. 18, No. 3, 1269-1283, 1990.

[82] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," *J. Mol. Biol.*, 48(3), 443-453, 1970.

[83] J. R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.

[84] G. Nuel, "LD-SPatt: Large deviations statistics for patterns on Markov chains," *J. Bomput. Biol.*, 11(6), 1023-33, 2004.

[85] D. S. Ornstein and B. Weiss, "How sampling reveals a process," *Ann. Probab.*, 18(3), 905-930, 1990.

[86] M. Pertea, X. Lin and S. L. Salzberg, "GeneSplicer: A new computational method for splice site detection," *Nucleic Acids Research*, 29(5), 1185-1190, 2001.

[87] O. Perron, "Zur Theorie der Matrizen," *Mathematische Annalen*, 64, 248-263, 1907.

[88] P. A. Pevzner, *Computational Molecular Biology*, MIT Press, Cambridge, MA, 2000.

[89] G. Picci, "On the internal structure of finite-state stochastic processes," in *Recent Developments in Variable Structure Systems*, R. Mohler and A. Ruberti (Eds.), Lecture Notes in Economics and Mathematical Systems, No. 162, Springer-Verlag, Heidelberg, 1978.

[90] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[91] L. W. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, 77(2), 257-285, February 1989.

[92] Z. Rached, F. Alalaji and L. L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Info. Thy.*, 50(5), 917-921, May 2004.

[93] R. T. Rockafellar, *Convexity.*

[94] G. Rozenberg and A. Salomaa, *Cornerstones in Undecidability*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[95] W. Rudin, *Real and Complex Analysis*, McGraw-Hill.

[96] S. L. Salzberg, A. L. Delcher, S. Kasif and O. White, "Microbial gene identification using interpolated Markov models," *Nucleic Acids Research*, 26(2), 544-548, 1998.

[97] S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner and H. Tettelin, "Interpolated Markov models for eukaryotic gene finding," *Genomics*, 59, 24-31, 1999.

[98] A. Sayre, *Rosalind Franklin and DNA*, Norton, New York, 2000 (reissue of 1975 edition).

[99] E. Seneta, *Non-negative Matrices and Markov Chains*, (Second Edition), Springer-Verlag, Berlin, 1981.

[100] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Amer. Math. Soc., Providence, RI, 1996.

[101] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, 147, 195-197, 1981.

[102] E. D. Sontag, "On certain questions of rationality and decidability," *Journal of Computer and System Science*, 11, 375-381, 1975.

[103] D. W. Stroock, *An Introduction to Markov Processes*, Springer-Verlag, Berlin, 2005.

[104] F. Topsoe, "Bounds for entropy and divergence over a two-element set," *J. Inequalities in Pure and Appl. Math.*, 2(2), 2001.

[105] V. N. Vapnik, *Statistical Learning Theory*, John Wiley, New York, 1998.

[106] J. C. Venter, M. D. Adams, E. W. Myers et al., "The sequence of the human genome," *Science*, 291, 1304-1351, 16 February 2001.

[107] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag, London, 1997.

[108] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*, Springer-Verlag, London, 2003.

[109] M. Vidyasagar, *Nonlinear Systems Analysis*, SIAM Publications, Philadelphia, PA, 2003.

[110] M. Vidyasagar, "Convergence of empirical means with alpha-mixing input sequences, and an application to PAC learning," *Proc. IEEE Conf. on Decision and Control*, Seville, Spain, 2005.

[111] M. Vidyasagar, “The realization problem for hidden Markov models: The complete realization problem,” *Proc. IEEE Conf. on Decision and Control*, Seville, Spain, 2005.

[112] M. Vidyasagar, “The 4M (Mixed Memory Markov Model) algorithm for stochastic modelling over a finite alphabet with applications to gene-finding algorithms,” *Proc. Symp. Math. Thy. Netw. and Sys.*, Kyoto, Japan, 443-459, July 2006.

[113] M. Vidyasagar, “A realization theory for hidden Markov models: The partial realization problem,” *Proc. Symp. Math. Thy. Netw. and Sys.*, Kyoto, Japan, 2145-2150, July 2006.

[114] M. Vidyasagar, “The 4M (Mixed Memory Markov Model) algorithm for finding genes from prokaryotic genomes,” *Proc. Conf. Decision and Control*, San Diego, CA, 1-6, Dec. 2006.

[115] J. D. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*, Touchstone, New York, 2001 (reprint of 1968 edition).

[116] T. A. Welch, “A technique for high performance data compression, *IEEE Computer*, 17(6), 8-19, 1984.

[117] H. Wielandt, “Unzerlegbare, nicht negativen Matrizen,” *Mathematische Zeitschrift*, 52, 642-648, 1950.

[118] M. Wilkins, *The Third Man of the Double Helix*, Oxford University Press, 2003.

[119] L. Xie, A. Ugrinovskii and I. R. Petersen, “Probabilistic distances between finite-state finite-alphabet hidden Markov models,” *IEEE Trans. Auto. Control*, **50(4)**, 505-511, April 2005.

[120] M. Q. Zhang, “Computational prediction of eukaryotic protein-coding genes,” *Nature Reviews*, 3, 698-710, September 2002.

[121] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Transactions on Information Theory*, 23(3), 337-343, 1977.

[122] J. Ziv and A. Lempel, “Compression of sequences via variable rate encoding,” *IEEE Transactions on Information Theory*, 24(5), 530-536, 1978.

[123] Figure 1.1 is from http://www.geocities.com/avinash_abhyankar/molecular/dna_files/image001.g

[124] Figure 1.2 is from http://wpcontent.answers.com/wikipedia/commons/thumb/e/e4/DNA_chemi
DNA_chemical_structure.svg.png

[125] Figure 1.3 is from http://www.bio.miami.edu/∼cmallery/150/gene/c16x6basepairs.jpg

[126] Figure 1.4 has been produced using the package Bio-Suite$^{\text{TM}}$.

[127] Figure 1.5 is from http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/

[128] Figure 1.6 is from http://plato.stanford.edu/entries/information-biological/GeneticCode.png

[129] Figure 1.7 is from http://genome.wellcome.ac.uk/assets/GEN10000676.jpg

[130] Figure 1.8 is from http://www.columbia.edu/cu/biology/courses/c2005/images/prot3imag.gif

[131] Figure 1.9 is from http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/qua