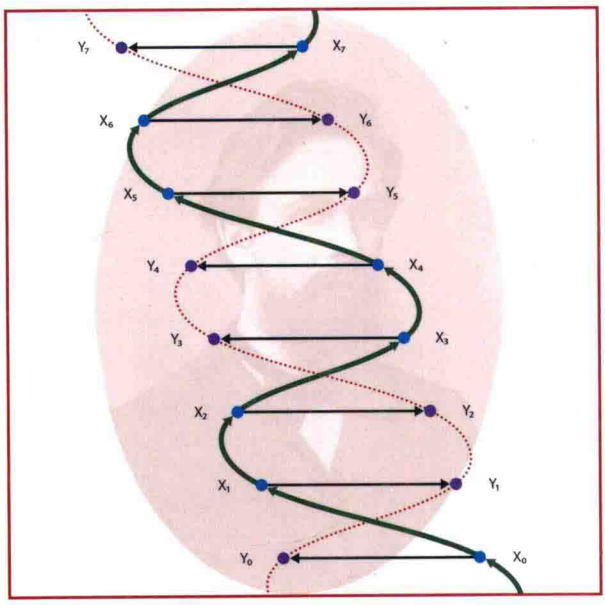
Princeton Series in APPLIED MATHEMATICS

Hidden Markov Processes

Theory and Applications to Biology



M. Vidyasagar

Hidden Markov Processes

Theory and Applications to Biology

M. Vidyasagar

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press, 6 Oxford Street,
Woodstock, Oxfordshire OX20 1TW

press.princeton.edu

Library of Congress Cataloging-in-Publication Data

Vidyasagar, M. (Mathukumalli), 1947–
Hidden Markov processes : theory and applications to biology / M. Vidyasagar.
p. cm. – (Princeton series in applied mathematics)
Includes bibliographical references and index.
ISBN 978-0-691-13315-7 (hardcover : alk. paper) 1. Computational biology. 2. Markov processes. I. Title.
QH324.2.V54 2014
570.285–dc23

2014009277

British Library Cataloging-in-Publication Data is available

This book has been composed in LaTeX.

The publisher would like to acknowledge the author of this volume for providing the camera-ready copy from which this book was printed.

Printed on acid-free paper ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Hidden Markov Processes

PRINCETON SERIES IN APPLIED MATHEMATICS

Series Editors

The Princeton Series in Applied Mathematics publishes high quality advanced texts and monographs in all areas of applied mathematics. Books include those of a theoretical and general nature as well as those dealing with the mathematics of specific applications areas and real-world situations.

A list of titles in this series appears at the back of the book.

To the memory of
Jan Willems (1939–2013)
Scholar, friend, and role model
He will be missed sorely

यद्यदाचरति श्रेष्टस्तत्तदेवेतरो जनः ।
स यत्प्रमाणम् कुरुते लोकस्तदनुवर्तते ॥

Whatever a great man does, that alone the other men do.
Whatever he sets up as the standard, that the world follows.

Bhagavadgita, Chapter 3, Verse 21.
English translation by Alladi Mahadeva Sastry in 1897.

Preface

Every author aspiring to write a book should address two fundamental questions: (i) Who is the targeted audience? (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 aimed at readers who are in the situation in which I found myself at the time that 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. I got very interested in hidden Markov processes (HMPs) when I realized that several researchers were applying HMPs to problems in computational biology. Thus, after spending virtually my entire research career in blissful ignorance of all matters stochastic, I got down to try to learn something about Markov processes and HMPs. At the same time, I was trying to learn enough basic biology, and to read the literature on the applications of Markov and hidden Markov methods in computational biology. The Markov process literature and the computational biology literature each presented its own sets of problems. The choices regarding the level and contents of the book have been dictated primarily by a desire to enable my readers to learn these topics more easily than I could.

Hidden Markov processes (HMPs) were introduced into the statistics literature as far back as 1966 [12]. Starting in the mid 1970s [9, 10], HMPs have been used in speech recognition, which is perhaps the earliest *application* of HMPs in a nonmathematical context. The paper [50] contains a wonderful survey of most of the relevant theory of HMPs. In recent years, Markov processes and HMPs have also been used in computational biology. Popular algorithms for finding genes from a genome are based either on Markov models, such as GLIMMER and its extensions [113, 36], or on hidden Markov models, such as GENSCAN [25, 26]. Methods for classifying proteins into one of several families make use of a very special type of hidden Markov model known as a "profile" hidden Markov model. See [142] for a recent survey. Finally, the BLAST algorithm, which is universally used to carry

out sequence alignment in an efficient though probabilistic manner, makes use of some elements of large deviation theory for i.i.d. processes. Thus any text aspiring to provide the mathematical foundations of these methods would need to cover all of these topics in mathematics.

Most existing books on Markov processes invariably focus on processes with *infinite* state spaces. Books such as [78, 121] 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 [121] is inaccessible to many if not most engineers, and certainly to most biologists. A typical engineer or mathematically inclined biologist can cope with Markov processes with *finite* state spaces because they can be analyzed using only matrices, eigenvalues, eigenvectors, and the like. At the same time, books on Markov processes with *finite* state spaces seldom go beyond computing stationary distributions, and generally ignore advanced topics such as ergodicity, parameter estimation, and large deviation theory. Yet, as I have stated above, these "advanced" ideas are used in computational biology, even if this fact is not always highlighted. There are some notable exceptions such as [116] that discusses ergodicity of Markov processes, and [100] that discusses parameter estimation; but neither of these books discusses large deviation theory, even for i.i.d. processes, let alone for Markov processes.

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 exclusively on stochastic processes assuming values in a finite set, so that technicalities are kept to an absolute minimum. By restricting attention to Markov processes with finite state spaces, I try to capture most of the interesting phenomena such as ergodicity and large deviation theory, while giving elementary proofs that are accessible to anyone who knows undergraduate-level linear algebra.

In the area of HMPs, most of the existing texts discuss only the computation of various likelihoods, such as the most likely state trajectory corresponding to a given observation sequence (also known as the Viterb algorithm), or to the determination of the most likely parameter set for a hidden Markov model *of a given, prespecified order* (also known as the Baum-Welch algorithm). In contrast, very little attention is paid to realization theory, that is, constructing a hidden Markov model on the basis of its statistics. Perhaps the reason is that, until recently, realization theory for hidden Markov processes was not in very good shape, despite the publication of the monumental paper [6]. In the present book, I have attempted to remedy the situation by including a thorough discussion of realization theory for hidden Markov processes, based primarily on the paper [133].

Finally, I decided to include a discussion of large deviation theory for both

i.i.d. processes as well as Markov processes. The material on large deviation theory for i.i.d. processes, especially the so-called method of types, is used in the proofs of the BLAST algorithm, which is among the most widely used algorithms in computational biology, used for sequence alignment. I decided also to include a discussion of large deviation theory for Markov processes, even though there aren't any applications to computational biology, at least as of now. This discussion is a compendium of various results that are scattered throughout the literature, and is based on the paper [135].

At present there are several engineers and mathematicians who would like to contribute to computational biology and suggest suitable algorithms. Such persons face 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. Perhaps nothing illustrates the terminology gap better than the fact that the very phrase "reductionist approach" means entirely different things to mathematicians and to biologists. To mathematicians reductionism means reducing a subject to its core principles. For instance, if a ring is defined as a set with two binary associative operations etc., then certain universal conclusions can be drawn that apply to *all* rings. In contrast, to biologists reductionism means constructing the simplest possible exemplar of a biological system, and studying it in great detail, in the hope that conclusions derived from the simplified system can be extrapolated to more complex exemplars. Or to put it another way, biological reductionism is based on the premise that a biological system is merely an aggregation of its parts, each of which can be studied in isolation. The difficulties with this premise are obvious to anyone familiar with "emergent behavior," whereby complex systems exhibit behavior that has not been explicitly programmed into them; but despite that, reductionism (in the biological sense) is widely employed, perhaps due to the inherent complexity of biological systems.

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 encyclopedic treatment of many aspects of computational biology are referred to [11, 53]. But despite laying out these boundary conditions, I still decided not to attempt a thorough and up-to-date treatment of all available algorithms in genomics and proteomics that are based on hidden Markov processes. The main reason is that the actual details of the algorithms keep changing very rapidly, whereas the underlying theory does not change very much over time.

For this reason, the material in Part 3 of the book on computational biology only *presents the flavor* of various algorithms, with up-to-date references.

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 also spur the engineering and statistics research community to study some new and interesting research problems.

I would like to conclude by dedicating this book to the memory of Jan Willems, who passed away just a few months before it was finished. Jan was a true scholar, a personal friend, and a role model for aspiring researchers everywhere. With his passing, the world of control theory is infinitely poorer. Dedicating this book to him is but a small recompense for all that I have learned from him over the years. May his soul rest in peace!

M. Vidyasagar
Dallas and Hyderabad
December 2013

Contents

Preface xi

PART 1. PRELIMINARIES 1

Chapter 1. Introduction to Probability and Random Variables 3

1.1 Introduction to Random Variables 3
1.1.1 Motivation 3
1.1.2 Definition of a Random Variable and Probability 4
1.1.3 Function of a Random Variable, Expected Value 8
1.1.4 Total Variation Distance 12
1.2 Multiple Random Variables 17
1.2.1 Joint and Marginal Distributions 17
1.2.2 Independence and Conditional Distributions 18
1.2.3 Bayes' Rule 27
1.2.4 MAP and Maximum Likelihood Estimates 29
1.3 Random Variables Assuming Infinitely Many Values 32
1.3.1 Some Preliminaries 32
1.3.2 Markov and Chebycheff Inequalities 35
1.3.3 Hoeffding's Inequality 38
1.3.4 Monte Carlo Simulation 41
1.3.5 Introduction to Cramér's Theorem 43

Chapter 2. Introduction to Information Theory 45

2.1 Convex and Concave Functions 45
2.2 Entropy 52
2.2.1 Definition of Entropy 52
2.2.2 Properties of the Entropy Function 53
2.2.3 Conditional Entropy 54
2.2.4 Uniqueness of the Entropy Function 58
2.3 Relative Entropy and the Kullback-Leibler Divergence 61

Chapter 3. Nonnegative Matrices 71

3.1 Canonical Form for Nonnegative Matrices 71
3.1.1 Basic Version of the Canonical Form 71
3.1.2 Irreducible Matrices 76
3.1.3 Final Version of Canonical Form 78
3.1.4 Irreducibility, Aperiodicity, and Primitivity 80
3.1.5 Canonical Form for Periodic Irreducible Matrices 86

3.2 Perron-Frobenius Theory 89
3.2.1 Perron-Frobenius Theorem for Primitive Matrices 90
3.2.2 Perron-Frobenius Theorem for Irreducible Matrices 95

PART 2. HIDDEN MARKOV PROCESSES 99

Chapter 4. Markov Processes 101
4.1 Basic Definitions 101
4.1.1 The Markov Property and the State Transition Matrix 101
4.1.2 Estimating the State Transition Matrix 107
4.2 Dynamics of Stationary Markov Chains 111
4.2.1 Recurrent and Transient States 111
4.2.2 Hitting Probabilities and Mean Hitting Times 114
4.3 Ergodicity of Markov Chains 122

Chapter 5. Introduction to Large Deviation Theory 129
5.1 Problem Formulation 129
5.2 Large Deviation Property for I.I.D. Samples: Sanov's Theorem 134
5.3 Large Deviation Property for Markov Chains 140
5.3.1 Stationary Distributions 141
5.3.2 Entropy and Relative Entropy Rates 143
5.3.3 The Rate Function for Doubleton Frequencies 148
5.3.4 The Rate Function for Singleton Frequencies 158

Chapter 6. Hidden Markov Processes: Basic Properties 164
6.1 Equivalence of Various Hidden Markov Models 164
6.1.1 Three Different-Looking Models 164
6.1.2 Equivalence between the Three Models 166
6.2 Computation of Likelihoods 169
6.2.1 Computation of Likelihoods of Output Sequences 170
6.2.2 The Viterbi Algorithm 172
6.2.3 The Baum-Welch Algorithm 174

Chapter 7. Hidden Markov Processes: The Complete Realization Problem 177
7.1 Finite Hankel Rank: A Universal Necessary Condition 178
7.2 Nonsufficiency of the Finite Hankel Rank Condition 180
7.3 An Abstract Necessary and Sufficient Condition 190
7.4 Existence of Regular Quasi-Realizations 195
7.5 Spectral Properties of Alpha-Mixing Processes 205
7.6 Ultra-Mixing Processes 207
7.7 A Sufficient Condition for the Existence of HMMs 211

PART 3. APPLICATIONS TO BIOLOGY 223

Chapter 8. Some Applications to Computational Biology 225
8.1 Some Basic Biology 226
8.1.1 The Genome 226

8.1.2 The Genetic Code 232
8.2 Optimal Gapped Sequence Alignment 235
8.2.1 Problem Formulation 236
8.2.2 Solution via Dynamic Programming 237
8.3 Gene Finding 240
8.3.1 Genes and the Gene-Finding Problem 240
8.3.2 The GLIMMER Family of Algorithms 243
8.3.3 The GENSCAN Algorithm 246
8.4 Protein Classification 247
8.4.1 Proteins and the Protein Classification Problem 247
8.4.2 Protein Classification Using Profile Hidden Markov Models 249

Chapter 9. BLAST Theory 255
9.1 BLAST Theory: Statements of Main Results 255
9.1.1 Problem Formulations 255
9.1.2 The Moment Generating Function 257
9.1.3 Statement of Main Results 259
9.1.4 Application of Main Results 263
9.2 BLAST Theory: Proofs of Main Results 264

Bibliography 273

Index 285

PART 1
Preliminaries

Chapter One

Introduction to Probability and Random Variables

1.1 INTRODUCTION TO RANDOM VARIABLES

1.1.1 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? If we make multiple draws, replacing the drawn ball each time and shaking the urn thoroughly before the next draw, what is the likelihood that we have to make at least ten draws before we draw a red ball for the first time? Probability theory provides a mathematical abstraction and a framework where such issues can be addressed.

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] This situation is studied in Section 1.3. 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 1930s precisely to address the issues thrown up by having uncountably many possible outcomes. Subsequent developments in probability theory have been based on the axiomatic foundation laid out in [81].

Example 1.1 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 as drawing any other color, neither more nor less. This leads to the observation that the likelihood of drawing a red ball (or any other ball) is $1/4 = 0.25$.

Example 1.2 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 1.3 There can be instances where such analytical reasoning can fail. Suppose that all balls have the same diameter, but 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*. Often the observed frequency of an outcome is referred to as its **empirical probability**, or the empirical estimate of the true but unknown probability based on a particular set of experiments. 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 would in general 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 important 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 the true probabilities? Such questions are addressed in Section 1.3.3.

1.1.2 Definition of a Random Variable and Probability

Suppose we wish to study the behavior of a "random" variable 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, 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 convention in many mathematics texts, where it is