

Abstract of “Approximation of Stochastic Processes by Hidden Markov Models,” by Athanasios Kehagias, Ph.D., Brown University, May 1992

In this thesis we restrict ourselves to stationary and discrete valued stochastic processes.

A pair of stochastic processes  $(X, Y)$  is a *Hidden Markov Model* (HMM) if  $X$  (the state process) is a Markov process and  $Y$  (the observable process) is an incomplete observation of  $X$ . The observation can be deterministic or noisy and the observable can be a state or a state transition. Hence we have four possible types of HMM's.

First we establish that *all types of HMM's are equivalent*, in the sense that, given any HMM of arbitrary type we can construct a HMM of any other arbitrary type, such that the two models have identical observable processes. Therefore all types of HMM's have the same modelling power.

Second, we consider the problem of *Representation*: what kind of stochastic processes can we approximate with Hidden Markov Models? To make the question meaningful we define two types of stochastic process approximation: (a) *weak approximation*, based on the weak convergence of probability measures and (b) *cross entropy approximation*, based on the Kullback-Leibler informational divergence. Then we prove that *there is a sequence of HMM's (of increasing size) that approximate any ergodic stochastic process in the weak and cross entropy sense*.

Third, we consider the problem of *Consistent Estimation*. To approximate an ergodic process we need a sequence of HMM's of increasing size. For a fixed size Hidden Markov Model we can use the very efficient Baum algorithm to find the Maximum Likelihood parameters estimate. But will the sequence of estimates be consistent (i.e. will it converge to the true process)? The answer is: *the sequence of Maximum Likelihood Estimates will be consistent if the original process is ergodic, has strictly positive probability and conditional probability bounded away from zero*.

Fourth, we develop HMM models of the *raw speech signal* and demonstrate numerically consistency of Maximum Likelihood estimation.

Finally, we develop *Hidden Gibbs Models*, an analogue of HMM, and use these to model one dimensional speech signals and two dimensional images.

# Approximation of Stochastic Processes by Hidden Markov Models

by

Athanasios Kehagias

Dipl. Eng., Aristotle University, Thessaloniki, Greece

M. Sc., Lehigh University

Thesis

Submitted in partial fulfillment of the requirements for  
the Degree of Doctor of Science  
in the Division of Applied Mathematics at Brown University

May 1992

© Copyright

by

Athanasios Kehagias

1992

This dissertation by Athanasios Kehagias is accepted in its present form by  
the Division of Applied Mathematics as satisfying the  
dissertation requirement for the degree of  
Doctor of Science

Date \_\_\_\_\_  
Stuart Geman

Recommended to the Graduate Council

Date \_\_\_\_\_  
Basilis Gidas

Date \_\_\_\_\_  
Donald McClure

Approved by the Graduate Council

Date \_\_\_\_\_  
Peder J. Estrup  
Dean of the Graduate School and Research

For Graziella

## The Vita of Athanasios Kehagias

Athanasios Kehagias was born in Thessaloniki, Greece in 1961. Before enrolling at Brown, he attended the Aristotle University of Thessaloniki, Greece, where he received his Diploma Eng. degree in Electrical Engineering, and Lehigh University in Bethlehem, Pennsylvania, where he received his M.Sc. degree in Applied Mathematics.

Kehagias is a member of the Technical Chamber of Greece, IEEE and AMS. He is interested in Statistics and Probability Theory, especially temporal and spatial stochastic processes, and in Artificial Neural Networks, especially from the Stochastic Control point of view. He is studying the application of the above theories to Speech Recognition, Image Processing and Oceanography.

## Preface

The general problem discussed in this dissertation is the approximation of stochastic processes by Hidden Markov Models. These models are extremely popular in the field of Speech Recognition and they are slowly beginning to be used in other applications as well. The dissertation problem was suggested to me by my advisor Stuart Geman when we realized (in the course of some investigations in Speech Recognition) that there were some mathematical questions about the properties of HMMs to which we did not know the answers. In particular, we wanted to know if consistent estimation was possible within the HMM framework.

A pair of stochastic processes  $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ , is called a **Hidden Markov Model** iff  $\{X_t\}_{t=-\infty}^{\infty}$  is a Markov process and, for all  $t$ ,  $Y_t$  is, roughly speaking, dependent only on  $X_t$  (but not on  $X_{t+k}$ ,  $k = \pm 1, \pm 2, \dots$ ). This definition will be made more precise later.

In this dissertation we will only consider discrete time stochastic processes that take values in a discrete (and usually finite) set. This is what is generally understood by a HMM, even though, from a more general point of view, incompletely observed stochastic dynamical systems (such as ARMA models, neural networks etc.) can be considered as continuous valued HMMs. Somewhat surprisingly, the treatment of continuous valued HMM is easier than that of discrete valued ones. For instance, take our Representation Theorem (Section 2) which says that any stationary ergodic (discrete valued) stochastic process can be approximated weakly by a sequence of HMMs. A close analogue of this theorem for the continuous valued case is Wald's Theorem (weak convergence is replaced by convergence in the mean), which has been known since 1935. In this spirit, it would be interesting to compare the results known about discrete valued HMMs (call these HMMs in the *strict sense*) and continuous valued ones (HMMs in the *wide sense*) and look for generalizations of known results in either direction. This is more than can be done in this dissertation, but a fairly extensive bibliography of the *theory* of strict sense HMMs is included in case the reader wants to further pursue the subject. However no claim of completeness is made. On the other hand, the literature on wide sense HMMs is vast and spreads over several disciplines (Control Theory,

Statistics, Neural Networks and many more) and no attempt was made to cover it.

The popularity of HMMs can be attributed to several factors. Partially observed phenomena occur in many instances and this leads naturally to the Hidden Markov mechanism. Also, from the modelling point of view, HMMs (unlike the standard Markov models) allow the model's memory to extend more than just one step back in the future. In fact, the general HMM has a “fading memory” of the infinite past. This extends the modelling power by far. Last, but not least, in the discrete valued case, the parameter estimation of HMMs can be done by a very efficient algorithm (the Backward - Forward algorithm of Baum).

Let us discuss the contents of the dissertation in some more detail. Consider a general stochastic process  $\{Y_t\}_{t=-\infty}^{\infty}$ , taking values in a finite set  $\Omega = \{0, \dots, L - 1\}$ .

First of all we define several different types of HMMs of the process  $Y$  and show that they are equivalent in the sense that HMMs of any two types can be found, such that they produce identical output processes.

We also introduce a new type of statistical models, closely related to HMMs. We call the new models *Hidden Gibbs Models*; they consist of an underlying Gibbs process  $X$  (it has Gibbs marginal probabilities) and a process  $Y$ , which is an incomplete local observation of  $X$ . The analogy with HMMs is obvious; however, unlike HMMs, HGMs can be generalized to spatial and higher dimensional processes. For the one dimensional case (time processes) we show that the class of HGMs is dense in the class of HMMs.

Therefore in the rest of the dissertation we can use the type of HMM that is most suitable for any particular question (e.g. Representation power, Consistency of Estimation) and generalize to all other types of HMMs (all HMMs are equivalent) or to HGMs (they can approximate HMMs arbitrarily well).

Then we ask the following question (**Representation**): Can we find a *sequence* of HMMs:  $(X_t^N, Y_t^N)_{t=-\infty}^{\infty}$   $N = 1, 2, \dots$  such that  $\{Y_t^N\}_{t=-\infty}^{\infty} \rightarrow \{Y_t\}_{t=-\infty}^{\infty}$  in some appropriate convergence sense? The answer, after considerable elaboration, turns out to be positive for the class of ergodic processes  $\{Y_t\}_{t=-\infty}^{\infty}$ . We prove a similar result for HGMs. Also we report an old result due to

Harris, which, cast in HMM terminology says that any stochastic process that satisfies certain mild conditions, can be represented *exactly* by a HMM with *infinite* state space.

The second question (**Estimation**) we ask is: given some observations  $y_1, \dots, y_n$  from  $\{Y_t\}_{t=-\infty}^{\infty}$  how to estimate an appropriate HMM  $\{(\hat{X}^{N(n)}, \hat{Y}^{N(n)})\}_{t=-\infty}^{\infty}$ ? Furthermore as we let  $n$  go to infinity, do we get a *consistent* sequence of estimates? That is, under what conditions on the process  $\{Y_t\}_{t=-\infty}^{\infty}$  and the method of estimation do we have  $\{\hat{Y}_t^{N(n)}\}_{t=-\infty}^{\infty} \rightarrow \{Y_t\}_{t=-\infty}^{\infty}$ ? The answer is, roughly, the following. Consistency is guaranteed if (i) the process  $\{Y_t\}_{t=-\infty}^{\infty}$  is ergodic and has conditional probabilities bounded away from zero and (ii)  $(\{\hat{X}_t^{N(n)}\}_{t=-\infty}^{\infty}, \{\hat{Y}_t^{N(n)}\}_{t=-\infty}^{\infty})$  is the estimate that maximizes  $n$  - sample Likelihood in a suitably restricted class of HMMs:

$$\left\{ \left\{ \hat{X}_t^{\phi} \right\}_{t=-\infty}^{\infty}, \left\{ \hat{Y}_t^{\phi} \right\}_{t=-\infty}^{\infty} \right\}_{\phi \in \Phi_{N(n)}}.$$

Having developed the theory of Maximum Likelihood Hidden Markov Modelling we return to the problem of Speech Recognition, that was our original motivation, and develop some new Hidden Markov Models of *raw speech*. In this **Speech Modelling** task, we model the human speech as a HMM stochastic process and we estimate its parameters via Maximum Likelihood. The speech is modelled at the *sub-phonemic* level of raw signal, which has not been modelled as a HMM before (phonemic HMMs are common). The speech HMMs produce output that resembles closely original speech data, both in the statistical sense and in its visual appearance.

Finally, we develop some HGM of speech (one dimensional, time process) and images (two dimensional, spatial process) and we estimate their parameters via Maximum Likelihood. We compare these results to Hidden Markov Modelling and find out that in the one dimensional case HMMs are superior to HGMs, because there is a very efficient estimation algorithm for HMMs, this algorithm does not generalize to the two dimensional case, where HGMs are the only viable modelling option.

This dissertation has two fairly distinct parts. The theoretical issues are discussed in Chapters 1- 3; on the other hand Chapters 4 and 5 are about practical modelling. Both approaches are useful and complementary of each other. Enough background material is included so that the applications oriented reader can (hopefully) get a pretty complete picture of the theory of Hidden

Markov Modelling.

## Acknowledgments

I want to offer my thanks to the following people for useful conversations and good company: Yali Amit, Graziella Bertocchi, Paul Dupuis, Basilis Gidas, Ulf Grenander, Nathan Intrator, Markos Katsoulakis, Hans Künsch, Dimitris Lainiotis, Donald McClure, Mike Miller, Krishna Nathan, Les Niles, Chris Raphael and many others. Special thanks go to my advisor Stuart Geman for pointing out the dissertation problem to me, guiding me expertly through its intricacies, helping me many times with proofs and exposition and, above all, showing infinite patience. And very special thanks to my parents.

# Contents

|          |   |           |
|----------|---|-----------|
| <b>0</b> | <b>Introduction</b>                                     | <b>1</b>  |
| <b>1</b> | <b>Preliminaries</b>                                    | <b>5</b>  |
| 1.1      | Preliminaries about Stochastic Processes . . . . .      | 5         |
| 1.2      | Probability Measures and Stochastic Processes . . . . . | 11        |
| 1.3      | Preliminaries about Hidden Markov Models . . . . .      | 18        |
| 1.4      | Preliminaries about Hidden Gibbs Models . . . . .       | 24        |
| 1.5      | Related work . . . . .                                  | 26        |
| <b>2</b> | <b>Representation</b>                                   | <b>29</b> |
| 2.1      | Approximation with Hidden Markov Models . . . . .       | 29        |
| 2.2      | Approximation with Hidden Gibbs Models . . . . .        | 41        |
| 2.3      | Exact Representation . . . . .                          | 45        |
| <b>3</b> | <b>Estimation</b>                                       | <b>49</b> |
| 3.1      | Preliminaries . . . . .                                 | 49        |
| 3.2      | Positivity Properties . . . . .                         | 55        |
| 3.3      | Mixing Properties . . . . .                             | 61        |
| 3.4      | An Auxiliary Hidden Markov Model . . . . .              | 72        |
| 3.5      | Consistency . . . . .                                   | 77        |

|          |   |            |
|----------|---|------------|
| <b>4</b> | <b>Estimation Experiments with Hidden Markov Models</b>             | <b>86</b>  |
| 4.1      | Hidden Markov Models for One Dimensional Processes . . . . .        | 86         |
| 4.2      | The Backward - Forward Likelihood Maximization Algorithm . . . . .  | 88         |
| 4.3      | Experiments with Hidden Markov Models and Artificial Data . . . . . | 89         |
| 4.4      | Experiments with Hidden Markov Models and Speech Data . . . . .     | 95         |
| 4.4.1    | About Speech Recognition Systems . . . . .                          | 97         |
| 4.4.2    | Amorphous Speech Models . . . . .                                   | 99         |
| 4.4.3    | Customized Speech Models . . . . .                                  | 105        |
| <b>5</b> | <b>Estimation Experiments with Hidden Gibbs Models</b>              | <b>114</b> |
| 5.1      | Hidden Gibbs Models for One Dimensional Processes . . . . .         | 115        |
| 5.2      | Likelihood Maximization Algorithms . . . . .                        | 116        |
| 5.3      | Experiments with Hidden Gibbs Models and Speech Data . . . . .      | 117        |
| 5.4      | Hidden Gibbs Models for Two Dimensional Processes . . . . .         | 119        |
| 5.5      | Experiments with Hidden Gibbs Models and Images . . . . .           | 121        |
| <b>6</b> | <b>Conclusion</b>   | <b>129</b> |

# List of Figures

|      |  |     |
|------|--|-----|
| 4.1  | $p, p_4$ . . . . .   | 91  |
| 4.2  | $p, \bar{p}_4$ . . . . .   | 93  |
| 4.3  | $p, p_6$ . . . . .   | 94  |
| 4.4  | A speech signal: “one” . . . . .                                   | 95  |
| 4.5  | A speech signal segment: [uh] . . . . .                            | 96  |
| 4.6  | A speech signal segment: [n] . . . . .                             | 96  |
| 4.7  | A quantized speech signal segment: [uh] . . . . .                  | 97  |
| 4.8  | A quantized speech signal segment: [n] . . . . .                   | 98  |
| 4.9  | Sample path from 4th order HMM estimate of [uh] process . . . . .  | 100 |
| 4.10 | Sample path from 8th order HMM estimate of [uh] process . . . . .  | 102 |
| 4.11 | Sample path from 12th order HMM estimate of [uh] process . . . . . | 103 |
| 4.12 | Sample path from 12th order HMM estimate of [n] process . . . . .  | 104 |
| 4.13 | Sample path from 16th order HMM estimate of [uh] process . . . . . | 105 |
| 4.14 | Sample path from 16th order HMM estimate of [n] process . . . . .  | 106 |
| 4.15 | A prototype component of [uh] . . . . .                            | 107 |
| 4.16 | A prototype component of [n] . . . . .                             | 107 |
| 4.17 | A deterministic model of [uh] . . . . .                            | 108 |
| 4.18 | A deterministic model of [n] . . . . .                             | 109 |
| 4.19 | A quasideterministic model of [uh] . . . . .                       | 110 |

|      |                                   |     |
|------|-----------------------------------|-----|
| 4.20 | A quasideterministic model of [n] | 110 |
| 4.21 | A probabilistic model of [uh]     | 111 |
| 4.22 | A probabilistic model of [n]      | 112 |
| 4.23 | A probabilistic model of [uh]     | 113 |
| 4.24 | A probabilistic model of [n]      | 113 |
|      |                                   |     |
| 5.1  | A probabilistic model of [uh]     | 118 |
| 5.2  | A probabilistic model of [n]      | 118 |
| 5.3  | A probabilistic model of [uh]     | 119 |
| 5.4  | A probabilistic model of [n]      | 120 |
| 5.5  | A picture of stripes              | 122 |
| 5.6  | A model of stripes                | 123 |
| 5.7  | A picture of straw                | 124 |
| 5.8  | A model of straw                  | 125 |
| 5.9  | A picture of paper                | 126 |
| 5.10 | A model of paper                  | 127 |

## List of Symbols

$\mathbf{1}_A(x)$  equals 1 iff string  $x$  is in set  $A$ ; otherwise it equals 0.

$\mathbf{1}_z(x)$  equals 1 iff string  $x$  is the same as string  $z$ ; otherwise it equals 0.

$A(x_1 \dots x_n)$  is the set (rectangle) of strings  $z_1 z_2 \dots$  that  $z_1 = x_1, \dots, z_n = x_n$ ; see Def. 14.

$\mathcal{A}(\Omega)$  is the class of all rectangles in  $\Omega$ ; see Def. 15.

$\mathcal{B}(\Omega)$  is the class of all finite disjoint unions of rectangles in  $\Omega$ ; see Def. 16.

$B_\epsilon(a)$  is a ball centered at  $a$  with radius  $\epsilon$ ; rectangles in  $\Omega$ ; see eq.(1.21).

$D_N, C_N$  are mixing constants; see Lemma 26.

$E_n(\cdot)$  is an expectation function defined for all  $n \geq 1$ ; see eq.(3.100).

$e_n(\cdot)$  is an expectation function defined for all  $n \geq 1$ ; see eq.(3.101).

$E(\cdot)$  is an expectation function; see eq.(3.102);

$H_n(p)$  is the  $n$ -th order entropy of  $p(\cdot)$ ; see Def. 9.

$H(p)$  is the entropy of  $p(\cdot)$ ; see Def. 10.

$H_n(q; p)$  is the  $n$ -th order cross entropy of  $q(\cdot)$  with respect to  $p(\cdot)$ ; see Def. 11.

$H(q; p)$  is the cross entropy of  $q(\cdot)$  with respect to  $p(\cdot)$ ; see Def. 12.

$L(\phi; y_1 \dots y_n)$  is the Likelihood function of observations  $y_1, \dots, y_n$ ; see Def. 43.

$N_\alpha$  is an integer dependent on the positivity bound of the original process  $Y$ ; see Theorem 20.

$p(\cdot)$  is a probability function; see Def. 4.

$p(\cdot|\cdot)$  is a conditional probability function (with finite or infinite conditioning); see Defs.6, 41.

$p_N$  is the probability function of the  $N$ -th order AR model; see Def. 40.

$p_{N,n}^y$  is the  $(N, n)$ -th order empirical probability of sample  $y_1, \dots, y_n$ ; see Def. 42.

$r_N$  is the probability function of the  $N$ -th order NAR model; see Def. 48.

$\mathbf{R}$  is the set of real numbers.

$(X^N, Y^N)$  denotes the  $N$ -th order Autoregressive (AR) model of  $Y$ ; see Def. 40.

$(\bar{X}^N, \bar{Y}^N)$  denotes the  $N$ -th order Nosiya Autoregressive (NAR) model of  $Y$ ; see Def. 48.

$\mathbf{Z}$  is the set of integers.

$\mathbf{Z}^2$  is the set of pairs of integers.

$\alpha$  is a positivity bound; see Theorem 3.1.

$\alpha_N$  is the positivity bound on the elements of class  $\Phi_N$ ; see Lemma 24.

$\beta_N$  equals  $(1 - \alpha_N^{8N})$ ; see Lemma 24.

$\pi(\cdot)$  is the probability measure associated with probability function  $p(\cdot)$ ; see Theorem 4.

$\sigma(\Omega)$  is the smallest  $\sigma$ -algebra generated by  $\mathcal{B}(\Omega)$ ; see Def. 17.

$\Phi_N(\Omega)$  or  $\Phi_N$  is the  $N$ -th class of  $(P, Q)$  parameters that describe a SNM model; see Def. 46.

$\Psi_{N,n}(y)$  is the set of ML estimates in the  $\Phi_N$ ; see Def. 47.

$\Omega$  is the alphabet of the original process  $Y$  that we want to model.

$\Omega^N$  is the set of  $N$ -long strings from  $\Omega$ ; see Def. 1.

$\Omega^*$  is the set of all finite length strings from  $\Omega$ ; see Def. 1.2.

$\Omega^\infty$  is the set of all infinite length strings from  $\Omega$ ; see Def. 3.

$\Omega_0$  is the class of zero probability states of a Markov process; see Def. 40.

$\Omega_+$  is the class of positive probability states of a Markov process; see Def. 40.

$x \rightsquigarrow z$  means  $z$  is a consequent of  $x$ ; see Def. 23.

$x \sim z$  means  $x$  and  $z$  are in the same ergodic class; see Def. 24.

$x \leftrightarrow z$  means  $x$  and  $z$  are neighbors; see Def. 33.

$\xrightarrow{w}$  indicates weak convergence; see Def. 8.

# Chapter 0

## Introduction

The subject of this dissertation is “Approximation of stochastic processes by Hidden Markov Models”. Here is a short, intuitive description of the problem: we are given an “original” stochastic process and a special class of stochastic processes (*Hidden Markov Models*); and we want to pick up a sequence of elements of the class that “converges” to the original stochastic process.

Let us first discuss the terms used in the previous paragraph. Then we will proceed to list the contents of the thesis chapter by chapter.

We assume that the reader is familiar with the concept of a stochastic process. *Note that in this thesis we limit ourselves to discrete-time, stationary stochastic processes that take values in a finite set.* Such processes can be mapped uniquely to suitable probability measures (this mapping is worked out in the Section 1.2). Hence, convergence of stochastic processes will be reduced to convergence of measures.

Hidden Markov Models are a particular class of stochastic processes. Strictly speaking, a HMM is a *pair* of stochastic processes  $(X, Y)$ , where  $X$  is a Markov process and  $Y$  has an *instantaneous* dependence on  $X$ . The term HMM has been applied to a very specific type of model, used mostly in Speech Recognition, but in a more general sense many of the popular engineering modelling methods can be considered as HMM (e.g. the type of modelling used in stochastic control theory [May88], autoregressive modelling [Aok87, LS83], image processing models [GG84], stochastic

recurrent neural networks [Keh91a], etc.). However, in this thesis we limit ourselves to “classical” HM Models (such as these used in speech recognition [R+83, R+85]), or models that can be shown to be equivalent to the classical ones.

We also develop a new model which we call a *Hidden Gibbs Model (HGM)*. It is a pair of processes  $(X, Y)$ , where  $X = \{X_t\}_{t \in T}$  is a Gibbs process (equivalently a Markov Random Field) and, for all  $t \in T$ ,  $Y_t$  depends only on  $X_t$  (local dependence).  $T$  is either  $\mathbf{Z}$  (the integers) or a higher dimensional version (typically  $\mathbf{Z}^2$ , the pairs of integers). The Hidden Gibbs Model is a generalization of the Hidden Markov Model to high dimensional processes, with potential applications to image processing. The modelling and estimation of such higher dimensional stochastic processes (e.g. processes over a square lattice) are discussed briefly in the last chapter. The work reported on higher dimensional processes is preliminary. It is presented as an indication of the modelling potential, generality and practicality of Hidden Markov Models.

A central topic of this thesis is approximation. In certain cases it is intuitively obvious when the approximation is good. For instance, when we plot natural speech data, every elementary speech unit (technically, a “phoneme”) has a distinct visual pattern of peaks and valleys; to a certain degree, different instances of the same phoneme retain this visual texture, which can be used to differentiate between different phonemes (see [Keh91a, Keh91b]). A “good” model should reproduce the same pattern. On the other hand, we want to be able to measure objectively the goodness of a model in terms of a statistical figure of merit. To achieve such a statistical evaluation we introduce a one-to-one correspondence between stationary stochastic processes and probability measures. Having established this correspondence we discuss approximation of stochastic processes in terms of weak convergence of measures or, alternatively, in terms of the cross entropy “distance” between probability measures.

Let us now discuss the contents of the thesis chapter by chapter.

Chapter 1 deals with definitions, preliminary results and so on. Most of the material is well known; there are however some original results, in particular the proof of equivalence of various types of HMM’s. The existence of many, slightly different, types of HMM’s (all of which we prove

to be equivalent) is in itself an interesting fact, which can be explained by looking at the history of HM modelling. With this in mind, we present an overview of work related to HMM's at the end of Chapter 1.

Chapter 2 deals with the following question: what types of stochastic processes can be approximated by HMM's? what types can be exactly represented? To make the question meaningful we define two types of stochastic process approximation: (a) *weak approximation*, based on the weak convergence of probability measures and (b) *cross entropy approximation*, based on the Kullback-Leibler informational divergence. Then we show that every stationary, ergodic stochastic process can be approximated arbitrarily well by a sequence of HMM's with finite state space. If the state space of the HMM's is allowed to be uncountably infinite and some additional conditions are imposed to the original process then exact representation is also possible.

Chapter 3 deals with estimation questions. It is true that we can approximate arbitrarily well a *completely known* stochastic process; however in practice all we know about the process is a finite number of observations  $y_1, \dots, y_n$  and we have to choose the optimal model based on these observations. Our optimality criterion is Maximization of the Likelihood function. This depends not only on the observations but also on the class of HMM's over which we maximize the Likelihood. The main result of Chapter 3 is the following. Suppose we are given an ergodic stochastic process  $Y$  that satisfies certain positivity conditions. For any fixed number  $n$  of observations we determine an appropriate class of HMM's  $\Phi_n$  such that the following convergence result holds. Denote by  $\hat{Y}^n$  the Maximum Likelihood estimate of  $Y$  in class  $\Phi_n$ . Then  $\hat{Y}^n \rightarrow Y$  in the cross entropy sense.

In Chapter 4 we present numerical results of a number of estimation experiments. In these experiments we start with observations of an original (one-dimensional) stochastic process and develop models that approximate it. The parameters of the model are the state transition probabilities (we use the DNM version of HMM) and their Maximum Likelihood values are estimated by use of the Backward / Forward algorithm of Baum. In some of the experiments the observations are artificially synthesized; in others we use natural speech data.

In Chapter 5 we also present numerical experiments. Here we use Hidden Gibbs Models to model speech signals (one dimensional processes), as well as images (two dimensional processes), In the case of HGM's the parameters are the coefficients of the energy function and the ML estimation is performed by stochastic gradient descent; the gradient is computed by stochastic relaxation. This algorithm is slower than the Backward / Forward algorithm, but also more versatile, as it can be generalized to processes of dimension greater than 1. The Backward / Forward algorithm depends on a factorization of marginals which is only possible for one dimensional processes.

In Chapter 6 we summarize the conclusions and results of this work. Namely, we have established some important mathematical properties of HMMs as well as developed some interesting and promising new models for specific types of processes (speech, images).

Finally, there is a Bibliography section and a List of Symbols.

# Chapter 1

## Preliminaries

In this chapter we deal with preliminary matters. In section 1.1 we define most of the terms that will be used later and establish some notation. In section 1.2 we work out the mapping of stochastic processes to probability measures. In section 1.3 we present fundamental concepts of Markov processes and Hidden Markov Models and prove the equivalence of the various types of HMM's. In section 1.4 we discuss Hidden Gibbs Models; finally in section 1.5 we briefly review previous work in speech modelling, HMM's and estimation.

### 1.1 Preliminaries about Stochastic Processes

In this section we present definitions, notation and basic facts about stochastic processes in general, that we will use repeatedly in later chapters. Most of the definitions can be skipped for the time being and read when needed.

In this thesis, unless specifically stated otherwise, we always deal with *discrete time* stochastic processes that take values in a *finite alphabet*. By finite alphabet we simply mean a set with a finite number of elements, which for convenience are usually taken to be consecutive integers. For example:  $X_t$ ,  $t = 0, \pm 1, \pm 2, \dots$ , and  $\forall t$ ,  $X_t$  equals some  $x \in \Omega = \{0, \dots, K - 1\}$ . So the **alphabet** of the stochastic process is  $\Omega$ . We will be interested in **strings** from  $\Omega$ , that is,

sequences of **characters** from  $\Omega$ . For example,  $x_1 \dots x_n$ , with  $x_1, \dots, x_n \in \Omega$ . Given two strings from  $\Omega$ ,  $x = x_1 \dots x_m$ ,  $z = z_1 \dots z_n$ , their **concatenation**  $xz$  is the string  $x_1 \dots x_m z_1 \dots z_n$ .

Given a string  $x = x_1 x_2 \dots$  we will use the notation  $[x]_n \doteq x_1 \dots x_n$ . Given a set of integers  $C$  and a string  $x$ ,  $x_C \doteq \{x_t\}_{t \in C}$ . E.g., for  $C = \{1, 4, 11, 12\}$ ,  $x_C = \{x_1, x_4, x_{11}, x_{12}\}$ . Given a string  $x$  and a set of strings  $A$ ,  $\mathbf{1}_A(x) = 1$  iff  $x \in A$  and zero otherwise. Given two strings  $x, z$ ,  $\mathbf{1}_x(z) = 1$  iff  $x = z$  and zero otherwise.

The following sets of strings are of interest:

**Definition 1** *Given an alphabet  $\Omega$ , for all  $n \geq 1$  we define the set  $\Omega^n$  as follows:*

$$\Omega^n = \{x = x_1 x_2 \dots x_n \text{ any } x_1, x_2, \dots, x_n \in \Omega\}. \quad (1.1)$$

**Definition 2** *Given an alphabet  $\Omega$ , we define the set  $\Omega^*$  as follows:*

$$\Omega^* = \{x = x_1 \dots x_n, \text{ any } n \geq 1, x_1, \dots, x_n \in \Omega\} \quad (1.2)$$

**Definition 3** *Given an alphabet  $\Omega$ , we define the set  $\Omega^\infty$  as follows:*

$$\Omega^\infty = \{x = x_1 x_2 \dots \text{ any } x_1, x_2, \dots \in \Omega\}. \quad (1.3)$$

We assume the definitions of a **stationary** and **ergodic** stochastic process to be known. See, for example, Billingsley [Bil65]. We also assume known the concept of a **probability measure**, or simply a **probability** [Bil76]. This is to be distinguished from a *probability function*, which will be defined presently. For a clarification of the difference between a probability and a probability function, see the remarks below.

A discrete time, stationary stochastic process with a finite alphabet is fully characterized by its probability function, which is defined as follows:

**Definition 4** *Given a stationary stochastic process  $X_t, t = 0, \pm 1, \pm 2, \dots$  with finite alphabet  $\Omega$  we*

define  $\forall t, \forall n \geq 0, \forall x_0, \dots, x_n \in \Omega$

$$p(x_0 \dots x_n) \doteq \Pr(X_t = x_0 \dots X_{t+n} = x_n). \quad (1.4)$$

The **probability function** of  $X_t, t = \dots, -1, 0, 1, \dots$  is the function  $p : \Omega^* \mapsto [0, 1]$ .

**Definition 5** Given a stationary stochastic process  $X = X_t, t = 0, \pm 1, \pm 2, \dots$  with probability function  $p$  and finite alphabet  $\Omega$ . Consider the restriction of  $p$  on  $\Omega^N : p(x_1 \dots x_N), x_1, \dots, x_N \in \Omega$ .

This is a probability on  $\Omega^N$  and is called the  **$N$ -th order marginal probability** of  $X$ .

**Remark:** Note that (because of stationarity)  $p(x_0 \dots x_n)$  in (1.4) is independent of  $t$ .

**Remark:** The stochastic process is fully defined by its probability function, so we use the following three symbolisms interchangeably to denote the stochastic process:  $\{X_t\}_{t=-\infty}^{\infty}, X, p$ .

**Remark:** Note the difference between a *probability*, e.g. the marginal probabilities of Definition 1.5, and a *probability function*. A probability  $p$  is defined over a restricted set  $\Omega^N$  and sums up to 1:  $\sum_{x \in \Omega^N} p(x) = 1$ . A probability function  $p$  is defined over an extended set  $\Omega^*$  and does *not* sum up to 1. However, we use the same symbol  $p$  for both of them, as the probability is a restriction of the probability function on an appropriate space.

Another important function for a stochastic process is the *conditional probability function*. We present here the definition that relates to *finite past* conditioning; the definition for *infinite past* conditioning will be given later.

**Definition 6** Given a stationary stochastic process  $X$  we define its **finite conditional probability function**  $p(\cdot | \cdot) : \Omega \times \Omega^* \mapsto [0, 1]$ , for  $n = 1, 2, \dots, x \in \Omega, z \in \Omega^n$ , by

$$p(x|z) \doteq \begin{cases} \frac{\Pr(X_t=x, X_{t-1} \dots X_{t-n}=z)}{\Pr(X_{t-1} \dots X_{t-n}=z)} & \text{when } \Pr(X_{t-1} \dots X_{t-n} = z) > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (1.5)$$

We also write, for all  $n \geq 1$ ,  $x_0, \dots, x_{-n}$  :

$$\Pr(X_t = x_0 | X_{t-1} = x_{-1}, \dots, X_{t-n} = x_{-n}) \doteq p(x_0 | x_{-1} \dots x_{-n}). \quad (1.6)$$

One more definition related to probability functions will be needed.

**Definition 7** A stationary stochastic process  $X$  with probability function  $p$  and finite alphabet  $\Omega$  will be called **positive** iff  $p(x) > 0 \quad \forall x \in \Omega^*$ .

Now we can describe the convergence of a sequence of stochastic processes  $X^n$  to  $X$  in terms of convergence of the respective probability functions. We use two alternative definitions of convergence:

**Definition 8** Given a sequence of stationary stochastic processes  $X^n$ , (with respective probability functions  $p_n$  and finite alphabet  $\Omega$ )  $n = 1, 2, \dots$  and a stochastic process  $X$  (with probability function  $p$  and finite alphabet  $\Omega$ ), we say  $X^n$  converges **weakly** to  $X$ , denoted by  $w\text{-}\lim_{n \rightarrow \infty} X^n = X$  or  $X^n \xrightarrow{w} X$ , iff

$$\forall x \in \Omega^* \quad p_n(x) \rightarrow p(x). \quad (1.7)$$

**Remark:** This type of convergence is called weak convergence for the following reason: if we map every stationary stochastic process to a probability measure, then weak convergence of stochastic processes corresponds to weak convergence of measures as defined in, e.g., Billingsley [Bil71]. A proof of this claim along with other facts about measures generated by stochastic processes will be found in Section 1.2.

Before we can define a second type of convergence, we need to recall the definitions of entropy and cross entropy of stochastic processes.

**Definition 9** Given a stationary stochastic process  $X$  with probability function  $p$  and finite al-

phabet  $\Omega$ , we define its **n-th order entropy**, denoted by  $H_n(p)$  or  $H_n(X)$ , as follows:

$$H_n(p) \doteq - \sum_{x \in \Omega, z \in \Omega^n} p(xz) \log p(x|z). \quad (1.8)$$

(We follow the usual convention that  $0 \log 0 = 0$ .)

The sequence  $H_n(p)$ ,  $n = 1, 2, \dots$  converges when  $p$  is a stationary process. The following theorem is proven in [Ash65] and will be used many times in what follows.

**Theorem 1** (*Entropy Convergence*) *Given a stationary process  $X$ , with finite alphabet and probability function  $p$  the limit  $\lim_{n \rightarrow \infty} H_n(p)$  always exists.*

Therefore the following definition makes sense:

**Definition 10** *Given a stationary process  $X$ , with finite alphabet and probability function  $p$ , we define its **entropy** by*

$$H(p) \doteq \lim_{n \rightarrow \infty} H_n(p). \quad (1.9)$$

The following two theorems will be used in later chapters:

**Theorem 2** *Given a stationary stochastic process  $X$ , with finite alphabet  $\Omega$  and probability function  $p$ , we have*

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{x \in \Omega^n} p(x) \log p(x) = H(p). \quad (1.10)$$

**Proof:** In Ash [Ash65]. •

**Theorem 3** (*Shannon-Breiman-Macmillan*) *Given a stationary ergodic process  $X$ , with finite alphabet and probability function  $p$ , we have*

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log p(X_1 \dots X_n) \stackrel{a.s.}{=} H(p). \quad (1.11)$$

**Proof:** In Billingsley [Bil65]. •

Similarly to entropy of a process we define the *cross entropy* of two processes:

**Definition 11** *Given two stationary stochastic processes  $X, Y$  with probability functions  $p, q$  respectively, and finite alphabet  $\Omega$ , we define the  **$n$ -th order cross entropy** of  $Y$  with respect to  $X$ , denoted by  $H_n(Y; X)$  or  $H_n(q; p)$ , as follows. Take all  $n \geq 1, p, q$  such that if  $x \in \Omega^{n+1}$  and  $q(x) = 0$  then  $p(x) = 0$ ; for such  $n, p, q$  we define*

$$H_n(q; p) \doteq \sum_{x \in \Omega, z \in \Omega^n} p(xz) \log \frac{p(x|z)}{q(x|z)}; \quad (1.12)$$

(We follow the convention  $0 \log \frac{0}{0} = 0$ .) For all other  $n \geq 1, p, q$ , we define  $H_n(p; q) \doteq \infty$ .

**Remark:** The  $n$ -th order cross entropy has a close connection with the Kullback - Leibler distance between probability distributions [Ku59]. To see this, rewrite  $H_n(q; p)$  as

$$H_n(q; p) = \sum_{z \in \Omega^n} p(z) \left\{ \sum_{x \in \Omega} p(x|z) \log \frac{p(x|z)}{q(x|z)} \right\}. \quad (1.13)$$

The interpretation is the following: consider the conditional probability  $p(\cdot|z)$  as a probability on  $\Omega$  which has a random dependence on the conditioning  $z$ . Then (1.13) says that  $H_n(q; p)$  is the expected value of the Kullback - Leibler distance between the (random) probabilities  $q(\cdot|z)$  and  $p(\cdot|z)$ .

Even if for all  $n \geq 1$ , and for all  $x \in \Omega^{n+1}$   $q(x) = 0 \Rightarrow p(x) = 0$ , the limit  $H_n(p; q)$  as  $n$  goes to infinity need not exist. Therefore we have the following definition:

**Definition 12** *Given two stationary processes  $X, Y$  with probability functions  $p, q$  respectively, and the same finite alphabet, we define the **cross-entropy** of  $Y$  with respect to  $X$ , denoted by  $H(Y; X)$  or  $H(q; p)$ , as follows:*

$$H(q; p) \doteq \lim_{n \rightarrow \infty} H_n(q; p) \quad (1.14)$$

when the limit exists; otherwise  $H(q; p)$  is undefined.

Now we are ready to define *cross entropy convergence*.

**Definition 13** *Given a sequence of stochastic processes  $X^n$  ( $n = 1, 2, \dots$ ) with probability functions  $p^n$  respectively, and finite alphabet  $\Omega$ , and a stochastic process  $X$ , with probability function  $p$ , and finite alphabet  $\Omega$ , we say that  $X^n$  tends to  $X$  in the **cross entropy** iff  $\lim_{n \rightarrow \infty} H(p^n; p)$  exists and equals 0.*

**Remark:** Weak convergence does *not* imply convergence in the cross entropy; conversely convergence in the cross entropy does *not* imply weak convergence. The two types of convergence are independent and supply complementary information: weak convergence guarantees convergence of all finite length marginals, whereas cross entropy convergence guarantees on the average convergence of infinite past dependence.

## 1.2 Probability Measures and Stochastic Processes

In this section we show how discrete time stationary stochastic processes with finite alphabet can be mapped to probability measures. This will be useful in understanding the notion of weak convergence of processes, as well as in talking about events that involve infinite samples from stochastic processes.

Consider a discrete time stationary stochastic process  $Y$  with finite alphabet  $\Omega$  and probability function  $p$ . We have defined  $\Omega^\infty$  in the previous section. Now consider a special class of subsets of  $\Omega^\infty$ :

**Definition 14** *A rectangle  $A(a_1 \dots a_n)$  in  $\Omega^\infty$  is defined by*

$$A(a_1 \dots a_n) \doteq \{b \in \Omega^\infty \text{ such that } b = b_1 b_2 b_3 \dots, \quad b_1 = a_1, \dots, b_n = a_n\}. \quad (1.15)$$

**Definition 15** *The class of rectangles of  $\Omega^\infty$   $\mathcal{A}(\Omega)$  is defined by*

$$\mathcal{A}(\Omega) \doteq \{A(a_1 \dots a_n), n \geq 1, a_1, \dots, a_n \in \Omega\} \quad (1.16)$$

It is easy to check that  $\mathcal{A}(\Omega)$  is a *semi-algebra* [Roy68]. The class  $\mathcal{A}(\Omega)$  is a set of infinite length strings, but they are “specified up to a finite length”. They are in a natural correspondence with the finite length strings  $a_1 \dots a_n$  that specify them. It is natural to define a probability-like set function on  $\mathcal{A}(\Omega)$  in terms of the probability functions of finite length strings. So we define a positive set function  $\pi : \mathcal{A}(\Omega) \mapsto [0, 1]$  by

$$\pi(A(a_1 \dots a_n)) \doteq p(a_1 \dots a_n). \quad (1.17)$$

Furthermore,  $\pi$  is *finitely additive*: for all  $A \in \mathcal{A}(\Omega)$  such that there is  $m \geq 1, A_1, \dots, A_m \in \mathcal{A}(\Omega)$  satisfying  $A = A_1 \cup \dots \cup A_m$ , and  $A_1, \dots, A_m$  disjoint, we have  $\pi(A) = \pi(A_1) + \dots + \pi(A_m)$ . The finite additivity follows from the standard probability consistency properties of  $p(\cdot)$ .

We would like to assign probabilities to sets that correspond to “infinite” events about  $Y$ , e.g. the event that  $Y_t = 0$  for  $t = 2, 4, \dots$ . Such sets are not readily described by the probability function  $p$ , but they can be handled in the framework of a probability space, consisting of a master set (in this case  $\Omega^\infty$ ), a sigma-algebra of its subsets and a probability measure thereon (in our case  $\pi$ ). For details on this standard probability theoretic setup see [Bil79]. The basic idea is to start from the class of “events”  $\mathcal{A}(\Omega)$  and extend it to a larger class of “events”, defining  $\pi$  on the extension. This extension will be the subject of Theorem 4; but first we need two additional definitions. First we define  $\mathcal{B}(\Omega)$ , the *algebra* generated by  $\mathcal{A}(\Omega)$ .

**Definition 16** *The class of cylinder sets of  $\Omega^\infty$ , denoted by  $\mathcal{B}(\Omega)$ , is defined to be the class of all sets that are finite disjoint unions of rectangles.*

(It is easy to check  $\mathcal{B}(\Omega)$  is an algebra.)

**Definition 17** Denote by  $\sigma(\Omega)$  the smallest sigma-algebra that contains all elements of  $\mathcal{B}(\Omega)$ .

The  $\sigma$ -algebra is a collection of “events”, i.e. sets of infinite length strings and we want to assign probabilities to all of these sets. This we do by extending  $\pi$  from  $\mathcal{A}(\Omega)$  to  $\sigma(\Omega)$ ; it is the subject of the Extension Theorem:

**Theorem 4 (Extension)** For every stationary stochastic process  $X$ , with probability function  $p$  and finite alphabet  $\Omega$ , there is exactly one probability measure  $\pi$  defined on  $\sigma(\Omega)$  such that

$$\forall N \geq 1, \forall a_1, \dots, a_N \in \Omega \quad p(a_1 \dots a_N) = \pi(A(a_1 \dots a_N)). \quad (1.18)$$

**Proof:** First we show that to every from every  $p$  we can construct *some*  $\pi$  that satisfies (1.18).

As  $\mathcal{B}(\Omega)$  is an algebra, by Proposition 9, p.260 in [Roy68], we can extend  $\pi$  *uniquely* from  $\mathcal{A}(\Omega)$  to  $\mathcal{B}(\Omega)$  (this extension requires *countable additivity* which follows from the probability function properties). As a result of this extension, if  $B \in \mathcal{B}(\Omega)$ ,  $B = \cup_{m=1}^{\infty} B_m$ , where  $B_1, B_2, \dots$  are in  $\mathcal{B}(\Omega)$  and disjoint, then  $\pi(B) = \sum_m \pi(B_m)$ . In other words we have countable additivity, over sets in the algebra.

One more step in the extension process yields the standard probability space setup [Bil79]: denote by  $\sigma(\Omega)$  the smallest sigma-algebra that contains all elements of  $\mathcal{B}(\Omega)$ . We can extend uniquely  $\pi$ , from  $\mathcal{B}(\Omega)$  to  $\sigma(\Omega)$  by Theorem 8, p.257 in [Roy68]. Now: (a)  $\pi$  is countably additive over the sets of the sigma-algebra and (b) it is a nonnegative set function and  $\pi(\Omega^\infty) = 1$ . Hence  $\pi$  is a *probability measure* over  $\sigma(\Omega)$  and  $(\Omega^\infty, \sigma(\Omega), \pi)$  is a probability space.

From our extension procedure it follows that for every stationary stochastic process  $Y$ ,  $p$  with discrete alphabet  $\Omega$ , there is a probability measure  $\pi$  on  $\sigma(\Omega)$ . Now we will prove that for every  $p$  there is *exactly one* measure  $\pi$  that satisfies (1.18).

Given a finite alphabet  $\Omega$  and a probability measure  $\pi$  on  $\sigma(\Omega)$  we can define a probability

function  $p$  (equivalently, a stochastic process  $Y$ ) on  $\Omega^*$  by

$$p(a_1 \dots a_n) = \pi(A(a_1 \dots a_n)). \quad (1.19)$$

Assume we have two measures  $\pi_1, \pi_2$  on  $\sigma(\Omega)$ , that generate the same probability function  $p$  on  $\Omega^*$ ; then they must agree on all rectangles of  $\Omega^\infty$ . But then, by the previous analysis, they agree on  $\mathcal{A}(\Omega)$ , hence on  $\mathcal{B}(\Omega)$  and finally on  $\sigma(\Omega)$ . This implies that  $\pi_1 = \pi_2$  and completes the proof.

•

**Remark:** Therefore, there is a one-to-one correspondence between probability measures and probability functions (equivalently, stationary stochastic processes). So, given a process  $p$ , we will sometimes use  $p$  to indicate the corresponding probability measure; e.g. we will talk about events being  $p$ -a.e. true etc.

Now we proceed to show that our definition of weak convergence of processes from Section 1.1 is equivalent to weak convergence of the corresponding *measures* in the usual sense (see e.g. Billingsley [Bil71]). To denote this weak convergence of measures we write, e.g.,  $\pi_n \rightarrow \pi$ .

Recall [Bil71] that weak convergence of measures is defined with respect to a fixed topology. For the space  $\Omega^\infty$ , with  $\Omega$  finite, the topology that comes to mind most naturally is the *product topology* (see [Roy67]). However, for convenience, we take a different route: we will equip  $\Omega^\infty$  with a metric and use the topology of the open sets associated with this metric; this turns out to be exactly the product topology. In fact there is nothing special about the particular metric we use; a number of other equally reasonable metrics would also yield the product topology.

Define the following metric  $\rho$  on elements of  $\Omega^\infty$ : given  $a = a_1 a_2 \dots, b = b_1 b_2 \dots \in \Omega^\infty$  we define  $\rho(a, b) \doteq \sum_{m=1}^{\infty} |a_m - b_m| \cdot 2^{-m}$ .<sup>1</sup> Note that if  $\rho(a^n, a) \rightarrow 0$  then

$$\forall m_0 \exists n(m_0) \text{ s.t. } \forall n \geq n(m_0) \forall m \leq m_0 \ a_m^n = a_m. \quad (1.20)$$

---

<sup>1</sup>There is nothing special about this metric - many other would do as well, e.g.  $\rho(a, b) = \sum_m |a_m - b_m| m^{-2}$ .

An open sphere  $B_\epsilon(a)$  is defined in the usual way:

$$B_\epsilon(a) = \{b \in \Omega^\infty : \rho(a, b) < \epsilon\}. \quad (1.21)$$

It is easy to check that  $\forall a, \epsilon, n$  such that  $\epsilon < 2^{-(n+1)}$ ,  $B_\epsilon(a) = A(a_1 \dots a_n)$  (where  $A(a_1 \dots a_n)$  is the rectangle defined in Definition 1.14). So  $\mathcal{A}(\Omega)$  is the set of all open spheres in  $\Omega^\infty$ . Then  $\mathcal{A}(\Omega)$  is a base for the topology induced on  $\Omega^\infty$  by  $\rho$ . We define weak convergence of measures with respect to this topology.

Now we prove the following theorem:

**Theorem 5** *Given a stationary stochastic process  $Y$  with probability function  $p$  and a sequence of stationary stochastic processes  $Y^n$ ,  $n = 1, 2, \dots$  with probability functions, respectively,  $p_n$ ,  $n = 1, 2, \dots$  (all the processes having finite alphabet  $\Omega$ ), and corresponding measures  $\pi, \pi_n$ ,  $n = 1, 2, \dots$ , the following two relationships are equivalent:*

$$\pi_n \xrightarrow{w} \pi, \quad (1.22)$$

$$p_n \xrightarrow{w} p. \quad (1.23)$$

To prove Theorem 5 we will make use of the following Theorem:

**Theorem 6** *Let  $\mathcal{U}$  be a class of sets such that (i)  $\mathcal{U}$  is closed under the formation of finite intersections and (ii) for every  $a$  in  $\Omega^\infty$  and every positive  $\epsilon$  there is an  $A$  in  $\mathcal{U}$  with  $a \in A^\circ \subset A \subset B_\epsilon(a)$ . If  $\Omega^\infty$  is separable and if  $\pi_n(A) \rightarrow \pi(A)$  for every  $A$  in  $\mathcal{U}$ , then  $\pi_n \xrightarrow{w} \pi$ .*

**Proof:** In Billingsley [Bil71], pp.14-15. •

**Proof of Theorem 5:** First we prove  $\pi_n \xrightarrow{w} \pi \Rightarrow p_n \xrightarrow{w} p$ . By definition

$$\pi_n \xrightarrow{w} \pi \Rightarrow$$

(because  $A(a_1 \dots a_m)$  are always both open and closed, their boundary is the empty set hence it has measure zero - then the following is true by definition)

$$\forall m \geq 1, a_1, \dots, a_m \in \Omega \quad \pi_n(A(a_1 \dots a_m)) \rightarrow \pi(A(a_1 \dots a_m)) \Rightarrow$$

$$\forall m \geq 1, a_1, \dots, a_m \in \Omega \quad p_n(a_1 \dots a_m) \rightarrow p(a_1 \dots a_m) \Rightarrow$$

$$p_n \xrightarrow{w} p. \tag{1.24}$$

Next we prove  $p_n \xrightarrow{w} p \Rightarrow \pi_n \xrightarrow{w} \pi$ . We will use Theorem 6; to apply Theorem 6 we need the following ingredients:

1. A class of sets  $\mathcal{U}$  s.t.  $\mathcal{U}$  is closed under finite intersections and  $\forall \epsilon > 0, a \in \Omega^\infty \exists A \in \mathcal{U}$  s.t.

$$a \in A^\circ \subset A \subset B_\epsilon(a). \tag{1.25}$$

2.  $\Omega^\infty$  is separable.

3.  $\pi_n(A) \rightarrow \pi(A) \forall A \in \mathcal{U}$ .

For (1), note we can identify  $\mathcal{U}$  with  $\mathcal{A}(\Omega)$ . It is easy to show that  $\mathcal{A}(\Omega)$  is closed under finite intersections and  $\forall a \in \Omega^\infty, \epsilon > 0 \exists m, b_1 \dots b_m, A = A(b_1 \dots b_m)$ , such that (1.25) holds.

For (2), consider the set  $\mathcal{W} \doteq \{a : \text{for some } n \ a = a_1 a_2 \dots a_n 0000 \dots\}$ .  $\mathcal{W}$  is countable and  $\forall b \in \Omega^\infty, \epsilon > 0 \exists a \in \mathcal{W}$  s.t.  $\rho(a, b) < \epsilon$ . This implies  $\mathcal{W}$  is a countable dense subset of  $\Omega^\infty$ , so  $\Omega^\infty$  is separable.

For (3), take any  $A \in \mathcal{A}(\Omega)$ . Then  $A = A(a_1 \dots a_m)$  for some  $m, a_1, \dots, a_m$ . Then

$$\pi(A) = p(a_1 \dots a_m), \tag{1.26}$$

$$\pi_n(A) = p_n(a_1 \dots a_m). \tag{1.27}$$

Now, by definition

$$\begin{aligned}
p_n &\xrightarrow{w} p \Rightarrow \\
\forall m \geq 1, a_1, \dots, a_m \in \Omega &\quad p_n(a_1 \dots a_m) \rightarrow p(a_1 \dots a_m) \Rightarrow \\
\forall m \geq 1, a_1, \dots, a_m \in \Omega &\quad \pi_n(A(a_1 \dots a_m)) \rightarrow \pi(A(a_1 \dots a_m)) \Rightarrow \\
\forall A \in \mathcal{A}(\Omega) &\quad \pi_n(A) \rightarrow \pi(A) \Rightarrow \\
\pi_n &\xrightarrow{w} \pi; \tag{1.28}
\end{aligned}$$

the last step follows from Theorem 4. This completes the proof. •

In the metric space  $(\Omega^\infty, \rho)$  we can define continuity and uniform continuity in the usual way that this is done for metric spaces. In fact, it is easy to check that the following, simpler definitions are equivalent to the usual definitions in terms of metric spaces.

**Definition 18** *A function  $F : \Omega^\infty \mapsto \mathbf{R}$  will be called **continuous** iff  $\forall \epsilon > 0$  and  $\forall a = a_1 a_2 \dots \in \Omega^\infty \exists n$  such that  $\forall b = b_1 b_2 \dots \in \Omega^\infty$  satisfying  $b_1 = a_1, \dots, b_n = a_n$  we have*

$$|F(a) - F(b)| < \epsilon. \tag{1.29}$$

**Definition 19** *A function  $F : \Omega^\infty \mapsto \mathbf{R}$  will be called **uniformly continuous** iff  $\forall \epsilon > 0 \exists n$  such that  $\forall a = a_1 a_2 \dots, b = b_1 b_2 \dots \in \Omega^\infty$  satisfying  $b_1 = a_1, \dots, b_n = a_n$  we have*

$$|F(a) - F(b)| < \epsilon. \tag{1.30}$$

### 1.3 Preliminaries about Hidden Markov Models

In this section we present definitions, notation and basic facts about Markov processes and Hidden Markov Models that we will use repeatedly in later chapters. Most of the definitions can be skipped for the time being and read when needed. However the reader should study now the equivalence theorem for the various types of HMM's (Theorem 9).

**Definition 20** A (first order) **Markov Process** is a stochastic process  $X$  that has the following property:

$$\forall m \geq 1 \quad Pr(X_t | X_{t-1}, X_{t-2}, \dots, X_{t-m}) = Pr(X_t | X_{t-1}). \quad (1.31)$$

**Definition 21** An  $N$ -th order **Markov Process** is a stochastic process  $X$  that has the following property:

$$\forall m \geq N \quad Pr(X_t | X_{t-1}, X_{t-2}, \dots, X_{t-m}) = Pr(X_t | X_{t-1}, \dots, X_{t-N}). \quad (1.32)$$

If  $X$  has a finite alphabet  $\Omega$ , then we have the following

**Definition 22** Given a stationary Markov process  $X$  over finite alphabet  $\Omega$  we define its **probability transition matrix**  $P$  by

$$\forall x, z \in \Omega \quad P_{x,z} \doteq Pr(X_t = z | X_{t-1} = x). \quad (1.33)$$

**Definition 23** Given a Markov probability transition matrix  $P_{x,z}$ ,  $x, z \in \Omega = \{0, 1, \dots, K-1\}$ , we say that  $z$  is a **consequent** of  $x$ , denoted by  $x \rightsquigarrow z$ , if there is some integer  $n$  such that  $P_{x,z}^n > 0$ .

**Definition 24** Given a stationary Markov process  $X$  with probability transition matrix  $P_{x,z}$ ,  $x, z \in \Omega = \{0, 1, \dots, K-1\}$ , we will define an **ergodic class** of  $P$  to be a set  $\Omega' \subset \Omega$  such that:

1.  $\forall x, z \in \Omega', x \rightsquigarrow z$  and  $z \rightsquigarrow x$ .
2.  $\forall x \in \Omega$ , if there is  $z \in \Omega'$  such that  $x \rightsquigarrow z$  and  $z \rightsquigarrow x$ , then  $x \in \Omega'$ .

If  $x, z$  are in the same ergodic class, we write  $x \sim z$ .

**Proof:** See Doob [Doo53]. •

**Definition 25** Given a transition probability matrix  $P_{x,z}$ ,  $x, z \in \Omega$  and a probability  $p(x)$ ,  $x \in \Omega$ , we say  $p$  is a **stationary probability** of  $P$  iff

$$\forall x \in \Omega \quad p(x) = \sum_{z \in \Omega} p(z)P_{z,x}. \quad (1.34)$$

**Theorem 7** Given a transition probability matrix  $P_{x,z}$ ,  $x, z \in \Omega$ , if  $P$  has exactly one ergodic class, then it has exactly one stationary probability  $p(x)$ ,  $x \in \Omega$ . Hence, there is exactly one stationary Markov process which has transition matrix  $P$ . This process is ergodic. Conversely, if  $X$  is a stationary ergodic Markov process with transition probability matrix  $P$ , then  $P$  has exactly one ergodic class.

**Proof:** See Doob [Doo53]. •

**Definition 26** We call a Markov process  $X$  with transition matrix  $P_{x,z}$ ,  $x, z \in \Omega$ , **regular** iff there is some  $n$  s.t.  $\forall x, z \in \Omega P_{x,z}^n > 0$ . In that case  $P$  is also called regular.

**Theorem 8** If a Markov transition matrix  $P$  is regular then it has exactly one ergodic class. Furthermore, there is exactly one stationary Markov process  $X$  (with probability function  $p$ ) having transition matrix  $P$ . This process is also ergodic. The limit  $\lim_{n \rightarrow \infty} P_{x,z}^n$  exists for all  $x, z \in \Omega$  and we have  $p(z) = \lim_{n \rightarrow \infty} P_{x,z}^n$ , for all  $x, z \in \Omega$ .

**Proof:** See Billingsley [Bil79]. •

If the stationary Markov process  $p$  is regular, then, for all  $n$  and all  $x_0, x_1, \dots, x_n \in \Omega$  we have:

$$p(x_1 \dots x_n) = \left( \lim_{n \rightarrow \infty} P_{x_0, x_1}^n \right) \cdot P_{x_1, x_2} \cdot P_{x_2, x_3} \cdot \dots \cdot P_{x_{n-1}, x_n}; \quad (1.35)$$

therefore the process is uniquely determined by  $P$ . In particular, there is a case when it is easy to establish regularity: If the Markov process  $X$  (over finite alphabet  $\Omega$ ) has transition matrix  $P$  such that  $P_{x,z} > 0 \forall x, z \in \Omega$  then  $X$  is a regular Markov Chain.

There are several types of Hidden Markov Models; they always involve a pair of stochastic processes  $(X, Y)$ . The **hidden** or **state** process  $X$  is Markov and the **observable** process  $Y$  depends *instantaneously* on  $X$ . We will distinguish four different types of dependence, which have all appeared in the literature and all deserve equally the name “Hidden Markov Model”.

**Definition 27** A **Deterministic Node Model** (DNM) is a pair of stochastic processes  $(X, Y)$  (over finite alphabets  $\Omega_X, \Omega_Y$  respectively) such that  $X$  is a Markov process and there is a deterministic function  $f : \Omega_X \mapsto \Omega_Y$  such that for all  $t$

$$Y_t = f(X_t). \quad (1.36)$$

**Definition 28** A **Deterministic Arc Model** (DAM) is a pair of stochastic processes  $(X, Y)$  (over finite alphabets  $\Omega_X, \Omega_Y$  respectively) such that  $X$  is a Markov process and there is a deterministic function  $f : \Omega_X^2 \mapsto \Omega_Y$  such that for all  $t$

$$Y_t = f(X_t, X_{t+1}). \quad (1.37)$$

**Definition 29** A **Stochastic Node Model** (SNM) is a pair of stochastic processes  $(X, Y)$  (over

finite alphabets  $\Omega, \Omega_Y$  respectively) such that  $X$  is a Markov process and for all  $t, m$

$$Pr(X_{t+1}, \dots, X_{t+m}, Y_{t+1}, \dots, Y_{t+m}) =$$

$$Pr(X_{t+1}) \cdot Pr(X_{t+2}|X_{t+1}) \cdot \dots \cdot Pr(X_{t+m}|X_{t+m-1}) \cdot Pr(Y_{t+1}|X_{t+1}) \cdot \dots \cdot Pr(Y_{t+m}|X_{t+m}). \quad (1.38)$$

**Definition 30** A **Stochastic Arc Model** (SAM) is a pair of stochastic processes  $(X, Y)$  (over finite alphabets  $\Omega_X, \Omega_Y$  respectively) such that  $X$  is a Markov process and for all  $t, m$

$$Pr(X_{t+1}, \dots, X_{t+m}, X_{t+m+1}, Y_{t+1}, \dots, Y_{t+m}) =$$

$$Pr(X_{t+1}) \cdot Pr(X_{t+2}|X_{t+1}) \cdot \dots \cdot Pr(X_{t+m+1}|X_{t+m})$$

$$\cdot Pr(Y_{t+1}|X_{t+1}, X_{t+2}) \cdot \dots \cdot Pr(Y_{t+m}|X_{t+m}, X_{t+m+1}). \quad (1.39)$$

By definition, any of the above types of models is a HMM:

**Definition 31** A **Hidden Markov Model** (HMM) is a pair of processes  $(X, Y)$  that is any of the following: DNM, DAM, SNM, SAM.

**Remark:** It must be pointed out that HMM do *not* have the Markov property. That is, in general

$$Pr(Y_t = y_0 | Y_{t-1} = y_{-1}, Y_{t-2} = y_{-2}, \dots) \neq Pr(Y_t = y_0 | Y_{t-1} = y_{-1}). \quad (1.40)$$

This means that, in a certain sense, HMMs are more general models than Markov processes, because they retain a *memory* of their infinite past. Also, in many cases HMMs have a *uniform mixing* (i.e. forgetting) property; this will prove to be very important in Chapter 3.

The model mostly considered by mathematicians is DNM (Gilbert [Gil59], Heller [Hel65], Dharmadikari [Dha63a, Dha63b], Rosenblatt [Ros59, Ros71]). DAM's are considered in Automata

Theory [Boo67]. Statisticians have paid special attention to DNM and SNM (Baum et al. [BP66, BE67, BS68, B+70], Petrie [Pet69]). In speech modeling both SNM (Levinson [L+83], Lee [L+90]) and SAM (Bahl [B+83]) are used. However all of these models are equivalent in the following sense:

**Theorem 9** (*Equivalence*) *Given any HMM  $(X, Y)$  where  $X$  is a Markov process, there exist a DNM  $(X^1, Y^1)$ , a DAM  $(X^2, Y^2)$ , a SNM  $(X^3, Y^3)$  and a SAM  $(X^4, Y^4)$  such that  $Y^1, \dots, Y^4$  all have the same probability function as  $Y$ .*

**Proof:** We prove the theorem in the following way:

1. We show that for every DNM  $(X, Y)$  there is a DAM  $(U, V)$  such that  $Y$  and  $V$  have the same probability function.
2. We show that for every DAM  $(X, Y)$  there is a SAM  $(U, V)$  such that  $Y$  and  $V$  have the same probability function.
3. We show that for every SAM  $(X, Y)$  there is a SNM  $(U, V)$  such that  $Y$  and  $V$  have the same probability function.
4. We show that for every SNM  $(X, Y)$  there is a DNM  $(U, V)$  such that  $Y$  and  $V$  have the same probability function.

1. Take  $(X, Y)$  to be a DNM with

$$Y_t = f(X_t) \quad f : \Omega_X \mapsto \Omega_Y. \quad (1.41)$$

Now define  $U_t = X_t$ ,  $V_t = Y_t$  and note that indeed (a)  $U_t$  is Markov and (b)  $V_t$  is a function of  $(U_t, U_{t+1})$ . In particular  $V_t = \hat{f}(U_t, U_{t+1}) \doteq f(U_t)$ .

2. Take  $(X, Y)$  to be a DAM with

$$Y_t = f(X_t, X_{t+1}) \quad f : \Omega_X^2 \mapsto \Omega_Y. \quad (1.42)$$

Now define  $U_t = X_t$ ,  $V_t = Y_t$  and note that indeed (a)  $U_t$  is Markov and (b)  $V_t$  satisfies for all  $t, m$

$$\begin{aligned} & Pr(U_{t+1}, \dots, U_{t+m}, U_{t+m+1}, V_{t+1}, \dots, V_{t+m}) = \\ & Pr(U_{t+1}) \cdot Pr(U_{t+2}|U_{t+1}) \cdot \dots \cdot Pr(U_{t+m+1}|U_{t+m}) \\ & \cdot Pr(V_{t+1}|U_{t+1}, U_{t+2}) \cdot \dots \cdot Pr(V_{t+m}|U_{t+m}, U_{t+m+1}). \end{aligned} \quad (1.43)$$

3. Take  $(X, Y)$  to be a SAM and define  $U_t = (X_t, X_{t+1})$ ,  $V_t = Y_t$  and note that indeed (a)  $U_t$  is Markov and (b)  $V_t$  satisfies for all  $t, m$

$$\begin{aligned} & Pr(U_{t+1}, \dots, U_{t+m}, V_{t+1}, \dots, V_{t+m}) = \\ & Pr(U_{t+1}) \cdot Pr(U_{t+2}|U_{t+1}) \cdot \dots \cdot Pr(U_{t+m}|U_{t+m-1}) \\ & \cdot Pr(V_{t+1}|U_{t+1}) \cdot \dots \cdot Pr(V_{t+m}|U_{t+m}). \end{aligned} \quad (1.44)$$

4. Take  $(X, Y)$  to be a SNM and define  $U_t = (X_t, Y_t)$ ,  $V_t = Y_t$ . Note that indeed (a)  $U_t$  is Markov and (b)  $V_t = \hat{f}(U_t) = \hat{f}(X_t, Y_t) \doteq Y_t$ .

This completes the proof of the theorem. •

**Remark:** The proof of Theorem 9 consists of repeated conversions from an original HMM to a new HMM of different type. It is easy to check that, in every case, if the original hidden process is stationary (respectively stationary ergodic) then the original observable process, the new hidden process and the new observable process are all stationary (respectively stationary ergodic).

In the sense of the previous theorem, all HMM's are equivalent. In the rest of this thesis we will use the type of models most convenient for the particular application at hand; the results can be generalized to all types of HMM's. In particular, we can prove all our results using only DNM's and SNM's. Further we can consider DNM's as a special case of SNM's as will be obvious from the following definition:

**Definition 32** For every SNM  $(X, Y)$ ,  $X$  having alphabet  $\Omega_X$ ,  $Y$  having alphabet  $\Omega_Y$ , define the **transition probability matrix**  $P$  by

$$\forall x, z \in \Omega_X \quad P_{x,z} \doteq Pr(X_{t+1} = z | X_t = x) \quad (1.45)$$

and define the **emission probability matrix**  $Q$  by

$$\forall x \in \Omega_X y \in \Omega_Y \quad P_{x,z} \doteq Pr(Y_t = y | X_t = x). \quad (1.46)$$

**Remark:** For any SNM  $(X, Y)$  such that  $P$  is regular, it is obvious that the pair  $(P, Q)$  completely defines  $(X, Y)$ .

**Remark:** It is also obvious that a DNM is a special case of a SNM where the elements of matrix  $Q$  can be only 1's or 0's.

## 1.4 Preliminaries about Hidden Gibbs Models

In this section we present definitions, notation and basic facts about Hidden Gibbs Models that we will use repeatedly in later chapters. Most of the definitions can be skipped for the time being and read when needed.

A Hidden Gibbs Model is a pair of processes  $(X, Y)$ , where  $X$  is a Gibbs process (equivalently a Markov Random Field) and  $Y$  is a function of  $X$ :  $Y_t = f(X_t)$ . We start by defining Gibbs Processes / Markov Random Fields.

**Definition 33** A **graph** is a pair  $(T, \mathcal{N})$ , where  $T$  is a (finite or infinite) countable set called the **node set** and  $\mathcal{N}$  is the **neighborhood set** defined by

$$\mathcal{N} \doteq \{N(t), t \in T\}, \quad N(t) \doteq \{s : s \leftrightarrow t\}. \quad (1.47)$$

Here  $\leftrightarrow$  is the **neighborhood** relationship that satisfies:  $s \leftrightarrow t \Leftrightarrow t \leftrightarrow s$ .

**Definition 34** A **Markov Random Field** on a graph  $(T, \mathcal{N})$  is a stochastic process  $\{X_t\}_{t \in T}$  over a finite alphabet  $\Omega$  such that:

1.  $\forall n, \forall t_1, \dots, t_n \in T, \forall x_1, \dots, x_n \in \Omega \quad \Pr(X_{t_1} \dots X_{t_n} = x_1 \dots x_n) > 0$ .
2.  $\Pr(X_t | X_s, s \neq t) = \Pr(X_t | X_{N(t)})$ ,

**Definition 35** A **potential**  $\mathcal{U}$  on a graph  $(T, \mathcal{N})$  is a family  $\{U_A, A \subset T\}$  of functions from  $\Omega^T$  to  $\mathbf{R}$  (the real numbers), with the property that  $U_A(x) = U_A(y)$  whenever  $x_t = y_t$  for all  $t \in A$ , and such that  $U_\emptyset = 0$ . The **energy**  $H$  of the potential  $\mathcal{U}$  is given by

$$H_{\mathcal{U}} = \sum_{A \subset T} U_A. \quad (1.48)$$

**Definition 36** A **clique** on a graph  $(T, \mathcal{N})$  is a set  $C \subset T$  such that  $s, t \in C \Rightarrow s \leftrightarrow t$ .

Denote by  $\mathcal{C}$  the set of all cliques in a graph.

**Definition 37** A potential  $\mathcal{U}$  on a graph  $(T, \mathcal{N})$  is called a **neighbor potential** (with respect to  $(T, \mathcal{N})$ ) iff  $U_A = 0$  whenever  $A \notin \mathcal{C}$ .

**Definition 38** A stochastic process  $\{X_t\}_{t \in T}$ , taking values in  $\Omega$ , is **Gibbs** on a graph  $(T, \mathcal{N})$  iff there exists a neighbor potential  $\mathcal{U}$  such that:  $\forall x = \{x_t\}_{t \in T}$ , and  $\forall t \in T, \forall n$  and  $t_1, \dots, t_n \in T$  satisfying (a)  $t_k \neq t$  (for  $k = 1, \dots, n$ ) and (b)  $N(t) \subset \{t_1, \dots, t_n\}$ , we have

$$\Pr(X_t = x_0 | X_{t_1} = x_1, \dots, X_{t_n} = x_n) = \frac{e^{-\sum_{C: C \in \mathcal{C}, t \in C} U_C(x)}}{Z} \quad (1.49)$$

where  $Z$  is a normalizing constant.

**Theorem 10** A stochastic process  $X$  is Gibbs on  $(T, \mathcal{N})$  iff it is a Markov Random Field on  $(T, \mathcal{N})$ .

**Proof:** See Griffeath, Theorem 12-21, in [SKK76]. •

**Theorem 11** Consider the following graph  $(T, \mathcal{N})$ :  $T = \{\dots, -1, 0, 1, \dots\}$ ,  $N(t) = \{t - 1, t + 1\} \forall t \in T$ . Any positive Markov process  $\{X_t\}_{t \in T}$  is a MRF on  $(T, \mathcal{N})$  and any MRF on  $(T, \mathcal{N})$  is a positive Markov process.

**Proof:** See Griffeath, Propositions 12-4 and 12-32, in Kemeny and Snell [SKK76]. •

**Definition 39** A **Hidden Gibbs Model** is a pair of processes  $(X, Y)$ , where  $X$  is a Gibbs process (equivalently MRF) on a graph  $(T, \mathcal{N})$  and  $Y$  satisfies:  $Y_t = f(X_t)$  for all  $t$ .

(It is obvious how to extend this definition for other types of HGM's, e.g. with stochastic emission, but this will not be necessary for our work.)

## 1.5 Related work

Hidden Markov Models became popular after their successful application to Speech Recognition in the late seventies and eighties, [JR85, B+83, L+83, L+90, R+83, R+85, Rab88, Bro87]. The most commonly used type of HMM's in speech is SNM [R+83]. Subsequently, HMM's have been used in modelling a number of other types of stochastic processes (shape recognition [KH89a, KH89c, MK91] arterial modelling [GM90] etc. An area where HMM's are widely used (under the names of *aggregated Markov Chains* or *compartmental models*) is biological modelling. The relevant theory is expounded in [FR85, FR87]. For a large bibliography of biological applications see [Kie89].

The great popularity of HMM's is due at least partly to the fact that their Maximum Likelihood parameters can be computed very efficiently. In particular, for Maximum Likelihood estimation Baum has developed the powerful Backward Forward algorithm which converges extremely fast. The relevant theory is developed in [BP66, BE67, BS68, Pet69, B+70]. This algorithm is a special case of the EM algorithm [D+77, Wu83].

Some of the mathematical theory of HMM's was developed in the late fifties and sixties. In particular, Gilbert [Gil59] and Blackwell [BK57] first posed the problem of identifiability of Markov processes. A little later, Kemeny and Snell discussed in [KS60] Markov processes such that groups of states can be *lumped* together and the resulting process still retains the Markov conditioning property. In other words, consider the Markov process  $X_t$  taking values in  $\Omega_X$ , take  $\Omega_Y \subset \Omega_X$ , and define the process  $Y_t = f(X_t)$  where  $f : \Omega_X \mapsto \Omega_Y$ . It is possible that  $X_t, f(\cdot)$  are such that  $Y_t$  is a Markov process. Such processes are called *lumpable* Markov processes; they are a special case of DNM's. Kemeny and Snell presented sufficient conditions that a Markov process is lumpable; but not every Markov process satisfies these conditions. A little later Dharmadikari [Dha63a] posed the following question: what stochastic processes can be represented as deterministic functions of Markov processes (in our terminology as DNM's)? He obtained sufficient conditions that this be true [Dha63b]. This led to further intense activity in the early sixties [BR58, Ros59, Hac63] which culminated in the discovery by Heller [Hel65] of *necessary and sufficient* conditions that a process must satisfy to be a DNM. Rosenblatt presents these results and other related work in [Ros71]. Later refinements and further elaboration appears in [Ley67, Fri67, Dha68, Eri70, W+74, Bos75, BT77, RP81].

We should also mention a result of Harris [Har55] that establishes conditions on a stationary stochastic process that are sufficient for it to be exactly representable as the observable of an *infinite state space* Markov process. A preliminary version of this result was anticipated in a paper by Doeblin and Fortet [DF37].

However, all of these results are about *exact representations* and apply only to limited classes of processes that satisfy rather restrictive conditions. Until recently it was still not known what class of stochastic processes can be *approximated* by HM processes. (However a very relevant result about approximation of processes by  $N$ -order Markov processes is described in [OW90].) Thus the first goal of this dissertation is to show that the class of HMM's is dense in the class of stationary ergodic stochastic processes.

We should note that there is an ongoing effort to apply HMM concepts to Artificial Neural

Networks. See [Bw88, BW89, Bri89] for examples of applications and [Keh90a, Keh90b, Keh91a, Keh91b, KH89a, KH89b, KH89c] for the relevant theory. Indeed, viewed in the appropriate way, Artificial Neural Networks *are* HMM. Therefore, the development of the theory of HMM's will have implications for ANN's theory as well. Similarly, HMM's have a close connection with stochastic finite state automata. Arbib points out the connection in [Arb66]. For more details see [Paz71, Boo67] which contain a voluminous bibliography on stochastic automata. In [Keh91a, Keh91b] *networks* of such finite state stochastic automata are considered and an algorithm is developed for estimating their parameters. Also prediction and classification algorithms are developed and applied to modelling of speech data. This work is reported in [Keh91a] and is strongly influenced by [HL69a] and [HL69b].

There is a lot of empirical work done on the question of estimating HMM's. However, the underlying theory is not very well developed. The previously mentioned work by Baum and collaborators develops the theory of a particular maximization technique (namely the BF algorithm). Also the question of consistency has so far been treated only for the case of estimating processes that actually *are* HMM's [BP66]. In this dissertation we show that consistent estimation is possible for a much larger class of processes.

## Chapter 2

# Representation

In this chapter we deal with the question of finding HMM's that reproduce a given stochastic process. In Section 2.1 we construct a sequence of Hidden Markov Models that approximate any given stationary ergodic process. In Section 2.2 we construct a sequence of Hidden Gibbs Models that approximate any given stationary ergodic process. In Section 2.3 we for once depart from the default assumption of finite state space HMM's and construct an infinite state HMM that reproduces exactly any stationary ergodic, positive stochastic process with continuous conditional probability function.

### 2.1 Approximation with Hidden Markov Models

In this and the following sections we will make use of a special type of HMM's which we call *autoregressive* models.

**Definition 40** *Given a stationary stochastic process  $Y$ , with finite alphabet  $\Omega_Y = \{0, 1, \dots, K-1\}$  and probability function  $p$ , an **N-th order autoregressive model** of  $Y$  (for short *AR model*) is a DNM  $(X^N, Y^N)$  that satisfies the following:*

1. *The observable alphabet  $\Omega_Y$  is the same as the alphabet of  $Y$ .*

2. The state alphabet  $\Omega_X = \Omega_Y^N$ . Hence we can write  $X_t^N = X_{t,1}^N \dots X_{t,N}^N$ , where for all  $N \geq 1$ ,  $t = 0, \pm 1, \pm 2, \dots$ ,  $n = 1, \dots, N$  we have  $X_{t,n}^N \in \Omega_Y$ .

3.  $X^N$  is stationary.

4.  $X^N$  determines  $Y^N$  by  $Y_t^N = f(X_t^N)$ ,  $t = 0, \pm 1, \pm 2, \dots$ . The "hiding function"  $f : \Omega_X \mapsto \Omega_Y$  is defined by:  $f(X_{t,1}^N \dots X_{t,N}^N) = X_{t,1}^N$ . (Hence, as  $X^N$  is stationary,  $Y^N$  is stationary as well.)

5.  $X^N$  has probability transition matrix  $P_N$ . To define  $P_N$  we need some auxiliary definitions.

First, partition  $\Omega_X$  into two sets:

$$\Omega_+ \doteq \{x \in \Omega_X : p(x) > 0\}, \quad (2.1)$$

$$\Omega_0 \doteq \{x \in \Omega_X : p(x) = 0\}. \quad (2.2)$$

Second, for all  $x \in \Omega_Y^N$ ,  $x = x_1 \dots x_N$ , with  $x_1, \dots, x_N \in \Omega_Y = \{0, 1, \dots, K-1\}$ , define the order number  $n(x)$  as follows:

$$n(x) \doteq x_1 + K \cdot x_2 + K^2 \cdot x_3 + \dots + K^{N-1} \cdot x_N. \quad (2.3)$$

Now define  $[P_N]_{x,z} \forall x = x_1 \dots x_N \in \Omega_+, z = z_1 \dots z_N \in \Omega_X$  by

$$[P_N]_{x_1 \dots x_N, z_1 \dots z_N} \doteq \begin{cases} \frac{p(x_1 \dots x_N z_N)}{p(x_1 \dots x_N)} & \text{when } x_2 = z_1, \dots, x_N = z_{N-1} \\ 0 & \text{otherwise.} \end{cases} \quad (2.4)$$

Denote by  $x_0$  the element of  $\Omega_+$  with smallest order number, and define  $[P_N]_{x,z} \forall x \in \Omega_0, z \in \Omega_X$  by

$$[P_N]_{x,x_0} = 1, \quad (2.5)$$

$$[P_N]_{x,z} = 0 \quad \forall z \neq x_0. \quad (2.6)$$

**Remark:** Here is an intuitive explanation of autoregressive HMMs. Essentially, the  $Y^N$  process is a  $N$ -th order Markov Chain obtained from a HMM. Every state in the hidden process is a sequence of  $N$  characters from the alphabet  $\Omega_Y$ . The only state transitions possible are shifts: state  $y_1 \dots y_N$  can only move to states  $y_2 \dots y_N y$ , where  $y$  is any character from  $\Omega_Y$ . These transitions take place with conditional probability  $Prob(Y_t = y | Y_{t-1} = y_N, \dots, Y_{t-N} = y_1)$ , i.e. they are conditioned the  $N$ -distant past of the original process. Finally, the observation mechanism ensures that the original process and the observable process of the  $N$ -th order autoregressive model have the same  $N + 1$  order marginals. This will be proven in Lemma 12.

**Remark:** Given a stationary stochastic process  $Y$  taking values in  $\Omega_Y$ , it is not immediately obvious from Definition 40 that an autoregressive model of order  $N$  exists at all; if it exists it is not clear that it is unique. In fact, using a compactness argument it is easy to show that for every  $N$  there is *at least one* autoregressive model. Uniqueness cannot be proven always, but holds true under special conditions, as in Theorems 13 and 14.

**Remark:** The models  $(X^N, Y^N)$  are called autoregressive because of the similarity with AR-models for real-valued stochastic processes [Aok87]. They are closely related to the *canonical  $N$ -order Markov representation* of Ornstein [OW90].

Before we prove these Theorems 2.1 and 2.2, we need the following Lemma, that sums up the properties of AR models.

**Lemma 12** *Given a stationary stochastic process  $Y$ , with finite alphabet  $\Omega_Y$  and probability function  $p$ , assume  $Y$  has exactly one  $N$ -th order AR model, call it  $(X^N, Y^N)$ . Call  $P_N$  the transition matrix of  $X^N$  and  $p_N$  the probability function of  $Y^N$ . Then:*

1.  $[p(y_1 \dots y_N)]_{y_1, \dots, y_N \in \Omega_Y}$  is the unique stationary probability of  $P_N$ . That is, for all  $y_1, \dots, y_N$  in  $\Omega_Y$

$$p(y_1 \dots y_N) = \sum_{z_1, \dots, z_N \in \Omega_Y} p(z_1 \dots z_N) [P_N]_{z_1 \dots z_N, y_1 \dots y_N}. \quad (2.7)$$

It follows that for all  $t = 0, \pm 1, \pm 2, \dots$  and for all  $y_1, \dots, y_N \in \Omega_Y$

$$Pr(X_t^N = y_1 \dots y_N) = p(y_1 \dots y_N). \quad (2.8)$$

2. For all  $M \geq N$ ,  $y_1, \dots, y_M \in \Omega_Y$ ,  $t = 0, \pm 1, \dots$

$$p_N(y_1 \dots y_M) = Pr(X_{t+1}^N = y_1 \dots y_N, \dots, X_{t+M-N+1}^N = y_{M-N+1} \dots y_M). \quad (2.9)$$

3. For all  $y_0, \dots, y_N \in \Omega_Y$

$$p_N(y_0 \dots y_N) = p(y_0 \dots y_N). \quad (2.10)$$

4. For all  $M \geq N$ ,  $y_1, \dots, y_M \in \Omega_Y$

$$p(y_1 \dots y_M) > 0 \Rightarrow p_N(y_1 \dots y_M) > 0. \quad (2.11)$$

5. For all  $M \geq N$ ,  $y_{-M}, \dots, y_0 \in \Omega_Y$  such that  $p_N(y_{-M} \dots y_{-1}) > 0$

$$p_N(y_0 | y_{-1} \dots y_{-M}) = p_N(y_0 | y_{-1} \dots y_{-N}). \quad (2.12)$$

6. For all  $M \geq N$   $y_{-M}, \dots, y_0 \in \Omega_Y$  such that  $p(y_{-M} \dots y_{-1}) > 0$

$$p_N(y_0 | y_{-1} \dots y_{-M}) = p(y_0 | y_{-1} \dots y_{-N}). \quad (2.13)$$

**Proof:**

1. For all  $y_1 \dots y_N \in \Omega_X$ , we have

$$\sum_{z_1, \dots, z_N \in \Omega_X} p(z_1 \dots z_N) [P_N]_{z_1 \dots z_N, y_1 \dots y_N} =$$

$$\begin{aligned}
& \sum_{z_1, \dots, z_N \in \Omega_+} p(z_1 \dots z_N) [P_N]_{z_1 \dots z_N, y_1 \dots y_N} = \\
& \sum_{z_1, \dots, z_N \in \Omega_+ : z_2 = y_1, \dots, z_N = y_{N-1}} p(z_1 \dots z_N) [P_N]_{z_1 \dots z_N, y_1 \dots y_N} = \\
& \sum_{z_1 \in \Omega_X : z_1 y_1 \dots y_{N-1} \in \Omega_+} p(z_1 y_1 \dots y_{N-1}) \frac{p(z_1 y_1 \dots y_N)}{p(z_1 y_1 \dots y_{N-1})} = p(y_1 \dots y_N). \tag{2.14}
\end{aligned}$$

So we are done.

2. For all  $M > N$ ,  $y_1, \dots, y_N \in \Omega_Y$ , we have

$$p_N(y_1 \dots y_M) = \sum_{z_1, \dots, z_M \in \Omega_Y^{N-1}} Pr(X_{t+1}^N = y_1 z_1) [P_N]_{y_1 z_1, y_2 z_2} \cdot \dots \cdot [P_N]_{y_{M-1} z_{M-1}, y_M z_M}. \tag{2.15}$$

We will now remove from the sum in (2.15) some summands (products) which equal 0, without changing the result.

First,  $Pr(X_{t+1}^N = y_1 z_1) = p(y_1 z_1)$  by (2.8). If  $y_1 z_1$  is in  $\Omega_0$   $p(y_1 z_1) = 0$ . Hence, in (2.15) we only need to sum over  $z_1$  such that  $z_1 y_1 \in \Omega_+$ . Repeating this argument, we see that we also need to sum only over  $z_2, \dots, z_M$  such that  $y_2 z_2, \dots, y_M z_M$  are in  $\Omega_+$ .

Second, if  $y_k z_k$  ( $k = 2, \dots, M$ ) are in  $\Omega_+$ , the only state transitions with positive probability are the ones where  $y_k z_k$  is a shift of  $y_{k-1} z_{k-1}$ , that is  $z_{k-1} = y_k u_{k-1}$ ,  $u_{k-1} \in \Omega_Y^{N-2}$ . So we only need consider state transitions of the form

$$y_1 \dots y_N \rightarrow y_2 \dots y_{N+1} \rightarrow \dots \rightarrow y_{M-N+1} \dots y_M \rightarrow y_{M-N+2} \dots y_M w_1 \rightarrow \dots \rightarrow y_M w_1 \dots w_{N-1} \tag{2.16}$$

where  $w_1, \dots, w_{N-1}$  belong to  $\Omega_Y$  and must be such that for  $k = 2, \dots, N$ ,  $y_{M-N+k} \dots y_M w_1 \dots w_{k-1}$  is in  $\Omega_+$ . If there are no such  $w$ 's, then both sides of (2.15) equal 0 and we are done. Otherwise (2.15) reduces to

$$p_N(y_1 \dots y_M) = \sum_{w_1, \dots, w_M} Pr(X_{t+1}^N = y_1 \dots y_N, X_{t+2}^N = y_2 \dots y_{N+1}, \dots, X_{t+M}^N = y_M w_1 \dots w_{N-1}) \Rightarrow$$

$$p_N(y_1 \dots y_M) = Pr(X_{t+1}^N = y_1 \dots y_N, X_{t+2}^N = y_2 \dots y_{N+1}, \dots \cdot X_{t+M-N+1}^N = y_{M-N+1} \dots y_M) \quad (2.17)$$

and we are done.

3. Use (2.17) with  $M = N + 1$  to get

$$p_N(y_1 \dots y_{N+1}) = Pr(X_t^N = y_1 \dots y_N) [P_N]_{y_1 \dots y_N, y_2 \dots y_{N+1}} = p(y_1 \dots y_N) [P_N]_{y_1 \dots y_N, y_2 \dots y_{N+1}}. \quad (2.18)$$

Now we have two cases:

- (a) If  $p(y_1 \dots y_N) = 0$ , then (i)  $p(y_1 \dots y_{N+1}) = 0$  and (ii) (by (2.18))  $p_N(y_1 \dots y_{N+1}) = 0$ . So we are done.
- (b) If  $p(y_1 \dots y_N) > 0$ , then  $[P_N]_{y_1 \dots y_N, y_2 \dots y_{N+1}} = p(y_1 \dots y_{N+1}) / p(y_1 \dots y_N)$ . This, together with (2.18) implies  $p_N(y_1 \dots y_{N+1}) = p(y_1 \dots y_{N+1})$ . So we are done.

4. If  $p(y_1 \dots y_M) > 0$ , then we have

$$p(y_1 \dots y_N) > 0 \quad p(y_1 \dots y_{N+1}) > 0 \quad \Rightarrow [P_N]_{y_1 \dots y_N, y_2 \dots y_{N+1}} > 0, \quad (2.19)$$

...

$$p(y_1 \dots y_N) > 0 \quad p(y_1 \dots y_{N+1}) > 0 \quad \Rightarrow [P_N]_{y_1 \dots y_N, y_2 \dots y_{N+1}} > 0. \quad (2.20)$$

Combining (2.19), ..., (2.20) we get

$$Pr(X_1^N = y_1 \dots y_N, \dots, X_{M-N+1}^N = y_{M-N+1} \dots y_M) > 0 \Rightarrow p_N(y_1 \dots y_M) > 0 \quad (2.21)$$

and we are done.

5. As  $p_N(y_{-M}\dots y_{-1}) > 0$ , we have:

$$p_N(y_0|y_{-1}\dots y_{-M}) = \frac{p_N(y_{-M}\dots y_0)}{p_N(y_{-M}\dots y_{-1})} =$$

(by 2.)

$$\frac{Pr(X_{-M}^N = y_{-M}\dots y_{-M+N-1}, \dots, X_{-N+1}^N = y_{-N+1}\dots y_0)}{Pr(X_{-M}^N = y_{-M}\dots y_{-M+N-1}, \dots, X_{-N}^N = y_{-N}\dots y_{-1})} =$$

$$Pr(X_0^N = y_{-N+1}\dots y_0 | X_{-1}^N = y_{-N}\dots y_{-1}, \dots, X_{-M}^N = y_{-M}\dots y_{-M+N-1}) =$$

(by Markov property)

$$Pr(X_0^N = y_{-N+1}\dots y_0 | X_{-1}^N = y_{-N}\dots y_{-1}). \quad (2.22)$$

Also,  $p_N(y_{-M}\dots y_{-1}) > 0$  implies  $p_N(y_{-N}\dots y_{-0}) > 0$ . Repeating the previous argument we get:

$$p_N(y_0|y_{-1}\dots y_{-N}) = Pr(X_0^N = y_{-N+1}\dots y_0 | X_{-1}^N = y_{-N}\dots y_{-1}). \quad (2.23)$$

Combining (2.22) and (2.23) we get (2.12) and we are done.

6. From 4. and 5. it follows that

$$p(y_{-M}\dots y_{-1}) > 0 \Rightarrow p_N(y_0|y_{-1}\dots y_{-M}) = p_N(y_0|y_{-1}\dots y_{-N}). \quad (2.24)$$

From 3. it follows that

$$p(y_{-M}\dots y_{-1}) > 0 \Rightarrow p_N(y_0|y_{-1}\dots y_{-M}) = p(y_0|y_{-1}\dots y_{-N}). \quad (2.25)$$

Combining (2.24) and (2.25) we are done.

This completes the proof of the Lemma. •

Now we can prove an approximation theorem for *positive* stochastic processes.

**Theorem 13** *Given a stationary positive stochastic process  $Y$ , with finite alphabet  $\Omega_Y$  and probability function  $p$ , for every  $N$  there is exactly one autoregressive model  $(X^N, Y^N)$ . Furthermore the sequence of autoregressive models  $(X^N, Y^N)$   $N = 1, 2, \dots$  (with observable probability functions  $p_N$ ) is such that:*

1.  $X^N$  is stationary ergodic.
2.  $Y^N$  is stationary ergodic and positive.
3.  $Y^N \xrightarrow{w} Y$ .
4.  $H(Y^N; Y) \rightarrow 0$ .

**Proof:**  $P_N$  (the probability transition matrix associated with every  $N$ -th order autoregressive model of  $Y$ ) is regular. To prove this, compute  $P_N^N$  and note it is positive. Hence, by Theorem 8,  $P_N$  has exactly one ergodic class and there is exactly one stationary and ergodic Markov process with  $P_N$  as transition matrix. Call this process  $X^N$ . We write in the usual way  $X_t^N = X_{t,1}^N \dots X_{t,N}^N$  and define  $Y^N$  by  $Y_t^N = X_{t,1}^N$ . Clearly  $(X^N, Y^N)$  is the unique  $N$ -th order autoregressive model of  $Y$ . Now we proceed to prove  $(X^N, Y^N)$  satisfies the properties (1)-(4). As it is the unique  $N$ -th AR model of  $Y$  (for any  $N$ ), we can use Lemma 12.

As already noted, since  $P_N$  is regular,  $X^N$  is stationary ergodic by Theorem 1.8.  $Y^N$  is stationary ergodic because it is a function of  $X^N$ .  $Y^N$  is also positive, by Lemma 12, eq.(2.11) and the positivity of  $p$ .

Next, weak convergence follows immediately from (2.10) in Lemma 12. Finally, to prove cross entropy convergence, note that

$$\lim_N H(Y^N; Y) = \lim_N \lim_n H_n(Y^N; Y). \quad (2.26)$$

Now take any  $N \geq 1$ , any  $n \geq N$ :

$$H_n(Y^N; Y) = \sum_{y_{-n}, \dots, y_0 \in \Omega_Y} p(y_{-n} \dots y_0) \log \frac{p(y_0 | y_{-1} \dots y_{-n})}{p_N(y_0 | y_{-1} \dots y_{-n})} =$$

$$\sum_{y_{-n}, \dots, y_0} p(y_{-n} \dots y_0) \log p(y_0 | y_{-1} \dots y_{-n}) - \sum_{y_{-n}, \dots, y_0} p(y_{-n} \dots y_0) \log p_N(y_0 | y_{-1} \dots y_{-n}) =$$

(by (2.13) in Lemma 12 and positivity of  $p$ )

$$\begin{aligned} & \sum_{y_{-n}, \dots, y_0} p(y_{-n} \dots y_0) \log p(y_0 | y_{-1} \dots y_{-n}) - \sum_{y_{-n}, \dots, y_0} p(y_{-n} \dots y_0) \log p(y_0 | y_{-1} \dots y_{-N}) = \\ & \sum_{y_{-n}, \dots, y_0} p(y_{-n} \dots y_0) \log p(y_0 | y_{-1} \dots y_{-n}) - \sum_{y_{-N}, \dots, y_0} p(y_{-N} \dots y_0) \log p(y_0 | y_{-1} \dots y_{-N}) = \\ & -H_n(Y) + H_N(Y). \end{aligned} \tag{2.27}$$

So we have

$$\begin{aligned} \lim_N H(Y^N; Y) &= \lim_N \lim_n [-H_n(Y) + H_N(Y)] = \\ & -\lim_n H_n(Y) + \lim_N H_N(Y) = 0. \end{aligned} \tag{2.28}$$

This shows cross entropy convergence and completes the proof of the theorem. •

In the previous theorem we used several times the assumption that  $Y$  is a positive process. This is a rather strong assumption. Now we will substitute ergodicity for positivity and get essentially the same results.

**Theorem 14** *Given a stationary ergodic stochastic process  $Y$ , with finite alphabet  $\Omega_Y$  and probability function  $p$ , for every  $N$  there is exactly one autoregressive model  $(X^N, Y^N)$ . Furthermore the sequence of autoregressive models  $(X^N, Y^N)$   $N = 1, 2, \dots$  (with observable probability functions  $p_N$ ) is such that:*

1.  $X^N$  is stationary ergodic.
2.  $Y^N$  is stationary ergodic.
3.  $Y^N \xrightarrow{w} Y$ .
4.  $H(Y^N; Y) \rightarrow 0$ .

To prove Theorem 14 we need the following Lemma.

**Lemma 15** *Given a stationary ergodic stochastic process  $Y$ , with finite alphabet  $\Omega_Y$  and probability function  $p$ , and two strings  $x, z \in \Omega_Y^N$  such that  $p(x) > 0$ ,  $p(z) > 0$  there is a string  $w \in \Omega_Y^*$  such that  $p(xwz) > 0$ .*

**Proof:** Define

$$A_x \doteq \{w \in \Omega_Y^\infty \text{ such that } w = xu, u \in \Omega_Y^\infty\}, \quad (2.29)$$

$$A_z \doteq \{w \in \Omega_Y^\infty \text{ such that } w = zu, u \in \Omega_Y^\infty\}, \quad (2.30)$$

$$A_{xz} \doteq \{w \in \Omega_Y^\infty \text{ such that } w = xuzv, u \in \Omega_Y^*, v \in \Omega_Y^\infty\}. \quad (2.31)$$

These are sets of infinite length strings.  $A_x$  is the set of all strings that start with the substring  $x$  (similarly for  $A_z$ ).  $A_{xz}$  is the set of all strings that start with an  $x$  substring, followed by *any* finite length substring, followed by a  $z$  substring. Obviously we have  $A_{xz} \subset A_x$ . Denote by  $\pi$  the probability measure associated with  $p$ . Then  $\pi(A_x) = p(x)$ ,  $\pi(A_z) = p(z)$ . Finally define the shift-to-left transformation (denoted by  $S : \Omega_Y^\infty \mapsto \Omega_Y^\infty$ ):

$$Sy_1y_2\dots \doteq y_2y_3\dots \quad (2.32)$$

We denote  $m$  repeated shifts to the left by  $S^m$ . Because  $Y$  is ergodic,  $S$  is an ergodic transformation with respect to  $\pi$ , the measure generated by  $p$  (see Billingsley [Bil65]), so

$$\lim_{M \rightarrow \infty} \frac{\sum_{m=0}^{M-1} \mathbf{1}_{A_z}(S^m x)}{M} \stackrel{\pi\text{-a.a.}x}{=} \pi(A_z) = p(z) > 0. \quad (2.33)$$

Now, assume  $p(xwz) = 0 \forall w \in \Omega_Y^*$ ; then  $\pi(A_{xz}) = 0$ .

This is so because  $\pi(A_{xz})$  can be approximated by a sum:  $\sum_w \pi(A(xwz)) = \sum_w p(xwz)$ . The sum is over a countable number of finite length  $w$  strings.  $A(xwz)$  denotes the *rectangle* associated with  $xwz$  - see Definition 1.14. Since all the  $\pi(A(xwz))$  equal zero, we also have  $\pi(A_{xz}) = 0$ .

As  $\pi(A_{xz}) = 0$ , we have

$$\pi(A_x - A_{xz}) = \pi(A_x) - \pi(A_{xz}) = \pi(A_x) - 0 = p(x) > 0. \quad (2.34)$$

Now, if  $x \in A_x - A_{xz}$  then  $x$  is an infinite length string which starts with an  $x$  substring and contains no  $z$  substring. Therefore  $\mathbf{1}_{A_z}(S^m x) = 0 \forall m > N$ , which implies

$$\forall x \in A_x - A_{xz} \quad \lim_{M \rightarrow \infty} \frac{\sum_{m=0}^{M-1} \mathbf{1}_{A_z}(S^m x)}{M} = 0. \quad (2.35)$$

But  $\pi(A_x - A_{xz}) > 0$ , so (2.35) contradicts (2.33). Consequently our assumption that  $p(xwz) = 0 \forall w \in \Omega_Y^*$  was false. In other words,  $\exists w \in \Omega_Y^*$  such that  $p(xwz) > 0$ . This completes the proof. •

Now we proceed to prove Theorem 14.

**Proof of Theorem 14:** We can apply Lemma 12 as soon as we prove the uniqueness of the autoregressive model. So we need to show that, for every  $N$ ,  $P_N$  has exactly one ergodic class. The rest of the proof proceeds very similarly to that of Definition 40, except for a minor technical detail which we will discuss later.

First we will prove uniqueness of autoregressive model. To do so, we show first that  $\Omega_+$  (as defined in Definition 13) is the only ergodic class of  $P_N$ .

Take any  $y_1 \dots y_N, z_1 \dots z_N \in \Omega_+$ . Then  $p(y_1 \dots y_N) > 0$ ,  $p(z_1 \dots z_N) > 0$  and, by Lemma 14,  $\exists M \geq 0$ ,  $u = u_1 \dots u_M \in \Omega_Y^M$  such that  $p(y_1 \dots y_N u_1 \dots u_M z_1 \dots z_N) > 0$ . But then:

$$p(y_1 \dots y_N) > 0 \quad p(y_1 \dots y_N u_1) > 0 \quad \Rightarrow \Pr(X_{t+1}^N = y_2 \dots u_1 | X_t^N = y_1 \dots y_N) > 0$$

$$p(y_2 \dots u_1) > 0 \quad p(y_2 \dots u_1 u_2) > 0 \quad \Rightarrow \Pr(X_{t+2}^N = y_3 \dots u_2 | X_{t+1}^N = y_2 \dots u_1) > 0$$

..

$$p(u_M z_1 \dots z_{N-1}) > 0 \quad p(u_M z_1 \dots z_N) > 0 \quad \Rightarrow \Pr(X_{t+M+N}^N = z_1 \dots z_N | X_{t+M+N-1}^N = u_M z_1 \dots z_{N-1}) > 0.$$

From this follows that  $Pr(X_{t+M+N}^N = z_1 \dots z_N | X_{t+1}^N = y_1 \dots y_N) > 0$  and so  $y_1 \dots y_N \rightsquigarrow z_1 \dots z_N$ . We can prove in exactly the same way that  $z_1 \dots z_N \rightsquigarrow y_1 \dots y_N$ . So for all  $y_1 \dots y_N, z_1 \dots z_N \in \Omega_+$  we have  $y_1 \dots y_N \sim z_1 \dots z_N$ .

On the other hand, take any  $y_1 \dots y_N$  in  $\Omega_0$ . It cannot be a consequent of any  $z_1 \dots z_N$  in  $\Omega_+$ . To see this, assume the opposite, namely  $z_1 \dots z_N \rightsquigarrow y_1 \dots y_N$ . Then there must exist  $M \geq 0$  and a sequence of states:  $u_1 \dots u_N, \dots, v_1 \dots v_N$  such that  $Pr(X_{t+1}^N = u_1 \dots u_N | X_t^N = z_1 \dots z_N) > 0, \dots, Pr(X_{t+M}^N = y_1 \dots y_N | X_{t+M-1}^N = v_1 \dots v_N) > 0$ . Somewhere along this sequence of states there must be some  $n$  ( $M \geq n \geq 0$ ) and states  $w_1 \dots w_N \in \Omega_+$  and  $x_1 \dots x_N \in \Omega_0$  such that  $Pr(X_{t+n+1}^N = x_1 \dots x_N | X_{t+n}^N = w_1 \dots w_N) > 0$ ; but this contradicts the definition of  $P_N$ . Therefore  $\Omega_+$  is an ergodic class of  $P_N$ .

Assume there is some other ergodic class  $\Omega' \subset \Omega_0$ . Say  $y_1 \dots y_N, z_1 \dots z_N$  are in  $\Omega'$ ; hence  $y_1 \dots y_N \rightsquigarrow z_1 \dots z_N$ . Using the same argument as above we reach the conclusion that a state from  $\Omega_0$  can transit to another state of  $\Omega_0$ . But from the definition of  $P_N$ , all  $\Omega_0$  states transit into the same state, which belongs to  $\Omega_+$ , and we have reached a contradiction; therefore there is no other ergodic class  $P_N$  except for  $\Omega_+$ .

From this we conclude the existence of a unique autoregressive model  $(X^N, Y^N)$ , with  $X^N$  ergodic,  $Y^N$  ergodic. Weak convergence follows from Lemma 12. All of this is exactly the same as in the proof of Theorem 13. The last step, proof of cross entropy convergence, is almost exactly the same as in Theorem 13, but there is a small detail to be taken care of. In Theorem 13 we write  $H_n(Y^N; Y) = \sum p(y_0 y) \log \frac{p(y_0 | y)}{p_N(y_0 | y)}$ ; both  $p(y_0 | y)$  and  $p_N(y_0 | y)$  are strictly positive for all  $y_0 \in \Omega_Y, y \in \Omega_Y^n$ . Here however we have to consider the possibility that  $p_N(y_0 | y) = 0$  and  $p(y_0 | y) > 0$ . However this cannot be! If  $p(y_0 | y) > 0$  for some  $y_0, y$  then  $p(y_0 y) > 0 \Rightarrow$  (from Lemma 12)  $p_N(y_0 y) > 0 \Rightarrow p_N(y_0 | y) > 0$  as well. The rest of the proof of cross entropy convergence proceeds just as in Theorem 13. This completes the proof of the theorem. •

This completes our discussion of approximation by HMMs. It must be noted that the  $N$ -th order AR model defined here is closely related to a model of Ornstein [OW90] which makes use of  $N$ -th order Markov processes. Ornstein calls these *canonical models* and, under stronger

conditions on the original process (*total ergodicity*), he proves a strong approximation result, namely that the sequence of canonical models converges to the original in the  $d$ -bar sense;  $d$ -bar convergence is stronger than weak or cross entropy convergence, but weaker than convergence in the *total variation*. For details see [OW90]. On the other hand, it is not obvious how to find a practical estimation scheme that is consistent in the  $d$ -bar sense; the estimation scheme proposed in [OW90] would require impractically large amounts of data.

## 2.2 Approximation with Hidden Gibbs Models

In this section we show that we can approximate ergodic processes by Hidden Gibbs Models. Given Theorems 10, 11 in Section 1.4 (which state that any positive HMM is a Hidden Gibbs Model), we only need to show that approximation is possible by a positive HMM.

**Theorem 16** *Given a stationary ergodic stochastic process  $Y$  with finite alphabet  $\Omega_Y$  and probability function  $p$ , there is a sequence of HMM's  $(\bar{X}^N, \bar{Y}^N)$   $N = 1, 2, \dots$  (with observable probability functions  $\bar{p}_N$ ) such that:*

1.  $\bar{X}^N$  is stationary ergodic, positive.
2.  $\bar{Y}^N$  is stationary ergodic, positive.
3.  $\bar{Y}^N \xrightarrow{w} Y$ .

**Remark:** Note that the theorem does not prove cross entropy convergence.

To prove Theorem 16 we need the following well known Lemma:

**Lemma 17** *Given a regular Markov transition matrix  $P$  and a sequence of regular Markov transition matrices  $P_M$ ,  $M = 1, 2, \dots$ ; assume that  $P, P_M, M = 1, 2, \dots$  are all  $L$ -by- $L$ . By regularity,  $P$  has unique stationary probability  $p = [p(1) \dots p(L)]$  and, for every  $M$ ,  $P_M$  has unique stationary probability  $p_M = [p_M(1) \dots p_M(L)]$ . Then if*

$$\lim_{M \rightarrow \infty} \max_{k,l=1,\dots,L} |P_{kl}^M - P_{kl}| = 0 \tag{2.36}$$

it follows that

$$\lim_{M \rightarrow \infty} \max_{l=1, \dots, L} |p_M(l) - p(l)| = 0. \quad (2.37)$$

**Proof:** Assume that (2.37) is wrong, i.e. the sequence  $p_M$  does not tend to  $p$ . Then, by a standard compactness argument, there is a sequence  $\{m_k\}_{k=1}^{\infty}$  and a probability  $\bar{p} = [\bar{p}(1) \dots \bar{p}(L)] \neq p$ , such that

1.  $p_{m_k}$  is a stationary probability of  $P_{m_k}$ ,  $k \geq 1$ .
2.  $\lim_k \max_{l=1, \dots, L} |p_{m_k}(l) - \bar{p}(l)| = 0$ .

Now for all  $k, l$  we have

$$\begin{aligned} \max_l \left| \sum_j \bar{p}(j) P_{j,l} - \bar{p}(l) \right| &\leq \max_l \left| \sum_j \bar{p}(j) P_{j,l} - \sum_j p_{m_k}(j) P_{j,l} \right| + \\ &\max_l \left| \sum_j p_{m_k}(j) P_{j,l} - \sum_j p_{m_k}(j) [P_{m_k}]_{j,l} \right| + \max_l \left| \sum_j p_{m_k}(j) [P_{m_k}]_{j,l} - \bar{p}(l) \right|. \end{aligned} \quad (2.38)$$

Now, for any  $\epsilon > 0$ , there is a  $k_0$  such that for all  $k \geq k_0$  we have:

1. The first term on the right side of (2.38) is smaller than

$$\max_l |\bar{p}(l) - p_{m_k}(l)| < \epsilon/3, \quad (2.39)$$

because  $p_{m_k}(l) \rightarrow \bar{p}(l)$  for all  $l$ .

2. The second term on the right side of (2.38) is smaller than

$$\max_{j,l} |P_{j,l} - [P_{m_k}]_{j,l}| < \epsilon/3, \quad (2.40)$$

because  $\max_{j,l} |P_{j,l} - [P_{m_k}]_{j,l}| \rightarrow 0$ .

3. The third term on the right side of (2.38) equals

$$\max_l |p_{m_k}(l) - \bar{p}(l)| \quad (2.41)$$

because  $p_{m_k}$  is a stationary probability of  $P_{m_k}$ . Now (2.41) is smaller than  $\epsilon/3$  because  $p_{m_k}(l) \rightarrow \bar{p}(l)$  for all  $l$ .

Therefore we have for all positive  $\epsilon$

$$\max_l \left| \bar{p}(l) - \sum_j \bar{p}(j) P_{j,l} \right| \leq \epsilon \quad (2.42)$$

which implies that  $\bar{p}$  is a stationary probability of  $P$ . But  $P$  was assumed to have a unique stationary probability  $p$  and we also assumed that  $\bar{p} \neq p$ . So we have reached a contradiction. It follows that

$$\lim_{M \rightarrow \infty} \max_{l=1, \dots, L} |p_M(l) - p(l)| = 0. \quad (2.43)$$

This completes the proof of the Lemma. •

Now we proceed to the proof of Theorem 16.

**Proof of Theorem 16:** Denote by  $\Omega_Y = \{0, 1, \dots, K-1\}$  the alphabet of  $Y$ . Since  $Y$  is ergodic, we know it has exactly one sequence of autoregressive models  $(X^N, Y^N)$ ,  $N = 1, 2, \dots$ . Call  $P_N$  the probability transition matrix of  $X^N$ . If  $P_N$  had a strictly positive stationary probability, then it would be easy to prove that  $Y^N$  has positive probability function. However,  $P_N$  need not have a strictly positive probability. So we will consider a perturbed version of  $P_N$  which has stationary probability that is positive *and* is close to the original stationary probability.

To this end, consider an auxiliary probability transition matrix  $E_N$ :  $[E_N]_{x,z} \doteq 1/K^N$ ,  $x, z \in T^N$ . Now define a class of probability transition matrices  $P_{Nn}$ ,  $n = 1, 2, \dots$  by

$$P_{Nn} \doteq (1 - 1/n) \cdot P_N + 1/n \cdot E_N. \quad (2.44)$$

Now,  $P_{Nn}$  is strictly positive and hence there is exactly one stationary Markov process corresponding to it. Call this process  $X^{Nn}$ . Use the same  $f$  as in Definition 40 to define  $Y_t^{Nn} = f(X_t^{Nn})$ .  $Y^{Nn}$  is stationary with probability function  $p_{Nn} : \Omega_Y^* \mapsto [0, 1]$  and  $P_{Nn}$  has exactly one stationary probability, namely  $[p_{Nn}(x)]_{x \in \Omega_Y^*}$ . Similarly,  $Y^N$  is stationary, ergodic with probability function  $p_N : \Omega_Y^* \mapsto [0, 1]$ ;  $P_N$  has exactly one ergodic class and hence exactly one stationary probability, namely  $[p_N(x)]_{x \in \Omega_Y^*}$ . Obviously for all  $N$

$$\lim_n \max_{x, z \in \Omega_Y^*} |[P_{Nn}]_{x, z} - [P_N]_{x, z}| = 0, \quad (2.45)$$

so, by Lemma 2.2,

$$\lim_n \max_{x \in \Omega_Y^*} |p_{Nn}(x) - p_N(x)| = 0. \quad (2.46)$$

It is easy to prove that for any fixed  $x \in \Omega_Y^*$  and any fixed  $N$ ,  $\lim_n p_{Nn}(x) = p_N(x)$  and for any fixed  $x \in \Omega_Y^*$   $\lim_N p_N(x) = p(x)$ . Then, for all  $N, \epsilon > 0, x$  we can define a number  $n(N, \epsilon, x)$  such that  $\forall n \geq n(N, \epsilon, x)$

$$|p_{Nn}(x) - p(x)| < \epsilon. \quad (2.47)$$

Define

$$n_N \doteq \max_{x \in \Omega \cup \Omega_Y^2 \cup \dots \cup \Omega_Y^N, 1 \leq k \leq N} n(k, 1/N, x) \quad (2.48)$$

and

$$\bar{p}_N \doteq p_{N, n_N}, \quad \bar{P}_N \doteq P_{N, n_N}. \quad (2.49)$$

In other words, for all strings  $x$  of length less than or equal to  $N$ , we know  $\bar{p}_N(x)$  is “within  $1/N$ ” to  $p(x)$ . From this follows immediately that  $\bar{p}_N \xrightarrow{w} p$ ; but  $\bar{p}_N$  is the observable probability function of the DNM  $(\bar{X}^N, \bar{Y}^N)$  where we define  $\bar{X}^N$  to be the unique stationary Markov process with transition probability matrix  $\bar{P}_N$ .  $\bar{Y}$  is defined by  $\bar{Y}_t^N = f(\bar{X}_t^N)$ ,  $t = 0, \pm 1, \pm 2, \dots$ , and has probability function  $\bar{p}_N$ . Since  $\bar{P}_N$  is positive (for all  $N$ ) it is regular and hence  $\bar{X}^N$  is ergodic.  $\bar{Y}^N$  is ergodic as a function of  $\bar{X}^N$ .

To complete the proof it remains to be shown that the probability functions of  $\bar{X}^N, \bar{Y}^N$  are positive. We first prove  $Pr(\bar{X}_t^N = x) > 0$  for all  $x \in \Omega_Y^N$ . We have

$$Pr(\bar{X}_t^N = x) = \sum_{z \in \Omega_Y^N} Pr(\bar{X}_{t-1}^N = z) [\bar{P}_N]_{z,x} \geq \min_{x,z \in \Omega_Y^N} [\bar{P}_N]_{z,x} \sum_{z \in \Omega_Y^N} Pr(\bar{X}_{t-1}^N = z) = \min_{x,z \in \Omega_Y^N} [\bar{P}_N]_{z,x} > 0 \quad (2.50)$$

since  $\bar{P}_N$  is positive. Next take any  $n \geq 1, x_0, \dots, x_n \in \Omega_Y^N$  and consider that

$$Pr(\bar{X}_t^N \dots \bar{X}_{t+n}^N = x_0 \dots x_n) = Pr(\bar{X}_t^N = x_0) \cdot [\bar{P}_N]_{x_0, x_1} \cdot \dots \cdot [\bar{P}_N]_{x_{n-1}, x_n} > 0 \quad (2.51)$$

since all the terms in the product are strictly positive.

So we have proven that  $\bar{X}^N$  has a positive probability function. To complete the proof, observe that for all  $N \geq 1, m \geq N, y_1, \dots, y_m \in T$

$$\bar{p}_N(y_1 \dots y_m) = \sum_{x_1, \dots, x_m \in \Omega_Y^N: f(x_1)=y_1, \dots, f(x_m)=y_m} Pr(\bar{X}_1^N = x_1, \dots, \bar{X}_m^N = x_m). \quad (2.52)$$

Every term in the sum is positive, so  $p_N(y_1 \dots y_m) > 0$  and the proof is complete. This completes the proof. •

## 2.3 Exact Representation

In this section we present an *exact representation result* for HMM's. We show that for every stationary stochastic process  $Y$  (subject to certain mild conditions) there is a Markov Process  $X$  with  $Y_t = f(X_t), t = 0, \pm 1, \dots$ . This result has essentially been discovered by Harris [Har55]; here we simply reformulate it in HMM terminology.

To prove this result, we make an important change in our usual assumptions and take the state process  $X$  to have uncountably infinite alphabet.

Before we can state the theorem, we need to define a new quantity, the *infinite conditional*

probability function.

**Definition 41** Given a stationary stochastic process  $Y$ , with finite alphabet  $\Omega$ , probability function  $p$  and associated probability measure  $\pi$ , we call any function  $p(\cdot|\cdot) : \Omega \times \Omega^\infty \mapsto [0, 1]$  that satisfies

$$\forall n \geq 1 \forall y_0, y_{-1}, \dots \in \Omega \quad \int_{A(y_{-n}, \dots, y_{-1})} p(y_0|z_{-1}z_{-2}\dots) d\pi(z_{-1}z_{-2}\dots) = p(y_{-n}\dots y_0). \quad (2.53)$$

a version of the infinite conditional probability function.

**Remark:** See [Bil65] for a proof that a version of the infinite conditional probability function always exists. Then it is easy to see that an infinity of such functions exist, because, if  $p(\cdot|\cdot)$  is a version and  $q(\cdot|\cdot)$  differs from  $p(\cdot|\cdot)$  only on a set of  $\pi$ -measure 0, then  $q(\cdot|\cdot)$  is a version, too. Therefore, strictly speaking, we identify the infinite conditional probability function itself with the class of functions that satisfy (2.53) and we call any individual function a version. However, for brevity, we will usually call any function that satisfies (2.53) an infinite conditional probability function, or just a conditional probability function. In fact we use the same symbolism (e.g.  $p(\cdot|\cdot)$ ) for both the infinite and finite (see Def.6) conditional probability functions; it will always be clear from the context which one is meant.

The following important theorem is proven in [Bil65].

**Theorem 18** If  $Y$  is a stationary stochastic process with finite alphabet  $\Omega$ , probability function  $p$  and associated probability measure  $\pi$ , then

$$\forall y_0 \in \Omega, \pi\text{-a.a. } y_{-1}y_{-2}\dots \in \Omega^\infty \quad \lim_{n \rightarrow \infty} p(y_0|y_{-1}\dots y_{-n}) = p(y_0|y_{-1}y_{-2}\dots). \quad (2.54)$$

**Proof:** In [Bil65]. •

**Remark:** In view of the previous theorem, sometimes we write the infinite conditional probability function as

$$Pr(Y_0 = y_0 | Y_{-1} = y_{-1}, Y_{-2} = y_{-2}, \dots). \quad (2.55)$$

Now we are ready to state the following theorem:

**Theorem 19** (*Representation Theorem*) *Given a stationary ergodic positive stochastic process  $Y$ , with finite alphabet  $\Omega$ , probability function  $p$  and associated probability measure  $\pi$ , suppose there is a version of the conditional probability function, call it  $p(\cdot|\cdot)$ , that satisfies the following:*

1. *There is a  $y_0 \in \Omega$  and a  $\delta > 0$  such that*

$$\forall y_{-1}, y_{-2} \dots \in \Omega \quad p(y_0|y_{-1}y_{-2}\dots) \geq \delta. \quad (2.56)$$

2.  $\sum_{m=1}^{\infty} \epsilon_m < \infty$ , where  $\epsilon_m, m = 1, 2, \dots$  is defined by

$$\epsilon_m \doteq \sup |p(y_0|y_{-1}y_{-2}\dots y_{-m}z_{-m-1}z_{-m-2}\dots) - p(y_0|y_{-1}y_{-2}\dots y_{-m}x_{-m-1}x_{-m-2}\dots)|; \quad (2.57)$$

*the supremum is taken over all  $y_0, y_{-1}, \dots, y_{-m}, z_{-m-1}, z_{-m-2}, \dots, x_{-m-1}, x_{-m-2}, \dots$  in  $\Omega$ .*

*Then the following are true:*

1.  *$Y$  is the **only** stationary stochastic process which has conditional probability  $p(\cdot|\cdot)$ . That is, **there is no other** stationary process  $Z$ , different from  $Y$ , for which we have:*

$$\forall y_0 \pi\text{-a.a. } y_{-1}y_{-2}\dots \in \Omega^\infty \quad Pr(Z_0 = y_0|Z_{-1} = y_{-1}, Z_{-2} = y_{-2}, \dots) = p(y_0|y_{-1}y_{-2}). \quad (2.58)$$

2. *There is a stochastic process  $X$ , which is stationary ergodic and Markov, taking values in  $\Omega^\infty$ , such that for all  $t = 0, \pm 1, \dots$   $Y_t = f(X_t)$ , where  $f : \Omega^\infty \mapsto \Omega$  is defined for all  $y_0, y_{-1}, \dots \in \Omega$  by  $f(y_0y_{-1}\dots) \doteq y_0$ .*

**Proof of Theorem 2.7:** Statement (1) is one of the results of Harris' Theorem 6 [Har55].

Statement (2) follows directly from the construction Harris uses to prove his theorem. •

**Remark:** What is the significance of the condition (2.57)? Here are two alternative points of view. On the one hand, we can consider it as a statement about the continuity properties (with

respect to  $y$ ) of the function  $p(y_0|y)$ . Here continuity is defined with respect to the metric defined in Section 1.2. Then (2.57) says that  $p(y_0|y)$  has a strong continuity property. Alternatively, we can consider (2.57) as a statement about the dependence of current values of  $Y$  on past conditioning; apparently the process  $Y$  does not retain a strong memory of its past (depends weakly on the past). So (2.57) is a *mixing* condition (for more details on mixing see [Bil65]). Of course the two points of view are related: continuity of the conditional probability function is a form of mixing.

**Remark:** A theorem similar to that of Harris is proven in [DF37]

# Chapter 3

## Estimation

In this chapter we prove that Maximum Likelihood Estimation over a large class of Hidden Markov Models is consistent. The result we prove is roughly the following: given a sequence of observations  $y_1, y_2, \dots$  from a stochastic process  $Y$  and a sequence of HMM classes  $\Phi_N$ ,  $N = 1, 2, \dots$ , we can find a sequence  $N(n)$  such that  $\hat{Y}^{N(n)} \rightarrow Y$  in the cross entropy sense. Here  $\hat{Y}^{N(n)}$  is the Maximum Likelihood model of  $Y$  in class  $\Phi_{N(n)}$ , with observations  $y_1, \dots, y_n$ . For the result to hold true, certain conditions must be imposed on  $Y$  and  $\Phi_N$ ,  $N = 1, 2, \dots$ .

The chapter is organized as follows: in Section 3.1 we state some preliminary definitions and the main *consistency theorem*. The proof of this theorem is postponed; first we need to prove several auxiliary lemmas. In Section 3.2 we prove a number of *positivity* properties of the elements of  $\Phi_N$ . In Section 3.3 we prove *mixing properties* of the elements of  $\Phi_N$ . In Section 3.4 we introduce a class of *auxiliary suboptimal models* and prove some of its properties. The results of Sections 3.2, 3.3 and 3.4 are combined in Section 3.5 to obtain the *proof of the consistency theorem*.

### 3.1 Preliminaries

In this section we start with an informal discussion of the Maximum Likelihood Estimation problem. Then we introduce some definitions that will be used in what follows, and present the

consistency theorem for Maximum Likelihood Estimation. This theorem is the main result of this chapter; as its proof requires considerable elaboration it will be given in Section 3.5.

As we have seen in Chapter 2, given the probability function of a stochastic process  $Y$  we can construct a sequence of HMM's that approximate  $Y$  arbitrarily well. However, in general we will not know the probability function. Typically we will only have a finite number of observations  $y_1, y_2, \dots, y_n$  from  $Y$ , and we will construct a HMM of  $Y$  based on these observations.

A strategy of modelling immediately suggests itself: use the observations to compute the probability function.

**Definition 42** *The  $(\mathbf{N}, \mathbf{n})$ -th empirical probability of  $Y$  is defined for all  $n \geq 1$ ,  $N \leq n$ ,  $y = y_1 y_2 \dots \in \Omega^\infty$ ,  $z_1, z_2, \dots, z_N \in \Omega$  by*

$$p_{N,n}^y(z_1 \dots z_N) \doteq \frac{\sum_{k=0}^{n-N} \mathbf{1}_{z_1 \dots z_N}(y_{k+1} \dots y_{k+N})}{n - N + 1}. \quad (3.1)$$

Having obtained the  $(N, n)$ -th empirical probability (for any  $N < n$ ), we can use it to build an  $N$ -th order AR model, similar to the one of Chapter 2. Call this model the **empirical  $(\mathbf{N}, \mathbf{n})$ -th order AR model**, and denote it by  $(X^{N,n,y}, Y^{N,n,y})$ . It is easy to show that a sequence  $N(n)$ ,  $n = 1, 2, \dots$ , exists such that the following are true:

$$Pr(\text{w-} \lim_{n \rightarrow \infty} Y^{N(n),n,y} = Y) = 1. \quad (3.2)$$

$$Pr(\lim_{n \rightarrow \infty} H(Y^{N(n),n,y}, Y) = 0) = 1. \quad (3.3)$$

**Remark:** Note that  $p_{N,n}^y$  is a random variable, depending on the observations  $y_1, \dots, y_n$ . Think of a finite sequence of observations  $y_1 \dots y_n$  as the truncation of an infinite string  $y = y_1 y_2 \dots$ ; this explains the  $y$ -superscript /  $n$ -subscript notation for  $p_{N,n}^y$ . From this follows that the estimator  $Y^{N(n),n,y}$  is also a random variable in  $y$ , and statements about the convergence properties of  $Y^{N(n),n,y}$  have to be specified probabilistically (convergence with probability 1, convergence in

probability etc.); vid. eqs.(3.2), (3.3). In fact, this will be true of any estimator that makes use of the  $y$  observations (and all reasonable estimators do!).

Eqs. (3.2), (3.3) are satisfying, because they imply that we can approximate the true process  $Y$  arbitrarily well using empirical AR models. However, empirical AR models are impractical for the following reason. Their performance depends on  $p_{N(n),n}^y(z)$  being close to  $p(z)$  for all  $z$  of length  $N(n)$ ; and the number  $n$  of  $Y$  observations needed to closely approximate the  $N(n)$ -th order marginals of  $Y$  will grow impractically fast. Therefore, other, more sophisticated, estimation methods are generally used for Hidden Markov Modelling.

The general idea is to define a parametric class of HMM's  $(X^\phi, Y^\phi)$ ,  $\phi$  belonging to an appropriate parameter set  $\Phi$ . (The usual parametrization is in terms of the transition and emission probabilities.) We also define a meaningful cost function  $J(\phi; y_1, \dots, y_n)$ , which depends on the model parameters  $\phi$  and the observations  $y_1, \dots, y_n$ . Then we choose the value  $\hat{\phi}(y_1 \dots y_n)$  that optimizes  $J(\phi; y_1, \dots, y_n)$ . This yields the optimal estimator  $(X^{\hat{\phi}}, Y^{\hat{\phi}})$ . Many variations on the main theme are possible, depending on the choice of parametrization and cost function, but in this thesis we limit ourselves to *Maximum Likelihood Estimation* which is defined as follows.

**Definition 43** *Given a stochastic process  $Y$ , with finite alphabet  $\Omega$ , a sequence of observations  $y_1, \dots, y_n$  from  $Y$ , and a class of stochastic processes  $Y^\phi$  (where  $\phi$  takes values in an appropriate parameter set  $\Phi$ ), each with probability function  $p_\phi$  and alphabet  $\Omega$ , we define the  **$n$ -th order Likelihood function** by*

$$L(\phi; y_1 \dots y_n) \doteq p_\phi(y_1 \dots y_n). \quad (3.4)$$

**Definition 44** *A **Maximum Likelihood** model (with respect to class  $\Phi$  and observations  $y_1, \dots, y_n$ ) is a pair of processes  $(X^{\hat{\phi}(y_1 \dots y_n)}, Y^{\hat{\phi}(y_1 \dots y_n)})$ , where  $\hat{\phi}(y_1 \dots y_n)$  maximizes  $L(\phi; y_1 \dots y_n)$  over  $\Phi$ .*

**Remark:** Note that the Maximum Likelihood model depends on the choice of parameter class  $\Phi$ .

**Remark:** There may be more than one  $\hat{\phi}$  elements in  $\Phi$  that maximize  $L(\phi; Y_1 \dots y_n)$ , so strictly speaking, we must refer to the *set of ML models*.

Now we specialize to ML estimation of HMM's. We define a sequence  $\Phi_N(\Omega)$ ,  $N = 1, 2, \dots$ , of parameter classes. Each parameter class consists of pairs of matrices:  $\phi = (P, Q)$ .  $P$  is a transition matrix and  $Q$  is an emission matrix; a pair  $\phi$  completely defines a SNM.

**Definition 45** For any finite set  $\Omega = \{0, 1, \dots, K-1\}$ , define the set  $\Phi_N(\Omega)$ ,  $N = 1, 2, \dots$  to be the set of all matrix pairs  $\phi = (P, Q)$  satisfying:

1. The matrix  $[P_{x,z}]_{x,z \in \Omega^N}$  is stochastic.
2. Define  $\alpha_N \doteq \frac{1}{N \cdot K^N}$ ; then

$$\forall x, z \in \Omega^N \quad P_{x,z} \geq \alpha_N \text{ or } = 0. \quad (3.5)$$

3. We have

$$\forall y_0, \dots, y_N \in \Omega \quad P_{y_0 \dots y_{N-1}, y_1 \dots y_N} \geq \alpha_N. \quad (3.6)$$

4. The matrix  $[Q_{x,y}]_{x \in \Omega^N, y \in \Omega}$  is stochastic.

5. Finally,

$$\forall x \in \Omega^N, y \in \Omega \quad Q_{x,y} \geq \alpha_N \quad (3.7)$$

**Remark:**  $\Phi_N(\Omega)$  is the parameter class. Note that, for every  $\phi = (P, Q) \in \Phi_N(\Omega)$ , we have  $P^N > 0$ . Hence  $P$  is regular and every  $\phi$  defines a *unique* stationary SNM. Therefore the next definition makes sense.

**Definition 46** For a fixed  $\Omega = \{0, \dots, K-1\}$ , for  $N = 1, 2, \dots$  and for every  $\phi = (P, Q)$  in  $\Phi_N(\Omega)$ , denote by  $(X^\phi, Y^\phi)$  the unique SNM satisfying:

1.  $X^\phi$  is a stationary Markov process with

$$\forall x, z \in \Omega^N \quad Pr(X_{t+1}^\phi = z | X_t^\phi = x) = P_{x,z}. \quad (3.8)$$

2.  $Y^\phi$  satisfies

$$\forall x \in \Omega^N, y \in \Omega \quad \Pr(Y_t^\phi = y | X_t^\phi = x) = Q_{x,y}. \quad (3.9)$$

We denote the probability function of  $Y^\phi$  by  $p_\phi$ .

Next we define the ML parameter estimates, with respect to class  $\Phi_N(\Omega)$ .

**Definition 47** Denote by  $\Psi_{N,n}(y)$  the set of **ML parameter estimates** (with respect to class  $\Phi_N(\Omega)$ , observations  $y_1, \dots, y_n$ ). They are defined as follows:  $\forall N, n \geq 1, y = y_1 y_2 \dots \in \Omega^\infty$

$$\Psi_{N,n}(y) \doteq \{\psi \in \Phi_N(\Omega) : L(\psi; y_1 \dots y_n) = \sup_{\phi \in \Phi_N(\Omega)} L(\phi; y_1 \dots y_n)\}. \quad (3.10)$$

The set of **ML models** (with respect to class  $\Phi_N(\Omega)$  and observations  $y_1, \dots, y_n$ ) is the set  $\{(X^\phi, Y^\phi)_{\phi \in \Psi_{N,n}(y)}\}$ .

**Remark:** In Section 3.5 we will show that  $\Psi_{N,n}(y)$  is not empty.

The basic idea we use in our estimation scheme is the following. First, note that the dimension of the  $P, Q$  matrices in class  $\Phi_N(\Omega)$  increases with  $N$ ; hence we expect that, for  $N > M$ , the modelling power of class  $\Phi_N(\Omega)$  is greater than that of class  $\Phi_M(\Omega)$ . On the other hand, the number of free parameters in  $(P, Q) \in \Phi_N(\Omega)$  also increases with  $N$ ; hence, using “statistics common sense” we expect that we will need more observations (bigger  $n$ ) for reliable estimation. The idea is to let  $N = N(n)$  grow as a function of  $n$ , in such a way that any sequence of ML estimates  $\{(X^{\hat{\phi}}, Y^{\hat{\phi}}), \hat{\phi} \in \Psi_{N(n),n}(y)\}_{n=1}^\infty$  will be consistent. This is the idea of estimation by the *method of sieves*, introduced by Grenander [Gre81]. Provided that the original process  $Y$  satisfies certain reasonable conditions, consistent estimation is indeed possible; this is the main result of this chapter and is summarized in the following theorem.

**Theorem 20** (*Consistency*) Given a stationary ergodic stochastic process  $Y$ , with finite alphabet  $\Omega = \{0, 1, \dots, K-1\}$ , probability function  $p$  and associated probability measure  $\pi$ , assume that

$$\exists \alpha > 0 \text{ such that } \forall m \geq 1, y_{-m}, \dots, y_{-1}, y_0 \in \Omega \quad p(y_0 | y_{-1} \dots y_{-m}) \geq \alpha. \quad (3.11)$$

Then, there is a sequence  $N(n)$ ,  $n = 1, 2, \dots$ ,  $N(n) \uparrow \infty$ , such that

$$\lim_{n \rightarrow \infty} \left\{ \sup_{\phi \in \Psi_{N(n)}(y)} H(p_\phi; p) \right\} \stackrel{\text{in prob.}}{=} 0. \quad (3.12)$$

**Remark:** Note that the limit in the consistency equation (3.12) is a limit in probability with respect to  $y$ . As we already mentioned, the limit has to be probabilistic, as the ML estimators are functions of the random observations.

**Remark:** Also, note that the supremum in (3.12) is taken over the set of all ML estimators in a given class  $\Phi_N(\Omega)$ .  $\Phi_N(\Omega)$  is a *big* class of processes, which is a mixed blessing. On the one hand we can model a lot of processes; on the other hand it is hard to compute the ML estimates in this big class. In this respect, see the next remark.

**Remark:** Theorem 20 shows consistency of ML estimation. It is an existence theorem; it does not tell us how to actually obtain the ML estimates. The theorem has a limited applicability to practical situation. To illustrate this consider how we would proceed to estimate a consistent sequence of models, and what difficulties we would encounter. For a fixed number of observations  $y_1, \dots, y_n$  we look for a ML estimate in class  $N(n)$ . If we truly compute an element of  $\Psi_{N(n)}(y)$  for every  $n$ , then we expect to approximate the true process in the cross entropy sense as  $n$  goes to infinity. To find a ML estimate in  $\Phi_{N(n)}$  we can use, the Baum Backward / Forward algorithm [BE67]. Another option is the use of stochastic relaxation methods. However, the BF algorithm computes a local maximum, and the stochastic relaxation algorithm is slow, so it is not clear that what we get is truly a global ML estimate. In addition, we have to ensure that we always remain in class  $\Phi_{N(n)}$ . This means that we have to modify certain estimates  $\hat{P}, \hat{Q}$  if some of their elements are smaller than  $\alpha_{N(n)}$ . This also detracts from the optimality of the estimates. In practice of course, we ignore all of these constraints and proceed in faith to obtain some suboptimal estimates. As we will see in Chapters 4 and 5, this suboptimal strategy still yields very good models.

**Remark:** Finally note that eq.(3.11) implies that the original process  $Y$  is positive.

The proof of Theorem 20 is complicated and requires a lot of preparatory work. We need to prove a number of properties of the  $\Phi_N(\Omega)$  classes (Sections 3.2, 3.3). Especially we need *mixing* properties which will allow us to relate the Likelihood function of a Hidden Markov Model to the cross entropy between the model and the original process. We also need to introduce a class of auxiliary HMMs (Section 3.4) which are *suboptimal* (do not maximize Likelihood) but converge to the original process in the cross entropy sense. There is one such model per class  $\Phi_N(\Omega)$ , and any ML estimator in  $\Phi_N(\Omega)$  is at least as good (in the cross entropy sense) as this model; therefore the sequence of ML estimators converges to  $Y$ , as stated in (3.12). This argument is presented in a rigorous manner in Section 3.5 and completes the proof of Theorem 20.

## 3.2 Positivity Properties

In this section we prove some properties of the elements of the classes  $\Phi_N(\Omega)$ ,  $N = 1, 2, \dots$ ; these properties are of interest in themselves and, more importantly, are necessary for the proof of the Consistency Theorem 20.

All of the results proven here assume that  $\Phi_N(\Omega)$  is fixed (as defined in Definition 45) for  $N = 1, 2, \dots$ . As  $\Omega$  will not change in what follows, we drop it from our notation and write just  $\Phi_N$ .

The first lemma we prove establishes positivity properties for marginal probabilities of the  $(X^\phi, Y^\phi)$  pair, for any  $\phi \in \Phi_N$ ,  $N = 1, 2, \dots$ .

**Lemma 21** *For all  $N \geq 1$ , for all  $\phi \in \Phi_N$ , for all  $M \geq 1$ , for all  $m, n, k_1, k_2, \dots, k_M$  such that*

$$m \leq k_1 \cdot N < k_2 \cdot N < \dots < k_M \cdot N \leq n, \quad (3.13)$$

*and for all  $y_m, y_{m+1}, \dots, y_n \in \Omega$ ,  $x_1, x_2, \dots, x_M \in \Omega^N$ , we have*

$$Pr(X_{k_1 \cdot N}^\phi = x_1, X_{k_2 \cdot N}^\phi = x_2, \dots, X_{k_M \cdot N}^\phi = x_M, Y_m^\phi = y_m, \dots, Y_n^\phi = y_n) > 0. \quad (3.14)$$

$$Pr(Y_m^\phi = y_m, \dots, Y_n^\phi = y_n) \geq \alpha_N^{n-m+1}. \quad (3.15)$$

**Proof:** Fix  $N \geq 1$ ,  $\phi = (P, Q) \in \Phi_N$ ,  $M \geq 1$ ,  $m, n, k_1, k_2, \dots, k_N, y_m, y_{m+1}, \dots, y_n \in \Omega$ ,  $x_1, x_2, \dots, x_M \in \Omega^N$ .

Let us first prove (3.14). We have

$$\begin{aligned} Pr(X_{k_1 \cdot N}^\phi = x_1, X_{k_2 \cdot N}^\phi = x_2, \dots, X_{k_M \cdot N}^\phi = x_M, Y_m^\phi = y_m, \dots, Y_n^\phi = y_n) = \\ \sum_{z_m, \dots, z_n} Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n) \cdot Q_{z_m, y_m} \cdot \dots \cdot Q_{z_n, y_n}; \end{aligned} \quad (3.16)$$

the sum being taken over all  $z_k \in \Omega^N$ ,  $k = m, \dots, k_1 \cdot N - 1, k_1 \cdot N + 1, \dots, k_2 \cdot N - 1, k_2 \cdot N + 1, \dots, n$ , while we set  $z_{k_1 \cdot N} = x_1, \dots, z_{k_M \cdot N} = x_M$ .

Now, for all  $z \in \Omega^N$ ,  $y \in \Omega$  we have  $Q_{z, y} \geq \alpha_N$ . Therefore if one of the terms  $Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n)$  in (3.16) is positive, the sum itself will be positive and we have proven (3.14).

To prove that at least one of the terms  $Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n)$  is positive, note that for all  $P$  in  $\Phi_N$ ,  $P^N > 0$ , hence

$$Pr(X_{k_1 \cdot N}^\phi = x_1, X_{k_2 \cdot N}^\phi = x_2, \dots, X_{k_M \cdot N}^\phi = x_M) > 0. \quad (3.17)$$

But the probability in (3.17) equals

$$\sum_{z_m, \dots, z_n} Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n), \quad (3.18)$$

the sum being taken over all  $z_k \in \Omega^N$ ,  $k = m, \dots, k_1 \cdot N - 1, k_1 \cdot N + 1, \dots, k_2 \cdot N - 1, k_2 \cdot N + 1, \dots, n$ , while we set  $z_{k_1 \cdot N} = x_1, \dots, z_{k_M \cdot N} = x_M$ .

Combining (3.17) and (3.18) it is obvious that at least one of the summands in (3.18) has to be positive. This completes the proof of (3.14).

To prove (3.15) we argue similarly. We have

$$Pr(Y_m^\phi = y_m, \dots, Y_n^\phi = y_n) = \sum_{z_m, \dots, z_n \in \Omega^N} Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n) \cdot Q_{z_m, y_m} \cdot \dots \cdot Q_{z_n, y_n}; \quad (3.19)$$

the sum being taken over all  $z_k \in \Omega^N$ ,  $k = m, \dots, n$ .

We also know that

$$\sum_{z_m, \dots, z_n \in \Omega^N} Pr(X_m^\phi = z_m, \dots, X_n^\phi = z_n) = 1 \quad (3.20)$$

and

$$\forall z_m, \dots, z_n \in \Omega^N, \forall y_m, \dots, y_n \in \Omega \quad Q_{z_m, y_m} \cdot \dots \cdot Q_{z_n, y_n} \geq \alpha_N^{n-m+1}. \quad (3.21)$$

Combining (3.20) and (3.21) we obtain (3.15) and complete the proof of the lemma. •

The next lemma establishes an important property of SNM, namely the way conditioning is truncated in conditional probabilities.

**Lemma 22** *For all  $N \geq 1$ , for all  $\phi \in \Phi_N$ , for all  $l, m$  such that  $l \leq -N < 0 \leq m$  and for all  $y_l, \dots, y_m \in \Omega$ ,  $x, z \in \Omega^N$  we have:*

$$Pr(Y_0^\phi = y_0, \dots, Y_m^\phi = y_m, X_0^\phi = x | Y_l^\phi = y_l, \dots, Y_{-1}^\phi = y_{-1}, X_{-N}^\phi = z) =$$

$$Pr(Y_0^\phi = y_0, \dots, Y_m^\phi = y_m, X_0^\phi = x | Y_{-N+1}^\phi = y_{-N+1}, \dots, Y_{-1}^\phi = y_{-1}, X_{-N}^\phi = z). \quad (3.22)$$

**Proof:** Fix  $N \geq 1$ ,  $\phi = (P, Q) \in \Phi_N$ ,  $l, m$  such that  $l \leq -N < 0 \leq m$  and  $x, z \in \Omega^N$ ,  $y_l, \dots, y_m \in \Omega$ . By Lemma 21, the following conditional probability can be written as the ratio of two *positive* marginals:

$$Pr(Y_0^\phi = y_0, \dots, Y_m^\phi = y_m, X_0^\phi = x | Y_l^\phi = y_l, \dots, Y_{-1}^\phi = y_{-1}, X_{-N}^\phi = z) =$$

$$\begin{aligned}
& \frac{\Pr(Y_l^\phi = y_l, \dots, Y_m^\phi = y_m, X_{-N}^\phi = z, X_0^\phi = x)}{\Pr(Y_l^\phi = y_l, \dots, Y_{-1}^\phi = y_{-1}, X_{-N}^\phi = z)} = \\
& \frac{\sum_{x_n, l \leq n \leq m, n \neq -N, 0} \Pr(X_l^\phi = x_l) Q_{x_n, y_n} \cdots P_{x_{m-1}, x_m} Q_{x_m, y_m}}{\sum_{x_n, -l \leq n \leq -1, n \neq -N} \Pr(X_l^\phi = x_l) Q_{x_n, y_n} \cdots P_{x_{-2}, x_{-1}} Q_{x_{-1}, y_{-1}}}; \tag{3.23}
\end{aligned}$$

in the fraction above we set  $x_{-N} = z$  and  $x_0 = x$ . Now, we write the numerator of (3.23) as

$$\begin{aligned}
& \left\{ \sum_{x_n, l \leq n \leq -N-1} \Pr(X_l^\phi = x_l) Q_{x_l, y_l} \cdots P_{x_{-N-1}, x_{-N}} Q_{x_{-N}, y_{-N}} \right\} \times \\
& \left\{ \sum_{x_n, -N+1 \leq n \leq m, n \neq 0} P_{x_{-N}, x_{-N+1}} Q_{x_{-N}, y_{-N}} \cdots P_{x_{m-1}, x_m} Q_{x_m, y_m} \right\} \tag{3.24}
\end{aligned}$$

and the denominator as

$$\begin{aligned}
& \left\{ \sum_{x_n, l \leq n \leq -N-1} \Pr(X_l^\phi = x_l) Q_{x_l, y_l} \cdots P_{x_{-N-1}, x_{-N}} Q_{x_{-N}, y_{-N}} \right\} \times \\
& \left\{ \sum_{x_n, -N+1 \leq n \leq -1} P_{x_{-N}, x_{-N+1}} Q_{x_{-N}, y_{-N}} \cdots P_{x_{-2}, x_{-1}} Q_{x_{-1}, y_{-1}} \right\}. \tag{3.25}
\end{aligned}$$

Substituting (3.24) and (3.25) in (3.23) and performing a cancellation of identical factors, we obtain

$$\begin{aligned}
& \frac{\sum_{x_n, -N+1 \leq n \leq m, n \neq 0} P_{x_{-N}, x_{-N+1}} Q_{x_{-N}, y_{-N}} \cdots P_{x_{m-1}, x_m} Q_{x_m, y_m}}{\sum_{x_n, -N+1 \leq n \leq -1} P_{x_{-N}, x_{-N+1}} Q_{x_{-N}, y_{-N}} \cdots P_{x_{-2}, x_{-1}} Q_{x_{-1}, y_{-1}}} = \\
& \frac{\Pr(Y_{-N+1}^\phi = y_{-N+1}, \dots, Y_m^\phi = y_m, X_0^\phi = x | X_{-N}^\phi = z)}{\Pr(Y_{-N+1}^\phi = y_{-N+1}, \dots, Y_{-1}^\phi = y_{-1} | X_{-N}^\phi = z)} = \\
& \Pr(Y_0^\phi = y_0, \dots, Y_m^\phi = y_m, X_0^\phi = x | Y_{-N+1}^\phi = y_{-N+1}, \dots, Y_{-1}^\phi = y_{-1}, X_{-N}^\phi = z). \tag{3.26}
\end{aligned}$$

This completes the proof of the lemma. •

The next lemma establishes positivity properties for conditional probabilities of the  $(X^\phi, Y^\phi)$  pair,  $\phi \in \Phi_N$ ,  $N = 1, 2, \dots$ .

**Lemma 23** *For all  $N \geq 1$ , for all  $\phi \in \Phi_N$  we have:*

1. For all  $m \geq 1$  and for all  $y_0, y_{-1}, y_{-2}, \dots, y_{-m} \in \Omega$

$$Pr(Y_0^\phi = y_0 | Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) \geq \alpha_N. \quad (3.27)$$

2. For all  $z, x \in \Omega^N$  and for all  $y_N, \dots, y_{2N-1} \in \Omega$

$$Pr(X_{2N}^\phi = x | Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1}, X_N^\phi = z) \geq \alpha_N^{2N}. \quad (3.28)$$

3. For all  $x \in \Omega^N$  and for all  $y_N, \dots, y_{2N-1} \in \Omega$

$$Pr(Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1} | X_N^\phi = x) \geq \alpha_N^{2N}. \quad (3.29)$$

4. For all  $x, z \in \Omega^N$  and for all  $y_1, \dots, y_N \in \Omega$

$$Pr(X_N^\phi = z | Y_1^\phi = y_1, \dots, Y_N^\phi = y_N, X_0^\phi = x) \geq \alpha_N^{2N}. \quad (3.30)$$

5. For all  $x \in \Omega^N$  and for all  $y_1, \dots, y_N \in \Omega$

$$Pr(Y_1^\phi = y_1, \dots, Y_N^\phi = y_N | X_0^\phi = x) \geq \alpha_N^{2N}. \quad (3.31)$$

**Proof:** Fix  $N \geq 1$ ,  $\phi = (P, Q) \in \Phi_N$ . To prove (3.27), fix  $m \geq 1$ ,  $y_0, y_{-1}, y_{-2}, \dots, y_{-m} \in \Omega$ . We have:

$$\begin{aligned} & Pr(Y_0^\phi = y_0 | Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) = \\ & \sum_{x \in \Omega^N} Pr(Y_0^\phi = y_0 | X_0^\phi = x, Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) \cdot Pr(X_0^\phi = x | Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) = \end{aligned}$$

(using Lemma 22 for truncation of conditioning)

$$\sum_{x \in \Omega^N} Q_{x, y_0} \cdot Pr(X_0^\phi = x | Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) \geq$$

$$\min_{x \in \Omega^N, y_0 \in \Omega} \{Q_{x, y_0}\} \sum_{x \in \Omega^N} \Pr(X_0^\phi = x | Y_{-1}^\phi = y_{-1}, \dots, Y_{-m}^\phi = y_{-m}) \geq \alpha_N. \quad (3.32)$$

To prove (3.28) fix  $x, z \in \Omega^N, y_N, \dots, y_{2N-1}$ . By Lemma 21 the conditional in (3.28) is positive.

Hence:

$$\begin{aligned} & \Pr(X_{2N}^\phi = x | Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1}, X_N^\phi = z) = \\ & \frac{\Pr(X_{2N}^\phi = x, Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1} | X_N^\phi = z)}{\Pr(Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1} | X_N^\phi = z)} \geq \\ & \Pr(X_{2N}^\phi = x, Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1} | X_N^\phi = z) = \\ & \sum_{x_{N+1}, \dots, x_{2N} \in \Omega^N} Q_{z, y_N} \cdot P_{z, x_1} \cdot Q_{x_{N+1}, y_{N+1}} \cdot P_{x_{N+1}, x_{N+2}} \cdot \dots \cdot Q_{x_{2N-1}, y_{2N-1}} \cdot P_{x_{2N-1}, x} \geq \\ & \alpha_N \cdot \dots \cdot \alpha_N = \alpha_N^{2N} \end{aligned} \quad (3.33)$$

(at least one product in the sum is positive, and, by assumption, every term in this product is greater than or equal to  $\alpha_N$ ).

To prove (3.29) fix  $x \in \Omega^N, y_N, \dots, y_{2N-1}$ . Now

$$\begin{aligned} & \Pr(Y_N^\phi = y_N, \dots, Y_{2N-1}^\phi = y_{2N-1} | X_N^\phi = x) = \\ & \sum_{x_{N+1}, \dots, x_{2N-1} \in \Omega^N} Q_{x, y_N} \cdot P_{x, x_{N+1}} \cdot Q_{x_{N+1}, y_{N+1}} \cdot P_{x_{N+1}, x_{N+2}} \cdot \dots \cdot P_{x_{2N-2}, x_{2N-1}} Q_{x_{2N-1}, y_{2N-1}} \geq \\ & \alpha_N^N \sum_{x_{N+1}, \dots, x_{2N} \in \Omega^N} P_{z, x_{N+1}} \cdot \dots \cdot P_{x_{N+1}, x_{N+2}} \cdot \dots \cdot P_{x_{2N-2}, x_{2N-1}} > \alpha_N^{2N} \end{aligned} \quad (3.34)$$

(as the sum equals 1 !).

Eq.(3.30) is proven exactly as (3.28) and Eq.(3.31) is proven exactly as (3.29). This completes the proof of the lemma. •

**Remark:** The bounds above are not *sharp*; e.g. in (3.29) we could have used  $\alpha_N^N$ , rather than  $\alpha_N^{2N}$ ; however we choose  $\alpha_N^{2N}$  to simplify superscript counting the result.

### 3.3 Mixing Properties

In this section we establish the *mixing* properties of all HMM's in the class  $\Phi_N$ ,  $N = 1, 2, \dots$ . Mixing refers to the property of some stochastic processes to “forget” their past. As we have pointed out in Section 2.3, Theorem 18, for every stochastic process  $Y$ , with alphabet  $\Omega$  and associated probability measure  $\pi$ , we have for  $\pi$ -a.a.  $y = y_0 y_{-1} \dots \in \Omega^\infty$

$$\lim_{n \rightarrow \infty} Pr(Y_0 = y_0 | Y_{-1} = y_{-1}, \dots, Y_{-n} = y_{-n}) = Pr(Y_0 = y_0 | Y_{-1} = y_{-1}, Y_{-2} = y_{-2}, \dots). \quad (3.35)$$

Eq.(3.35) indicates that for any fixed sequence  $y = y_0 y_{-1} \dots \in \Omega^\infty$ , conditioning on the  $n$ -long past approaches conditioning on the infinite past as  $n$  gets big. Here we will prove something stronger, namely that for every  $N \geq 1$  the limit in (3.35) is uniform for all  $y$  and for all  $Y^\phi$ ,  $\phi \in \Phi_N$ .

We prove four lemmas in this section. The first two are technical and rather non-intuitive, but necessary for the proof of the third one, Lemma 26, which is our main goal. This lemma will be crucial for the proof of the Consistency Theorem 20. We also use this lemma to prove Lemma 27, which establishes uniform convergence of cross entropy, for all elements of class  $\Phi_N$ .

The mixing property depends only on the properties of class  $\Phi_N$ ,  $N = 1, 2, \dots$ ; in the proofs below we will use the positivity bounds inherent in the definition of  $\Phi_N$ , as well as the ones developed in the previous section.

**Lemma 24** *For all  $N \geq 1$ , define  $\beta_N \doteq \alpha_N^{8N} / K^N$ .<sup>1</sup> Then for all  $N \geq 1$ ,  $\forall \phi \in \Phi_N$ ,  $\forall n \geq 2$  and for all  $x, z \in \Omega^N$ ,  $y = y_1 y_2 \dots \in \Omega^\infty$  we have:*

$$Pr(X_N^\phi = z | X_0^\phi = x, Y_1^\phi = y_1, \dots, Y_{n \cdot N}^\phi = y_{n \cdot N}) \geq \beta_N. \quad (3.36)$$

**Proof:** Fix  $N, n, y, \phi$ . For the rest of the proof drop  $Y^\phi$  and  $\phi$  from the notation (for brevity).

Now I claim the following:

---

<sup>1</sup>Recall that  $K$  is the size of the alphabet  $\Omega$ , wherein all  $Y^\phi$  processes take values.

**Claim:** To prove (3.36) it suffices to prove

$$\forall x, z, u \in \Omega^N \quad \frac{Pr(X_N = z | X_0 = x, y_1, \dots, y_{n \cdot N})}{Pr(X_N = u | X_0 = x, y_1, \dots, y_{n \cdot N})} \leq \frac{1}{\alpha_N^{8N}}. \quad (3.37)$$

Let us first prove this claim. Assume (3.37) is true and (3.36) is wrong. That is, there are  $x, \bar{z} \in \Omega^N$  such that

$$Pr(X_N = \bar{z} | X_0 = x, y_1, \dots, y_{n \cdot N}) < \frac{\alpha_N^{8N}}{K^N}. \quad (3.38)$$

Define  $\tilde{z}$  by

$$\tilde{z} \doteq \arg \max_{z \in \Omega^N} Pr(X_N = z | X_0 = x, y_1, \dots, y_{n \cdot N}). \quad (3.39)$$

Then

$$\begin{aligned} & \frac{Pr(X_N = \tilde{z} | X_0 = x, y_1, \dots, y_{n \cdot N})}{Pr(X_N = \bar{z} | X_0 = x, y_1, \dots, y_{n \cdot N})} > \\ & \frac{Pr(X_N = \tilde{z} | X_0 = x, y_1, \dots, y_{n \cdot N})}{\alpha_N^{8N} / K^N} = \\ & \frac{K^N \cdot Pr(X_N = \tilde{z} | X_0 = x, y_1, \dots, y_{n \cdot N})}{\alpha_N^{8N}} \geq \\ & \frac{\sum_{z \in \Omega^N} Pr(X_N = z | X_0 = x, y_1, \dots, y_{n \cdot N})}{\alpha_N^{8N}} = \frac{1}{\alpha_N^{8N}}. \end{aligned} \quad (3.40)$$

But (3.40) contradicts (3.37)! This completes the proof of the claim. All that remains to prove the lemma, is to prove that (3.37) is true.

Let us prove (3.37). Choose any  $x, z, u \in \Omega^N$ . Then

$$\begin{aligned} & \frac{Pr(X_N = z | X_0 = x, y_1, \dots, y_{n \cdot N})}{Pr(X_N = u | X_0 = x, y_1, \dots, y_{n \cdot N})} = \\ & \frac{Pr(X_N = z, y_1, \dots, y_{n \cdot N} | X_0 = x)}{Pr(X_N = u, y_1, \dots, y_{n \cdot N} | X_0 = x)} = \\ & \frac{\sum_{v \in \Omega^N} Pr(X_{2N} = v, X_N = z, y_1, \dots, y_{n \cdot N} | X_0 = x)}{\sum_{v \in \Omega^N} Pr(X_{2N} = v, X_N = u, y_1, \dots, y_{n \cdot N} | X_0 = x)}. \end{aligned} \quad (3.41)$$

We want to bound (3.41) from above. To do so we need to define some auxiliary quantities. Define

for fixed  $z, u$  and for all  $v \in \Omega^N$

$$F_{zv} \doteq Pr(X_{2N} = v | y_N, \dots, y_{2N-1}, X_N = z) Pr(y_N, \dots, y_{2N-1} | X_N = z) Pr(X_N = z | y_1, \dots, y_N, X_0 = x), \quad (3.42)$$

$$F_{uv} \doteq Pr(X_{2N} = v | y_N, \dots, y_{2N-1}, X_N = u) Pr(y_N, \dots, y_{2N-1} | X_N = u) Pr(X_N = u | y_1, \dots, y_N, X_0 = x), \quad (3.43)$$

$$G_v \doteq Pr(y_{2N}, \dots, y_{nN} | X_{2N} = v). \quad (3.44)$$

Also define  $M \doteq \max_{v \in \Omega^N}$ . It follows from Lemma 23 that

$$M \leq \frac{1}{\alpha^{8N}}. \quad (3.45)$$

Using the definitions of  $F_{uv}, F_{zv}, G_v$  and the conditioning truncation properties proven in Lemma 22, it is easy to prove that (3.41) equals

$$\frac{\sum_{v \in \Omega^N} G_v F_{zv}}{\sum_{v \in \Omega^N} G_v F_{uv}}. \quad (3.46)$$

We will now show that (3.46) is less than or equal to  $M$ . This, in conjunction with (3.45) will complete the proof of the lemma.

$$\begin{aligned} 0 &= \sum_{v \in \Omega^N} G_v \left\{ F_{uv} \frac{F_{zv}}{F_{uv}} - F_{zv} \right\} \leq \sum_{v \in \Omega^N} G_v \{ F_{uv} \cdot M - F_{zv} \} \Rightarrow \\ &0 \leq M \cdot \sum_{v \in \Omega^N} G_v F_{uv} - \sum_{v \in \Omega^N} G_v F_{zv} \Rightarrow \\ &\frac{\sum_{v \in \Omega^N} G_v F_{zv}}{\sum_{v \in \Omega^N} G_v F_{uv}} \leq M. \end{aligned} \quad (3.47)$$

This completes the proof of the lemma. •

Now we use the previous lemma to prove Lemma 25. This lemma, which will be used in the proof of the main result of this section (Lemma 26), basically says that conditional probabilities

of any pair  $(X^\phi, Y^\phi)$  converge as the conditioning recedes to the infinite past. It is a modification of a standard argument used to prove the existence of a unique equilibrium probability for regular Markov Chains. The method of proof is also very similar to the one used by Baum and Petrie in [BP66].

**Lemma 25** For all  $N \geq 1$ ,  $\forall \phi \in \Phi_N$ ,  $\forall m \geq 2$  and for all  $y = y_0 y_{-1} y_{-2} \dots \in \Omega^\infty$ , define

$$D_m^\phi(y) \doteq \max_{x \in \Omega^N} Pr(Y_0^\phi = y_0 | Y_{-1}^\phi = y_{-1}, \dots, Y_{-mN}^\phi = y_{-mN}, X_{-mN}^\phi = x), \quad (3.48)$$

$$d_m^\phi(y) \doteq \min_{x \in \Omega^N} Pr(Y_0^\phi = y_0 | Y_{-1}^\phi = y_{-1}, \dots, Y_{-mN}^\phi = y_{-mN}, X_{-mN}^\phi = x) \quad (3.49)$$

Then, for all  $m \geq 2$ , we have

$$0 \leq D_m^\phi(y) - d_m^\phi(y) \leq (1 - 2\beta_N)^{m-1}. \quad (3.50)$$

**Remark:** As  $\alpha_N > 0$ ,  $\beta_N < 1$  and so  $(1 - 2\beta_N)^{m-1}$  goes to zero as  $m$  goes to infinity.

**Proof:** Fix  $N \geq 1$ ,  $\phi \in \Phi_N$  and  $y = y_0 y_{-1} \dots \in \Omega^N$ . For the rest of the proof drop  $Y^\phi$ ,  $\phi$  from the notation (for brevity).

Next, for all  $m \geq 2$ ,  $x, z \in \Omega^N$  define

$$p_{m,z}(y) \doteq Pr(y_0 | y_{-1}, y_{-2}, \dots, y_{-mN}, X_{-mN} = z), \quad (3.51)$$

$$q_{m,x,z}(y) \doteq Pr(X_{-mN} = z | y_{-1}, y_{-2}, \dots, y_{-(m+1)N}, X_{-(m+1)N} = x). \quad (3.52)$$

Note that, because of Lemma 22

$$p_{m,z}(y) = Pr(y_0 | y_{-1}, y_{-2}, \dots, y_{-(m+1)N}, X_{-mN} = z, X_{-(m+1)N} = x). \quad (3.53)$$

Also define

$$\mu_m(y) \doteq \min_{x,z \in \Omega^N} q_{m,x,z}(y), \quad (3.54)$$

$$\bar{z} \doteq \arg \min_{z \in \Omega^N} p_{m,z}(y), \quad (3.55)$$

$$\tilde{z} \doteq \arg \max_{z \in \Omega^N} p_{m,z}(y). \quad (3.56)$$

Now

$$\begin{aligned} Pr(y_0|y_{-1}y_{-2}, \dots, y_{-(m+1)N}, X_{-(m+1)N} = x) &= \\ &= \sum_{z \in \Omega^N} q_{m,x,z}(y) \cdot p_{m,z}(y) = \\ &= \left\{ \sum_{z \in \Omega^N, z \neq \bar{z}} q_{m,x,z}(y) \cdot p_{m,z}(y) \right\} + \{q_{m,x,\bar{z}} - \mu_m(y)\} \cdot p_{m,\bar{z}}(y) + \mu_m(y) \cdot p_{m,\bar{z}}(y) \leq \\ &= D_m(y) \cdot \left\{ \sum_{z \in \Omega^N, z \neq \bar{z}} q_{m,x,z}(y) \right\} + \{q_{m,x,\bar{z}} - \mu_m(y)\} \cdot D_m(y) + \mu_m(y) \cdot d_m(y) = \\ &= D_m(y) \cdot \sum_{z \in \Omega^N} q_{m,x,z}(y) p_{m,z}(y) - D_m(y) \cdot \mu_m(y) + \mu_m(y) \cdot d_m(y) = \\ &= (1 - \mu_m(y)) \cdot D_m(y) + \mu_m(y) \cdot d_m(y) \Rightarrow \end{aligned}$$

$$Pr(y_0|y_{-1}y_{-2}, \dots, y_{-(m+1)N}, X_{-(m+1)N} = x) \leq (1 - \mu_m(y)) \cdot D_m(y) + \mu_m(y) \cdot d_m(y). \quad (3.57)$$

In exactly the same way (using  $\tilde{z}$  instead of  $\bar{z}$ ) we can prove

$$Pr(y_0|y_{-1}y_{-2}, \dots, y_{-(m+1)N}, X_{-(m+1)N} = x) \geq (1 - \mu_m(y)) \cdot d_m(y) + \mu_m(y) \cdot D_m(y). \quad (3.58)$$

Taking the maximum / minimum in (3.57) / (3.58), we get:

$$D_{m+1}(y) \leq (1 - \mu_m(y)) \cdot D_m(y) + \mu_m(y) \cdot d_m(y), \quad (3.59)$$

$$d_{m+1}(y) \geq (1 - \mu_m(y)) \cdot d_m(y) + \mu_m(y) \cdot D_m(y). \quad (3.60)$$

Subtracting (3.60) from (3.59) we get

$$0 \leq D_{m+1}(y) - d_{m+1}(y) \leq (1 - 2\mu_m(y)) \cdot (D_m(y) - d_m(y)). \quad (3.61)$$

From Lemma 24 we know that, for all  $m \geq 2$ , for all  $y \in \Omega^\infty$ , we have  $\mu_m(y) \leq \beta_N$ . This, together with (3.61) gives:

$$0 \leq D_{m+1}(y) - d_{m+1}(y) \leq (1 - 2\beta_N) \cdot (D_m(y) - d_m(y)). \quad (3.62)$$

Finally, repeating this argument for  $m, m-1, \dots, 2$  we get

$$0 \leq D_{m+1}(y) - d_{m+1}(y) \leq (1 - 2\beta_N)^{m-1} \cdot (D_2(y) - d_2(y)) \leq (1 - 2\beta_N)^{m-1} \cdot . \quad (3.63)$$

This completes the proof of the lemma. •

Finally we get to the mixing result which has been our main goal.

**Lemma 26** (*Mixing*) *For all  $N \geq 1$  there are constants  $C_N, D_N$  such that for all  $\phi \in \Phi_N, k, l$  such that  $k \geq l \geq mN, y = y_0 y_{-1} y_{-2} \dots \in \Omega^\infty$*

$$|p_\phi(y_0|y_{-1} \dots y_{-k}) - p_\phi(y_0|y_{-1} \dots y_{-l})| < C_N \cdot (1 - 2\beta_N)^m, \quad (3.64)$$

$$\left| \log \frac{p_\phi(y_0|y_{-1} \dots y_{-k})}{p_\phi(y_0|y_{-1} \dots y_{-l})} \right| < D_N \cdot (1 - 2\beta_N)^m \quad (3.65)$$

**Proof:** Fix  $N \geq 1, \phi \in \Phi_N$ . For the rest of the proof drop  $Y^\phi$  and  $\phi$  from the notation (for brevity). Using Lemma 22 for the truncation of conditional probabilities, we have

$$p_\phi(y_0|y_{-1}, \dots, y_{-k}) = \sum_{x \in \Omega^N} Pr(y_0|y_{-1}, \dots, y_{-mN}, X_{-mN} = x) \cdot Pr(X_{-mN} = x|y_{-1}, \dots, y_{-mN}). \quad (3.66)$$

From Lemma 25,  $Pr(y_0|y_{-1}, \dots, y_{-mN}, X_{-mN} = x)$  lies between  $d_m(y)$  and  $D_m(y)$ . This and

(3.66) imply

$$d_m(y) \leq p_\phi(y_0|y_{-1}, \dots, y_{-mN}) \leq D_m(y). \quad (3.67)$$

Similarly,

$$p_\phi(y_0|y_{-1}, \dots, y_{-(m+1)N}) = \sum_{x \in \Omega^N} Pr(y_0|y_{-1}, \dots, y_{-mN}, X_{-mN} = x) Pr(X_{-mN} = x|y_{-1}, \dots, y_{-(m+1)N}). \quad (3.68)$$

From Lemma 25,  $Pr(y_0|y_{-1}, \dots, y_{-mN}, X_{-mN} = x)$  lies between  $d_m(y)$  and  $D_m(y)$ . This and (3.68) imply

$$d_m(y) \leq p_\phi(y_0|y_{-1}, \dots, y_{-(m+1)N}) \leq D_m(y). \quad (3.69)$$

Combining (3.67) and (3.52), and using Lemma 25 we get

$$|p_\phi(y_0|y_{-1}, \dots, y_{-mN}) - p_\phi(y_0|y_{-1}, \dots, y_{-(m+1)N})| \leq (1 - 2\beta_N)^{m-1}. \quad (3.70)$$

Similarly

$$|p_\phi(y_0|y_{-1}, \dots, y_{-(m+1)N}) - p_\phi(y_0|y_{-1}, \dots, y_{-(m+2)N})| \leq (1 - 2\beta_N)^m. \quad (3.71)$$

...

$$|p_\phi(y_0|y_{-1}, \dots, y_{-nN}) - p_\phi(y_0|y_{-1}, \dots, y_{-k})| \leq (1 - 2\beta_N)^{n-1}. \quad (3.72)$$

(In (3.72)  $n$  is the largest integer such that  $nN < k$ .) Combining (3.70), (3.71), ... , (3.72), we get

$$\begin{aligned} |p_\phi(y_0|y_{-1}, \dots, y_{-k}) - p_\phi(y_0|y_{-1}, \dots, y_{-l})| &\leq ((1 - 2\beta_N)^{m-1} + (1 - 2\beta_N)^m + \dots + (1 - 2\beta_N)^{n-1}) < \\ &(1 - 2\beta_N)^m \cdot \left( \frac{1 + (1 - 2\beta_N) + (1 - 2\beta_N)^2 + \dots}{1 - 2\beta_N} \right). \end{aligned} \quad (3.73)$$

Defining  $C_N \doteq (1 + (1 - 2\beta_N) + (1 - 2\beta_N)^2 + \dots) / (1 - 2\beta_N)$ , we have completed the proof of (3.64).

To prove (3.65) we use the Mean Value Theorem. We have

$$\left| \log \frac{p_\phi(y_0|y_{-1}, \dots, y_{-k})}{p_\phi(y_0|y_{-1}, \dots, y_{-l})} \right| < \frac{1}{\zeta} |p_\phi(y_0|y_{-1}, \dots, y_{-k}) - p_\phi(y_0|y_{-1}, \dots, y_{-l})|, \quad (3.74)$$

with  $\zeta$  lying between  $p_\phi(y_0|y_{-1}, \dots, y_{-k})$  and  $p_\phi(y_0|y_{-1}, \dots, y_{-l})$ . By Lemma 23, both  $p_\phi(y_0|y_{-1}, \dots, y_{-k})$  and  $p_\phi(y_0|y_{-1}, \dots, y_{-l})$  are greater than  $\alpha_N$ . Defining  $D_N \doteq C_N/\alpha_N$  and using (3.64) yields (3.65) and completes the proof of the lemma.  $\bullet$

**Remark:** An immediate consequence of Lemma 26, that is very easy to prove is the following: for all  $N \geq 1$ ,  $\phi \in \Phi_N$ ,  $k \geq mN$  and for all  $y_0, y_{-1}, \dots$ , we have

$$|p_\phi(y_0|y_{-1}\dots y_{-k}) - p_\phi(y_0|y_{-1}y_{-2}\dots)| < C_N \cdot (1 - 2\beta_N)^m. \quad (3.75)$$

From this follows immediately that for all  $N \geq 1$ ,  $\phi \in \Phi_N$ ,  $k \geq mN$  and for all  $y, z$  in  $\Omega^\infty$  with  $y_0 = z_0, \dots, y_{-k} = z_{-k}$ , we have

$$|p_\phi(y_0|y_{-1}y_{-2}\dots) - p_\phi(z_0|z_{-1}z_{-2}\dots)| < C_N \cdot (1 - 2\beta_N)^m, \quad (3.76)$$

which specifies the continuity behavior of  $p_\phi(y_0|y)$ . As we have already indicated mixing and continuity properties are very closely related.

**Remark:** Another consequence of Lemma 26 is that we can approximate any process  $Y^\phi$ ,  $\phi \in \Phi_N$  with HGMs, and the approximation is both in the cross entropy *and* the weak sense. (Recall that in Section 2.2 we only proved approximation in the weak sense by HGMs.) We will not pursue this here, but the proof is easy.

Lemma 26 will be used extensively in Section 3.5. However we will also use it here to prove a property of cross entropy convergence, which hold for every  $(X^\phi, Y^\phi)$ ,  $\phi \in \Phi^N$ .

**Lemma 27** Under the assumptions of Theorem 20, for all  $N \geq 1$  and for all  $\phi \in \Phi_N$

$$\lim_{n \rightarrow \infty} H_n(p_\phi; p) = \int \log \frac{p(y_0|y_{-1}y_{-2}\dots)}{p_\phi(y_0|y_{-1}y_{-2}\dots)} d\pi(y_0y_{-1}y_{-2}\dots). \quad (3.77)$$

Then, by definition,

$$H(p_\phi; p) = \int \log \frac{p(y_0|y_{-1}y_{-2}\dots)}{p_\phi(y_0|y_{-1}y_{-2}\dots)} d\pi(y_0y_{-1}y_{-2}\dots). \quad (3.78)$$

**Proof:** Fix  $N \geq 1$  and  $\phi \in \Phi_N$ . As the conditional probabilities of both  $p$  and  $p_\phi$  are always positive, the  $n$ -th order cross entropy of  $p_\phi$  with respect to  $p$  equals

$$H_n(p_\phi; p) = \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0|y_{-1} \dots y_{-n}) - \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p_\phi(y_0|y_{-1} \dots y_{-n}). \quad (3.79)$$

We will show that

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0|y_{-1} \dots y_{-n}) = \int \log p(y_0|y_{-1}y_{-2}\dots) d\pi(y_0y_{-1}y_{-2}\dots) \quad (3.80)$$

and

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p_\phi(y_0|y_{-1} \dots y_{-n}) = \int \log p_\phi(y_0|y_{-1}y_{-2}\dots) d\pi(y_0y_{-1}y_{-2}\dots). \quad (3.81)$$

Then (3.79) follows immediately from (3.80), (3.81). In fact, recalling that, by definition,  $H(p_\phi; p) = \lim_n H_n(p_\phi; p)$ , we can write

$$H(p_\phi; p) = \int \log \frac{p(y_0|y_{-1}y_{-2}\dots)}{p_\phi(y_0|y_{-1}y_{-2}\dots)} d\pi(y_0y_{-1}y_{-2}\dots). \quad (3.82)$$

Let us now prove (3.80). First of all, note that for fixed  $n$  and  $y_0, y_{-1}, \dots, y_{-n}$ , the function  $f(y) = p(y_0|y_{-1} \dots y_{-n})$  (with  $y = y_0y_{-1} \dots y_{-n}y_{-n-1} \dots$ ) is independent of  $y_{-n-1}, y_{-n-2}, \dots$ .

Therefore,

$$\sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0 | y_{-1} \dots y_{-n}) = \int \log p(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots). \quad (3.83)$$

Furthermore, by Theorem 18 (convergence of conditional probability) and the continuity of the logarithm function, for all  $y = y_0 y_{-1} y_{-2} \dots$  we have

$$\lim_{n \rightarrow \infty} \log p(y_0 | y_{-1} \dots y_{-n}) = \log p(y_0 | y_{-1} y_{-2} \dots). \quad (3.84)$$

Finally, by assumption (eq.(3.11) in Theorem 20), for all  $n$  and all  $y = y_0 y_{-1} y_{-2} \dots$  we have

$$p(y_0 | y_{-1} \dots y_{-n}) \geq \alpha > 0, \quad (3.85)$$

which implies

$$0 \geq \log p(y_0 | y_{-1} \dots y_{-n}) \geq \log \alpha, \quad (3.86)$$

From (3.83), (3.84), (3.86) and the Dominated Convergence Theorem we conclude that

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0 | y_{-1} \dots y_{-n}) = \int \lim_{n \rightarrow \infty} \log p(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots) \Rightarrow \quad (3.87)$$

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0 | y_{-1} \dots y_{-n}) = \int \log p(y_0 | y_{-1} y_{-2} \dots) d\pi(y_0 y_{-1} y_{-2} \dots). \quad (3.88)$$

This completes the proof of (3.80).

The proof of (3.81) is very similar. There are two differences. First of all, the bound on  $p_\phi(y_0 | y_{-1} \dots y_{-n})$  is  $\alpha_N$  (rather than  $\alpha$ ). Second, we will compute an explicit rate of convergence, which will be uniform for all  $\phi \in \Phi_N$ .

Once again, note that, for fixed  $\phi$ ,  $n$  and  $y_0, y_{-1}, \dots, y_{-n}$ , the function  $f(y) = p_\phi(y_0 | y_{-1} \dots y_{-n})$

(with  $y = y_0 y_{-1} \dots y_{-n} y_{-n-1} \dots$ ) is independent of  $y_{-n-1}, y_{-n-2}, \dots$ . Therefore,

$$\sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p_\phi(y_0 | y_{-1} \dots y_{-n}) = \int \log p_\phi(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots). \quad (3.89)$$

Furthermore, by Lemma 26 (Mixing Lemma), for all  $y_0 y_{-1} y_{-2} \dots$ , and all  $k, n \geq mN$  ( $m \geq 2$ ) we have

$$\log p_\phi(y_0 | y_{-1} \dots y_{-k}) - D_N \cdot (1 - 2\beta_N)^m < \log p_\phi(y_0 | y_{-1} \dots y_{-n}) < \log p_\phi(y_0 | y_{-1} \dots y_{-k}) + D_N \cdot (1 - 2\beta_N)^m. \quad (3.90)$$

Letting  $k \rightarrow \infty$  and integrating, we get

$$\begin{aligned} \int \log p_\phi(y_0 | y_{-1} y_{-2} \dots) d\pi(y_0 y_{-1} y_{-2} \dots) - D_N \cdot (1 - 2\beta_N)^m < \\ \int \log p_\phi(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots) < \\ \int \log p_\phi(y_0 | y_{-1} y_{-2} \dots) d\pi(y_0 y_{-1} y_{-2} \dots) + D_N \cdot (1 - 2\beta_N)^m. \end{aligned} \quad (3.91)$$

Finally, by Lemma 23 for all  $y = y_0 y_{-1} y_{-2} \dots$  we have

$$p_\phi(y_0 | y_{-1} \dots y_{-n}) \geq \alpha_N > 0. \quad (3.92)$$

which implies

$$0 \geq \log p(y_0 | y_{-1} \dots y_{-n}) \geq \log \alpha_N, \quad (3.93)$$

From (3.89), (3.91), (3.93) and the Dominated Convergence Theorem we conclude that

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p_\phi(y_0 | y_{-1} \dots y_{-n}) = \int \lim_{n \rightarrow \infty} \log p_\phi(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots) \Rightarrow \quad (3.94)$$

$$\lim_{n \rightarrow \infty} \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0 | y_{-1} \dots y_{-n}) = \int \log p(y_0 | y_{-1} y_{-2} \dots) d\pi(y_0 y_{-1} y_{-2} \dots) \quad (3.95)$$

and in fact, using (3.91), we get

$$\left| \sum_{y_{-n}, \dots, y_0 \in \Omega} p(y_{-n} \dots y_0) \cdot \log p(y_0 | y_{-1} \dots y_{-n}) - \int \log p(y_0 | y_{-1} \dots) d\pi(y_0 y_{-1} \dots) \right| < D_N \cdot (1 - 2\beta_N)^{n/N}. \quad (3.96)$$

This completes the proof of (3.81) and of the lemma. •

### 3.4 An Auxiliary Hidden Markov Model

In this section we introduce a sequence of auxiliary SNM's,  $(\bar{X}^N, \bar{Y}^N)$ ,  $N = 1, 2, \dots$ , where  $(\bar{X}^N, \bar{Y}^N)$  has transition matrix  $\bar{P}_N$  and emission matrix  $\bar{Q}_N$ ;  $(\bar{P}_N, \bar{Q}_N)$  is a member of  $\Phi_N$ . This sequence of SNM's is matched to a specific process  $Y$ . We denote by  $(\bar{X}^N, \bar{Y}^N)$  the (unique) SNM that has probability transition matrix  $\bar{P}_N$  and probability emission matrix  $\bar{Q}_N$ . We call these models *Noisy Autoregressive* (NAR) models; as the name suggests, the NAR models are *noisy* versions of the AR models of Chapter 2. By “noisy” we mean that the observable process is a probabilistic observation of the state process.

The NARs are constructed in such a way that the noisy observation effect diminishes as  $N$  goes to infinity. Thus the  $N$ -th order NAR approaches the  $N$ -th order AR as  $N$  goes to infinity. As the AR sequence converges to the original process in the cross entropy sense, so does the NAR sequence.

On the other hand the  $N$ -th order is in  $\Phi_N$ , so any  $N$ -th order ML estimate has better Likelihood than the  $N$ -th order NAR. As will be shown in Section 3.5, the  $N$ -th order Likelihood and cross entropy values get closer as  $N$  grows, so ML estimates must be closer to  $Y$  in the cross entropy sense than NARs. But the NARs tend to  $Y$  in the cross entropy sense by the argument of the previous paragraph, so the ML estimates also tend to  $Y$ .

The argument of the previous paragraph will be developed in a rigorous manner in Section 3.5. In this section we will define the NARs and show that the NAR sequence  $(\bar{X}^N, \bar{Y}^N)$  converges to  $Y$  in the cross entropy sense as  $N$  goes to infinity.

**Definition 48** Given a positive stationary stochastic process  $Y$ , with finite alphabet  $\Omega = \{0, 1, \dots, K-1\}$  and probability function  $p$ , the  **$N$ -th order Noisy Autoregressive Model** of  $Y$  (for short NAR model) is defined to be the unique SNM  $(\bar{X}^N, \bar{Y}^N)$  that satisfies the following

1.  $\bar{X}^N$  is the unique Markov process with transition probability matrix  $\bar{P}_N$ , where  $\bar{P}_N$  is defined for all  $y_1, \dots, y_N, z_1, \dots, z_N \in \Omega$  by

$$[\bar{P}_N]_{y_1 \dots y_N, z_1 \dots z_N} \doteq \begin{cases} \frac{p(y_1 \dots y_N z_N)}{p(y_1 \dots y_N)} & \text{when } y_2 = z_1, \dots, y_N = z_{N-1} \\ 0 & \text{otherwise.} \end{cases} \quad (3.97)$$

2. For  $t = 0, \pm 1, \pm 2, \dots$ , and for all  $y_1, \dots, y_N, y \in \Omega$   $(\bar{X}^N, \bar{Y}^N)$  satisfies

$$Pr(\bar{Y}_t^N = y | \bar{X}_t^N = y_1 \dots y_N) = [\bar{Q}_N]_{y_1 \dots y_N, y}, \quad (3.98)$$

where the emission probability matrix  $\bar{Q}_N$  is defined for all  $y_1, \dots, y_N, y \in \Omega$  by

$$[\bar{Q}_N]_{y_1 \dots y_N, y} \doteq \begin{cases} \frac{1}{N \cdot (K-1)} & \text{when } y_1 \neq y \\ 1 - \frac{1}{N} & \text{when } y_1 = y \end{cases} \quad (3.99)$$

The observable probability function of  $\bar{Y}^N$  is denoted by  $\bar{p}_N$ .

**Remark:** Note that the AR and NAR  $N$ -th order models of a process have the same transition probability matrix:  $P_N = \bar{P}_N$ .

**Remark:** As we only consider NAR models of positive processes  $Y$ , the corresponding matrix  $\bar{P}_N$  is regular and hence the  $N$ -th order NAR model of  $Y$  is uniquely defined.

**Remark:** As we usually have a specific original process  $Y$  in mind, we will just talk about the  $N$ -th order NAR model  $(\bar{X}_N, \bar{Y}_N)$ , with the understanding it is the NAR model of  $Y$ .

**Remark:** If a process  $Y$  satisfies the positivity assumption (3.11) of Theorem 20, with positivity bound  $\alpha$ , there is an integer  $N_\alpha$  such that, for all  $N \geq N_\alpha$ ,  $(\bar{P}_N, \bar{Q}_N)$  belongs to  $\Phi_N$ .

For the proof of the next Lemma we need some new notation. Given the original process

$Y$ , with probability function  $p$  and associated measure  $\pi$ , for any conditional probability  $q(\cdot|\cdot)$  :  $\Omega \times \Omega^\infty \mapsto [0, 1]$ , define

$$\forall n \geq 1 \quad E_n(q) \doteq \sum_{y_0, y_{-1}, \dots, y_{-n} \in \Omega} p(y_{-n} \dots y_{-1} y_0) \cdot \log q(y_0 | y_{-1} \dots y_{-n}), \quad (3.100)$$

$$\forall n \geq 1 \quad e_n(q) \doteq \frac{\sum_{y_1, y_2, \dots, y_n \in \Omega} p(y_1 \dots y_n) \cdot \log q(y_1 \dots y_n)}{n}. \quad (3.101)$$

$$E(q) \doteq \int \log q(y_0 | y_{-1} \dots y_{-n}) d\pi(y_0 y_{-1} y_{-2} \dots). \quad (3.102)$$

**Lemma 28** *Under the assumptions of Theorem 20*

$$\lim_{N \rightarrow \infty} H(\bar{p}_N; p) = 0. \quad (3.103)$$

**Proof:** Note that  $H(\bar{p}_N, p)$  equals by definition  $\lim_{n \rightarrow \infty} H_n(\bar{p}_N, p)$ ; also  $-H_n(\bar{p}_N; p) = E(\bar{p}_N) - E(p)$ . Hence, eq.(3.103) is equivalent to

$$\lim_{N \rightarrow \infty} \lim_{n \rightarrow \infty} \{E_n(\bar{p}_N) - E_n(p)\} = 0, \quad (3.104)$$

assuming that the limits exist. On the other hand

$$\{E_n(\bar{p}_N) - E_n(p)\} \geq$$

$$\{E_n(\bar{p}_N) - e_n(\bar{p}_N)\} + \{e_n(\bar{p}_N) - e_n(p_N)\} + \{e_n(p_N) - E_N(p)\} + \{E_N(p) - E_n(p)\}. \quad (3.105)$$

(Here  $p_N$  denotes the observable probability function of the  $N$ -th order AR model of Section 2.1.)

Therefore, to prove (3.104) we will consider separately each of the four summands in the rightmost side of (3.105). Namely, we will show that each summand goes to zero as  $n$  and  $N$  go to infinity.

**First**, take  $\{E_n(\bar{p}_N) - e_n(\bar{p}_N)\}$ . We know that, for all  $N \geq N_\alpha$ ,  $(\bar{P}_N, \bar{Q}_N) \in \Phi_N$ . Therefore,

from Lemma 27, for all  $N \geq N_\alpha$  we have

$$\lim_{n \rightarrow \infty} E_n(\bar{p}_N) = E(\bar{p}_N). \quad (3.106)$$

On the other hand

$$e_n(\bar{p}_N) = \frac{E_{n-1}(\bar{p}_N) + E_{n-2}(\bar{p}_N) + \dots + E_1(\bar{p}_N) + e_1(\bar{p}_N)}{n}. \quad (3.107)$$

From (3.106) and (3.107) we get

$$\lim_{n \rightarrow \infty} e_n(\bar{p}_N) = E(\bar{p}_N). \quad (3.108)$$

As  $E_n(\bar{p}_N)$  and  $e_n(\bar{p}_N)$  have a common limit, we also have for all  $N \geq 1$

$$\lim_{n \rightarrow \infty} \{E_n(\bar{p}_N) - e_n(\bar{p}_N)\} = 0. \quad (3.109)$$

**Second**, take  $\{e_n(\bar{p}_N) - e_n(p_N)\}$ . Recall that  $(\bar{X}^N, \bar{Y}^N)$  is the  $N$ -th order NAR model of  $Y$  and that  $(X^N, Y^N)$  is the  $N$ -th order AR model of  $Y$ . Denote by  $\bar{p}_N^X$  the probability function of the state process  $\bar{X}^N$ , by  $\bar{p}_N$  the probability function of the observable process  $\bar{Y}^N$ , by  $p_N^X$  the probability function of the state process  $X^N$  and by  $p_N$  the probability function of the observable process  $Y^N$ . We have

$$\begin{aligned} e_n(\bar{p}_N) &= \frac{\sum_{y_1, \dots, y_n \in \Omega} p(y_1 \dots y_n) \log \bar{p}_N(y_1 \dots y_n)}{n} = \\ &= \frac{\sum_{y_1, \dots, y_n \in \Omega} \log \left( \sum_{x_1, \dots, x_n \in \Omega^N} \bar{p}_N^X(x_1 \dots x_n) \cdot [\bar{Q}_N]_{x_1, y_1} \cdot \dots \cdot [\bar{Q}_N]_{x_n, y_n} \right)}{n}. \end{aligned} \quad (3.110)$$

Now

$$\begin{aligned} &\sum_{x_1, \dots, x_n \in \Omega^N} \bar{p}_N^X(x_1 \dots x_n) \cdot [\bar{Q}_N]_{x_1, y_1} \cdot \dots \cdot [\bar{Q}_N]_{x_n, y_n} \geq \\ &\sum_{u_1, \dots, u_n \in \Omega} \bar{p}_N^X(y_1 \dots y_n) \cdot [\bar{Q}_N]_{y_1 \dots y_n, y_1} \cdot \dots \cdot [\bar{P}_N]_{y_{n-1} y_n \dots u_{N-2}, y_n u_1 \dots u_{N-1}} \cdot [\bar{Q}_N]_{y_n u_1 \dots u_{N-1}, y_n} = \end{aligned}$$

$$\bar{p}_N^X(y_1 \dots y_N, y_2 \dots y_{N+1}, \dots, y_{n-N+1} \dots y_n) \cdot \left(1 - \frac{1}{N}\right)^n. \quad (3.111)$$

Combining (3.110) and (3.111) we get

$$e_n(\bar{p}_N) \geq \frac{\sum_{y_1, \dots, y_n \in \Omega} p(y_1 \dots y_n) \log \bar{p}_N^X(y_1 \dots y_n, \dots, y_{n-N+1} \dots y_n)}{n} + \log\left(1 - \frac{1}{N}\right). \quad (3.112)$$

But the  $N$ -th order AR model and NAR model have the same transition matrix ( $P_N = \bar{P}_N$ ) and hence the same state probability function ( $p_N^X = \bar{p}_N^X$ ). Also, from Lemma 2.1,

$$\begin{aligned} p_N^X(y_1 \dots y_n, \dots, y_{n-N+1} \dots y_n) &= p_N(y_1 \dots y_n) \Rightarrow \\ \bar{p}_N^X(y_1 \dots y_n, \dots, y_{n-N+1} \dots y_n) &= p_N(y_1 \dots y_n). \end{aligned} \quad (3.113)$$

Substituting this into (3.112) we get

$$\begin{aligned} e_n(\bar{p}_N) &\geq \frac{\sum_{y_1, \dots, y_n \in \Omega} p(y_1 \dots y_n) \log p_N(y_1 \dots y_n, \dots, y_{n-N+1} \dots y_n)}{n} + \log\left(1 - \frac{1}{N}\right) \Rightarrow \\ e_n(\bar{p}_N) &\geq e_n(p_N) + \log\left(1 - \frac{1}{N}\right). \end{aligned} \quad (3.114)$$

On the other hand, from the cross entropy inequality  $H_n(\bar{p}_N; p) \geq 0$  we obtain

$$e_n(p) \geq e_n(\bar{p}_N). \quad (3.115)$$

Combining (3.114) and (3.115) we get

$$e_n(p) \geq e_n(\bar{p}_N) \geq e_n(p_N) + \log\left(1 - \frac{1}{N}\right). \quad (3.116)$$

Now, from Theorem 13,  $\lim_n e_n(p_N) = E(p_N)$  and  $\lim_N E(p_N) = E(p)$ . Therefore

$$\lim_{N \rightarrow \infty} \lim_{n \rightarrow \infty} e_n(p_N) = E(p). \quad (3.117)$$

Similarly,

$$\lim_{n \rightarrow \infty} e_n(p) = E(p). \quad (3.118)$$

Combining (3.116) , (3.117), (3.118) we get

$$\lim_{N \rightarrow \infty} \left\{ \lim_{n \rightarrow \infty} e_n(\bar{p}_N) - e_n(p_N) \right\} = 0. \quad (3.119)$$

**Third**, take  $\{e_n(p_N) - E_N(p)\}$ . By Lemma 12, for all  $n \geq N$ ,  $E_n(p_N) = E_N(p)$ . This, together with

$$e_n(p_N) = \frac{E_{n-1}(p_N) + E_{n-2}(p_N) + \dots + E_1(p_N) + e_1(p_N)}{n} \quad (3.120)$$

imply

$$\lim_{n \rightarrow \infty} \{e_n(p_N) - E_N(p)\} = 0. \quad (3.121)$$

**Fourth**, take  $\{E_N(p) - E_n(p)\}$ . Note that, in fact,  $E_N(p) = -H_N(p)$  and  $E_n(p) = -H_n(p)$ .

Therefore we have

$$\lim_{N \rightarrow \infty} \lim_{n \rightarrow \infty} \{E_N(p) - E_n(p)\} = 0. \quad (3.122)$$

Now, in (3.105) take first the  $n$  limit and then the  $N$  limit. Substituting appropriately from (3.109), (3.119), (3.121) and (3.122) we get

$$\lim_{N \rightarrow \infty} \lim_{n \rightarrow \infty} \{E_n(\bar{p}_N) - E_n(p)\} = 0; \quad (3.123)$$

this completes the proof of the lemma. •

### 3.5 Consistency

In this section we prove the consistency theorem stated in Section 3.1. For completeness, we repeat the statement of the theorem here.

**Theorem 1** (*Consistency*) *Given a stationary ergodic stochastic process  $Y$ , with finite alphabet*

$\Omega = \{0, 1, \dots, K - 1\}$ , probability function  $p$  and associated probability measure  $\pi$ , assume that

$$\exists \alpha > 0 \text{ such that } \forall m \geq 1, y_{-m}, \dots, y_{-1}, y_0 \in \Omega \quad p(y_0 | y_{-1} \dots y_{-m}) \geq \alpha. \quad (3.124)$$

Then, there is a sequence  $N(n)$ ,  $n = 1, 2, \dots$ ,  $N(n) \uparrow \infty$ , such that

$$\lim_{n \rightarrow \infty} \left\{ \sup_{\phi \in \Psi_{N,n}(y)} H(p_\phi; p) \right\} \stackrel{\text{in prob.}}{=} 0. \quad (3.125)$$

As we will use several quantities defined in previous sections, we list them here to refresh the reader's memory.

- $\Omega$  is the alphabet of the original process  $Y$  that we want to model.
- $K$  is the size of  $\Omega$ .
- $\Phi_N(\Omega)$  is the  $N$ -th class of  $(P, Q)$  parameters that describe a SNM model, defined in Definition 46. It will be, from now on, abbreviated as  $\Phi_N$ , as  $Y$  and its alphabet  $\Omega$  are fixed.
- $\Psi_{N,n}(y)$  is the set of ML estimates in the  $\Phi_N$ , defined in Definition 47.
- $N_\alpha$  is an integer dependent only on the positivity bound of the original process  $Y$  (see (3.11)).  $N_\alpha$  has the following property: for all  $N \geq N_\alpha$ , the  $N$ -th order NAR model (see Lemma 28) is in  $\Phi_N$ .
- $\alpha_N$  is the positivity bound on the elements of class  $\Phi_N$ . For all  $N \geq 1$  we have  $\alpha_N > 0$ .
- $\beta_N$  was defined to be equal to  $(1 - \alpha_N^{8N})$ . It is important because it is *mixing constant* of class  $\Phi_N$ . That is, dependence on the greater-than- $mN$  past decays as  $(1 - \beta_N)^m$ , which goes to zero as  $m$  goes to infinity.
- $D_N$  is a constant associated with the mixing rate, defined in Lemma 26.
- $p_N$  is the probability function of the  $N$ -th order AR model, defined in Definition 40.
- $\bar{p}_N$  is the probability function of the  $N$ -th order NAR model, defined in Definition 48.

- $p_{N,n}^y$  is the  $(N, n)$ -th order empirical probability of sample  $y_1, \dots, y_n$ , defined in Definition 42.
- $E_n(\cdot)$  is an expectation function defined for all  $n \geq 1$  in eq.(3.100).
- $e_n(\cdot)$  is an expectation function defined for all  $n \geq 1$  in eq.(3.101).
- $E(\cdot)$  is an expectation function defined in eq.(3.102).

To prove this theorem, we need an auxiliary lemma.

**Lemma 2** *Under the conditions of Theorem 1, there are*

1. a sequence of integers  $\{n_N\}_{N=1}^{\infty}$  with  $n_N \uparrow \infty$ ,
2. a sequence of reals  $\{\zeta_N\}_{N=1}^{\infty}$  with  $\zeta_N \downarrow 0$ ,
3. a double sequence of sets  $\{E_{Nn}\}_{N,n=1}^{\infty}$  with  $E_{Nn} \subset \Omega^\infty$  for all  $N, n \geq 1$ ,

that satisfy for all  $N \geq N_\alpha$ ,  $n \geq n_N$ ,  $y \in E_{Nn}$

1.  $\pi(E_{Nn}^c) \leq 2^{-N}$ .
2.  $\sup_{\phi \in \Psi_{N,n}(y)} H(p_\phi; p) < \zeta_N$ .

The bulk of the argument is presented in the proof of Lemma 2. Once this Lemma is established, we only need a diagonalization argument (for the sequence  $\{n_N\}_{N=1}^{\infty}$ ) to obtain the proof of Theorem 1.

**Proof of Lemma 2** We want to show that, for every  $N \geq N_a$ , for large enough  $n$  and all  $y$  outside of a set of small probability, we have

$$\sup_{\phi \in \Psi_{N,n}(y)} |E(p) - E(p_\phi)| < \zeta_N. \quad (3.126)$$

From the entropy inequality  $H(p_\phi; p) \geq 0$  we know

$$E(p) - E(p_\phi) \geq 0. \quad (3.127)$$

On the other hand, for any integer  $k_N \geq 1$  we have

$$\begin{aligned}
& E(p) - E(p_\phi) = \\
& \{E(p) - E(\bar{p}_N)\} + \{E(\bar{p}_N) - E_{k_N}(\bar{p}_N)\} + \{E_{k_N}(\bar{p}_N) - E_{k_N}(p_\phi)\} + \{E_{k_N}(p_\phi) - E(p_\phi)\} \leq \\
& |E(p) - E(\bar{p}_N)| + |E(\bar{p}_N) - E_{k_N}(\bar{p}_N)| + |E_{k_N}(\bar{p}_N) - E_{k_N}(p_\phi)| + |E_{k_N}(p_\phi) - E(p_\phi)|. \quad (3.128)
\end{aligned}$$

We will select the sequence  $k_N$ ,  $N = 1, 2, \dots$  so that every term in (3.128) will be less than a number  $\theta_N$ , with  $\theta_N \downarrow 0$ . Then, (3.127) and (3.128) imply (3.126) (with  $\zeta_N = 4\theta_N$ ), and we are done.

Define

$$\forall N \geq 1 \quad \theta_N \doteq E(p) - E(\bar{p}_N). \quad (3.129)$$

From Lemma 28 we know  $\theta_N \downarrow 0$ . Now choose the sequence  $\{k_N\}_{N=1}^\infty$  as follows: choose  $k_1, \dots, k_{N_\alpha-1}$  arbitrarily, and  $k_{N_\alpha}, k_{N_\alpha+1}$  etc. such that for all  $N \geq N_\alpha$  and for all  $\phi \in \Phi_N$  the following three relations hold true:

$$|E(\bar{p}_N) - E_{k_N}(\bar{p}_N)| < \theta_N, \quad (3.130)$$

$$|E_{k_N}(p_\phi) - E(p_\phi)| < \theta_N, \quad (3.131)$$

$$D_N \cdot (1 - 2 \cdot \beta_N)^{k_N/N} < \frac{\theta_N}{4}. \quad (3.132)$$

Eqs. (3.130) and (3.131) are possible because of Lemma 27 (note that for all  $N \geq N_\alpha$  the  $N$ -th order NAR model is in  $\Phi_N$ ). Eq.(3.132) is possible because  $\beta_N < 1$ .  $D_N$  is the constant of eq.(3.65), in Lemma 26.

Now look again at (3.128), for any  $N \geq N_\alpha$  The first term in it equals  $\theta_N$  by definition. The second term is less than  $\theta_N$  by (3.130). The fourth term is less than  $\theta_N$  by (3.131). What remains, is to show that the third term is less than  $\theta_N$ . Let us now prove this.

For all  $N \geq N_\alpha$ , the  $N$ -th order NAR model is in  $\Phi_N$ . By definition of the ML set  $\Psi_{N,n}(y)$ , for all  $N \geq N_\alpha$ , for all  $n > k_N$ , for all  $y$ , for all  $\phi \in \Psi_{N,n}(y)$  we have

$$p_\phi(y_1 \dots y_n) \geq \bar{p}_N(y_1 \dots y_n). \quad (3.133)$$

For (3.133) to really hold, we must know that the set  $\Psi_{N,n}(y)$  is not empty. This is indeed true because for all  $N$  it is easy to check that  $\Phi_N$  is a compact set (think of it as a subset of  $R^{2K^N}$  which consists of a union of closed and bounded sets). Then (3.133) implies

$$\begin{aligned} \frac{\log p_\phi(y_1 \dots y_{k_N})}{n - k_N + 1} + \frac{\sum_{k=k_N}^n \log p_\phi(y_k | y_{k-1} \dots y_1)}{n - k_N + 1} &\geq \\ \frac{\log \bar{p}_N(y_1 \dots y_{k_N})}{n - k_N + 1} + \frac{\sum_{k=k_N}^n \log \bar{p}_N(y_k | y_{k-1} \dots y_1)}{n - k_N + 1}. \end{aligned} \quad (3.134)$$

The first term of the inequality (3.134) is negative, so we can drop it and strengthen the inequality. We strengthen the inequality further by truncating the conditioning on the conditionals of the second term and *adding* an  $(1 - 2\beta_N)^{k_N/N}$  term (making use of Lemma 26). We strengthen the inequality further by truncating the conditioning on the conditionals of the fourth term and *subtracting* an  $(1 - 2\beta_N)^{k_N/N}$  term (making use of Lemma 26). Finally, we bound the marginal probability on the third term, using Lemma 22. When we are done, we have for all  $N \geq N_\alpha$ , for all  $n > k_N$ , for all  $y$ , for all  $\phi \in \Psi_{N,n}(y)$

$$\begin{aligned} \frac{\sum_{k=k_N}^n \log p_\phi(y_k | y_{k-1} \dots y_{k-k_N})}{n - k_N + 1} + D_N \cdot (1 - 2\beta_N)^{k_N/N} &\geq \\ \frac{\log \alpha_N^{2k_N}}{n - k_N + 1} + \frac{\sum_{k=k_N}^n \log \bar{p}_N(y_k | y_{k-1} \dots y_{k-k_N})}{n - k_N + 1} - D_N \cdot (1 - 2\beta_N)^{k_N/N}. \end{aligned} \quad (3.135)$$

Next, for all  $z_{-k_N}, z_{-k_N+1}, \dots, z_0$  in  $\Omega$ , count all  $k_N$  long strings  $y_k \dots y_{k+k_N}$  that equal  $z_{-k_N} z_{-k_N+1} \dots z_0$ , and obtain the  $(k_N, n)$  empirical probability function  $p_{k_N, n}^y$ . Then we have

for all  $N \geq N_\alpha$ , for all  $n > k_N$ , for all  $y$ , for all  $\phi \in \Psi_{N,n}(y)$

$$\begin{aligned} & \sum_{z_{-k_N}, \dots, z_0 \in \Omega} p_{N,n}^y(z_{-k_N} \dots z_{-1} z_0) \log p_\phi(z_0 | z_{-1} \dots z_{-k_N}) + (1 - 2\beta_N)^{k_N/N} \geq \\ & \frac{\log \alpha_N^N}{n - k_N + 1} + \sum_{z_{-k_N}, \dots, z_0 \in \Omega} p_{N,n}^y(z_{-k_N} \dots z_{-1} z_0) \log \bar{p}_N(z_0 | z_{-1} \dots z_{-k_N}) - (1 - 2\beta_N)^{k_N/N}. \end{aligned} \quad (3.136)$$

However, we know that, for all  $k_N, z_{-k_N}, \dots, z_0 \in \Omega$ ,  $p_{k_N,n}^y(z_{-k_N} \dots z_0)$  goes to  $p(z_{-k_N} \dots z_0)$  with  $\pi$  probability 1, and hence also in probability.

Also, it is clear that, for all  $N \geq 1$ , the quantity  $\log \alpha_N^{k_N} / (n - k_N + 1)$  goes to 0 as  $n \rightarrow \infty$ . In particular, we can choose  $n$  large enough so that  $\log \alpha_N^{k_N} / (n - k_N + 1)$  is greater than  $-\theta_N/2$ .

From the above facts it follows that there exists

1. a sequence of integers  $\{n_N\}_{N=1}^\infty$ ,  $n_N \uparrow \infty$ , and
2. a double sequence of sets  $\{E_{Nn}\}_{N,n=1}^\infty$ , with  $E_{Nn} \subset \Omega^\infty$  for all  $N, n \geq 1$

that satisfy the following:

1. for all  $N, n \geq 1$ ,  $\pi(E_{Nn}^c) < 2^{-N}$ ,
2. for all  $N \geq 1$   $n_N \geq k_N$ , and
3. for all  $N \geq N_\alpha$ , for all  $n > k_N$ , for all  $y \in E_{Nn}$ , for all  $\phi \in \Psi_{N,n}(y)$

$$\begin{aligned} & \sum_{z_{-k_N}, \dots, z_0 \in \Omega} p(z_{-k_N} \dots z_{-1} z_0) \log p_\phi(z_0 | z_{-1} \dots z_{-k_N}) + D_N \cdot (1 - 2\beta_N)^{k_N/N} \geq \\ & \frac{\theta_N}{2} + \sum_{z_{-k_N}, \dots, z_0 \in \Omega} p(z_{-k_N} \dots z_{-1} z_0) \log \bar{p}_N(z_0 | z_{-1} \dots z_{-k_N}) - D_N \cdot (1 - 2\beta_N)^{k_N/N}. \end{aligned} \quad (3.137)$$

Rearranging the terms in (3.137) and taking into account that  $D_N \cdot (1 - 2\beta_N)^{k_N/N}$  is less than  $\theta_N/4$ , we get

$$\sum_{z_{-k_N}, \dots, z_0 \in \Omega} p(z_{-k_N} \dots z_{-1} z_0) \log \frac{\bar{p}_N(z_0 | z_{-1} \dots z_{-k_N})}{p_\phi(z_0 | z_{-1} \dots z_{-k_N})} < \theta_N. \quad (3.138)$$

This is the desired bound on  $E_{k_N}(\bar{p}_{k_N}) - E_{k_N}(p_\phi)$  and completes the proof of the lemma.  $\bullet$

Now we can prove Theorem 1.

**Proof of Theorem 1**

Define  $n_0 \doteq 0$  and, using this and the sequence  $\{n_N\}_{N=1}^\infty$  of Lemma 2, define  $\forall n \geq 1$

$$N(n) \doteq \sup\{K \geq 0 \text{ such that } n_K \leq n\}. \quad (3.139)$$

Let us first prove that

$$N(n) \uparrow \infty. \quad (3.140)$$

It is clear that  $N(n)$  is increasing, because

$$\{K : n_K \leq n\} \subset \{K : n_K \leq n+1\} \Rightarrow N(n) \leq N(n+1). \quad (3.141)$$

To show that  $N(n)$  goes to infinity, observe that  $n_N < \infty$  for all  $N$ . This, together with  $n_N \uparrow \infty$  implies

$$\cup_{n=1}^\infty \{K : n_K \leq n\} = \{0, 1, 2, \dots\} \Rightarrow N(n) \rightarrow \infty. \quad (3.142)$$

Combining (3.141) and (3.142) we get (3.140).

Also, from the definition of  $N(n)$  it is obvious that

$$n \geq n_{N(n)}. \quad (3.143)$$

Now fix  $\epsilon > 0$ . We need to show

$$\lim_{n \rightarrow \infty} Pr \left( \sup_{\phi \in \Psi_{N(n), n}(y)} H(p_\phi; p) > \epsilon \right) = 0. \quad (3.144)$$

Fix  $\delta > 0$  and choose  $\bar{n}$  such that, for all  $n > \bar{n}$  and for the sequence  $\zeta_N$ ,  $N = 1, 2, \dots$  of Lemma 2, we have  $\zeta_{N(n)} < \epsilon$  and  $2^{-N(n)} < \delta$ . Then,  $n > \bar{n}$  implies

$$\begin{aligned} Pr \left( \sup_{\phi \in \Psi_{N(n),n}(y)} H(p_\phi; p) > \epsilon \right) &\leq \\ Pr \left( \sup_{\phi \in \Psi_{N(n),n}(y)} H(p_\phi; p) > \zeta_{N(n)} \right) &\leq \\ Pr \left( E_{N(n),n}^c \right) &< 2^{-N(n)} < \delta. \end{aligned} \tag{3.145}$$

This proves (3.144) and completes the proof of the Consistency Theorem. •

**Remark:** This completes the proof of Theorem 1, which was the main goal of this chapter. What we have shown is that any stationary ergodic stochastic process that satisfies the positivity assumption (3.11) can be approximated consistently by a sequence of SNMs. The exact value of  $\alpha$  in (3.11) does not matter; all it matters is that  $\alpha$  is strictly positive. The positivity bounds  $\alpha_N$  drop to zero as  $N$  goes to infinity; therefore, for all  $N \geq N_\alpha$  we have  $\alpha > \alpha_N$ . For all such  $N$ , the class  $\Phi_N$  is large enough to contain good models of  $Y$  (namely the  $N$ -th order NAR model) and the ML models can only be better.

**Remark:** Another important factor is the sample size  $n$ , especially as the model size  $N(n)$  varies with the size of the available sample. Unfortunately we cannot compute an explicit rate of growth for  $N(n)$  as a function of  $n$ ; we conjecture that this rate is inversely related to the rate of convergence of the  $n$ -th order entropy  $H_n(Y)$ , in other words the rate at which information is produced by observing  $Y$ . Of course this can be arbitrarily small, depending on the process  $Y$ .

**Remark:** In fact the consistency result for SNM estimation implies that consistent estimation is possible with any other type of HMM (DNM, DAM or SAM). This is so because of the Equivalence Theorem 8. We have chosen to work with SNM just because the proof of consistency was easier. The consistency of an arbitrary-type HMM estimation scheme has important (and reassuring) implications for applied, real-world Hidden Markov Modelling. Of course, for the estimation to be consistent, the increase of model size  $N(n)$  as a function of  $n$  must be slow enough. However,

as the growth rate for  $N(n)$  is not known, we cannot suggest specific values  $N, n$ .

**Remark:** As a final remark, it should be clear that the choice of the sequence  $\{\alpha_N\}_{N=1}^{\infty}$  was up to a point arbitrary; any other sequence would do as long as (a)  $\alpha_N < K^N$  and (b)  $\alpha_N \cdot K^N$  goes to zero as  $N$  goes to infinity.

## Chapter 4

# Estimation Experiments with Hidden Markov Models

In this chapter we perform a sequence of numerical experiments. The goal is to compute the Maximum Likelihood Hidden Markov Model of a sampled stochastic process. That is, we are given a class of HMMs  $\Phi$  and a finite sequence of measurements  $y_1, \dots, y_n$  from the **original** process  $Y$ . Now we seek a point  $\hat{\phi}$  that maximizes the Likelihood function

$$L(\phi; y_1 \dots y_n) \doteq Pr(Y_1^\phi = y_1, \dots, Y_n^\phi = y_n). \quad (4.1)$$

The chapter is organized as follows. In Section 4.1 we present the type of HMMs we will use. In Section 4.2 we present the Likelihood maximization algorithm. In Section 4.3 we present experiments with artificial data. In Section 4.4 we present experiments with real speech data.

### 4.1 Hidden Markov Models for One Dimensional Processes

The stochastic processes we consider in this chapter are *one dimensional*. That is, they are processes indexed by time. These are the only processes we have discussed so far; we will consider

*spatial* processes in Section 5.5.

All the estimation experiments performed in this chapter involve DNMs. That is, the model is parametrized by a transition probability matrix  $P$ . The emission matrix  $Q$  is fixed (it can be chosen judiciously, to improve the model, as will be discussed on a case-by-case basis) and is not estimated. We choose DNMs mainly for convenience; in case our results were poor we would use other types of models, e.g. SNMs. But this turned out to not be necessary, as we obtained very good DNMs for the modelling tasks we report. From the Equivalence Theorem 9 and the Consistency Theorem 20 we expect to be able to find a consistent sequence of DNMs for most reasonable stochastic processes. Indeed in several of the experiments reported here we compute a sequence of Maximum Likelihood DNMs of increasingly high dimensionality and obtain successively better results; this can be taken as a practical demonstration of the truth of our theorems.

Evaluation of the models is not a simple matter. In some cases we evaluate models by looking at their transition matrices. In other cases we compute the cross entropy (over a finite length of marginals) between the original process and the model. To compute the cross entropy we need to know the probability function of the original process and the model. When we use a HMM to generate the original process, we compute its probability function (finite length marginals) explicitly; otherwise we use the empirical probability function computed from the observations. We always compute explicitly the probability function of the model explicitly. However, it is computationally impractical to compute high order marginals; on the other hand, the speech processes we consider in Section 4.4 have a long term correlation structure which would not be captured in the low order cross entropy. Therefore, in such cases we depend on the visual appearance of the original and model processes to evaluate the goodness of the model. It will become evident in later sections that this is a good criterion.

## 4.2 The Backward - Forward Likelihood Maximization Algorithm

In this section we discuss the maximization algorithm we use to maximize the Likelihood function.

First of all we should explain that we do not use the class of models  $\Phi_N$ , which was used for the Consistency Theorem of Chapter 3. Instead we fix the state transition topology of our model (i.e. we decide in advance which transition probabilities will be positive and which ones will be zero) and then maximize Likelihood by the Backward - Forward algorithm of Baum. This algorithm picks a probability transition matrix which is locally ML in the space of all matrices with the same state transition topology. We are choosing this estimation strategy only for convenience. We just do not have an algorithm which will maximize Likelihood globally over  $\Phi_N$ . It is obvious that we are acting on faith, but we believe our results justify our approach.

For a derivation and description of the BF algorithm, see [L+83]. Here we give only a very brief summary. The algorithm consists of iterative reestimation formulas for the parameters of a HMM. We use the reestimation formulas only for the transition probabilities  $P_{ij}$ ,  $i, j = 1, \dots, M$ , but not for the emission probabilities  $Q_{ij}$ ,  $i = 1, \dots, M$ ,  $j = 1, \dots, N$  which we assume fixed. The question of choosing the emission probabilities is an interesting one which will be discussed on a case-by-case basis; proper selection of the emission matrix can simplify the estimation problem very much. At any rate, given the  $l$ -step estimate of the  $i$  to  $j$  transition probability, call it  $P_{ij}^{\phi_l}$ , the fixed emission probabilities  $Q_{ij}$  and the fixed observations  $y_1 \dots y_T$ , the  $l + 1$  step reestimation formulas for the parameters  $P_{ij}$ , call them  $P_{ij}^{\phi_{l+1}}$ , are the following:

$$p_{ij}^{\phi_{l+1}} = \frac{\sum_{t=1}^T Pr(Y_1 = y_1, \dots, Y_t = y_t, X_t = i) p_{ij}^{\phi_l} q_{j, y_{t+1}}^{\phi_l} Pr(Y_{t+2} = y_{t+2}, \dots, Y_T = y_T | X_{t+1} = j)}{\sum_{t=1}^T Pr(Y_1 = y_1, \dots, Y_t = y_t, X_t = i) Pr(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T | X_t = i)} \quad (4.2)$$

The quantities  $Pr(Y_1 = y_1, \dots, Y_t = y_t, X_t = i)$ ,  $Pr(Y_{t+2} = y_{t+2}, \dots, Y_T = y_T | X_{t+1} = j)$  etc. are computed in a recursive manner using the values  $P_{ij}^{\phi_l}$ ,  $i, j = 1, \dots, N$ ,  $Q_{ij}$ ,  $i = 1, \dots, M$ ,  $j = 1, \dots, N$ .

The recursion has the following form:

$$Pr(Y_1 = y_1, \dots, Y_{t+1} = y_{t+1}, X_{t+1} = i) = \sum_j Pr(Y_1 = y_1, \dots, Y_t = y_t, X_t = j) p_{ji}^\phi q_{iy_{t+1}}^\phi, \quad (4.3)$$

$$Pr(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T | X_t = i) = \sum_j Pr(Y_{t+2} = y_{t+2}, \dots, Y_T = y_T | X_{t+1} = j) p_{ij}^\phi q_{jy_{t+1}}^\phi. \quad (4.4)$$

All of the computational experiments reported in this and the next chapter were performed on a SUN/4 computer. Some of them were programmed in MATLAB, a high level interpreted language and some of them in FORTRAN 77. MATLAB has a number of high speed built in operations (e.g. matrix multiplication, inversion etc.) but in general is much slower than FORTRAN. A reasonable conversion rule is that 1 minute of FORTRAN time is approximately equal to 100 minutes of MATLAB time.

### 4.3 Experiments with Hidden Markov Models and Artificial Data

In this section we perform ML estimation experiments with the data being artificial data obtained from a DNM. This experiment is meant to illustrate how gradual increase of the model size produces a sequence of consistent estimates.

The original process in this experiment is a DNM with

$$P = \begin{bmatrix} .01 & .97 & .01 & .01 \\ .01 & .01 & .97 & .01 \\ .01 & .01 & .01 & .97 \\ .97 & .01 & .01 & .01 \end{bmatrix} \quad (4.5)$$

and emission matrix

$$Q = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (4.6)$$

It is easy to see that the state process is almost periodic, namely we expect with high probability state sequences of the form 12341234123... . Consequently we also expect to see with high probability observable sequences of the form 11121112... . We use  $P$  and  $Q$  above to produce a sequence of 200 samples  $y_1, \dots, y_{200}$ . These data will be used for the BF algorithm in all experiments of this section.

1. First we compute a fourth order model; the initial value of the transition matrix  $P_4$  is selected randomly. The emission matrix is

$$Q_4 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (4.7)$$

in other words the same as that of the original process. Hence we would like to obtain a transition matrix  $P_4$  that is very close to the original matrix  $P$ . Indeed, after 50 iterations of the BF algorithm, which took about 10 minutes of MATLAB time, we get the following final value for  $P_4$

$$P_4 = \begin{bmatrix} 0.0050 & 0.9918 & 0.0032 & 0.0000 \\ 0.0092 & 0.0410 & 0.0028 & 0.9470 \\ 0.9939 & 0.0040 & 0.0010 & 0.0011 \\ 0.0110 & 0.0306 & 0.9534 & 0.0050 \end{bmatrix}. \quad (4.8)$$

It is not immediately obvious, but  $P_4$  is very close to  $P$  in the output sense. To see this, perform the following permutation: map original state 2 to model state 2, original state 3 to model state

Figure 4.1:  $p, p_4$

4, original state 4 to model state 1, and original state 1 to model state 3. Then the transition probabilities of  $P$  and  $P_4$  are almost equal. The form of the hiding function  $f$  is such that, under this permutation, the observable probabilities should be almost equal, too. To further evaluate the goodness of the model, we consider all the 4-long binary strings: 1111,1112,1121,..., 2222. There are sixteen of them. Using  $P, Q$  we compute the probability function  $p$  at  $p(1111), p(1112)$  etc. Similarly, using  $P_4, Q_4$  we compute the probability function  $p_4$  at  $p_4(1111), \dots$ . We plot these two functions against the sixteen points in Fig.4.1. We can see in the figure a pretty good agreement between the original process and the second order model. The cross entropy between these two probabilities ( $H_4(p_4; p)$ ) is .0577.

**2.** Next we use a **wrong** fourth order model; the initial value of the transition matrix  $\bar{P}_4$  is

selected randomly. The emission matrix is

$$\bar{Q}_4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (4.9)$$

Obviously, the structure of the original process and the model are quite different, so we do not expect to get a good model. After 50 iterations of the BF algorithm, which took about 10 minutes of MATLAB time, we get the following final value for  $\bar{P}_4$

$$\bar{P}_4 = \begin{bmatrix} 0.1831 & 0.7816 & 0.0289 & 0.0064 \\ 0.0542 & 0.4421 & 0.0616 & 0.4420 \\ 0.3782 & 0.2894 & 0.2951 & 0.0372 \\ 0.2614 & 0.3197 & 0.2582 & 0.1606 \end{bmatrix} \quad (4.10)$$

We consider all the 4-long binary strings: 1111,1112,1121,..., 2222. There are sixteen of them. Using  $P, Q$  we compute the probability function  $p$  at  $p(1111), p(1112)$  etc. Similarly, using  $\bar{P}_4, \bar{Q}_4$  we compute the probability function  $\bar{p}_4$  at  $\bar{p}_4(1111), \dots$ . We plot these two functions against the sixteen points in Fig.4.2. The agreement between the original and the model is not very good. The cross entropy between these two probabilities ( $H_4(\bar{p}_4; p)$ ) is 2.5232.

**3.** Now we use a sixth order model; the initial value of the transition matrix  $P_6$  is selected randomly. The emission matrix is

$$Q_6 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (4.11)$$

Figure 4.2:  $p, \bar{p}_4$

In this case we have a “wrong” model, in the sense that it has different structure from the original. It is however easy to fit this model to the data by a “grouping” of the states. For instance we could have model states 1,3,5 take the role of original states 1,2,3 (as both groups emit observable 1) and model states 2,4,6 take the role of original state 4. In that case we would also expect an appropriate transition pattern. This is exactly what happens. After 50 iterations which took about 60 minutes of MATLAB time, we get the following final value for  $P_6$

$$P_6 = \begin{bmatrix} 0.0059 & 0.0000 & 0.0039 & 0.0001 & 0.9878 & 0.0023 \\ 0.9667 & 0.0008 & 0.0209 & 0.0001 & 0.0047 & 0.0067 \\ 0.0022 & 0.4376 & 0.0306 & 0.3567 & 0.0131 & 0.1596 \\ 0.9371 & 0.0003 & 0.0521 & 0.0000 & 0.0068 & 0.0037 \\ 0.0009 & 0.0012 & 0.9826 & 0.0001 & 0.0098 & 0.0054 \\ 0.9979 & 0.0000 & 0.0015 & 0.0000 & 0.0004 & 0.0002 \end{bmatrix}. \quad (4.12)$$

Figure 4.3:  $p, p_6$

We see that model state 1 goes to 5, 5 goes to 3 and 3 goes to 2, 4 or 6. This is just like original states 1 going to 2, 2 going to 3, 3 going to 4. Finally model state group 2,4,6 returns to state 1, which is just like original state 4 going to 1.

We consider all the 4-long binary strings: 1111,1112,1121,..., 2222. There are sixteen of them. Using  $P, Q$  we compute the probability function  $p$  at  $p(1111), p(1112)$  etc. Similarly, using  $P_6, Q_6$  we compute the probability function  $p_6$  at  $p_6(1111), \dots$ . We plot these two functions against the sixteen points in Fig.4.3. We can see a pretty good agreement between the original process and the second order model. The cross entropy is .0090.

This concludes the section on artificial data and HMM estimation.

Figure 4.4: A speech signal: “one”

## 4.4 Experiments with Hidden Markov Models and Speech Data

In this section we perform ML estimation experiments with the data being a real-world, pre-processed speech signal. This signal is obtained as follows: we start with an utterance of the word “one”. This utterance contains three “phonemes” i.e. elementary units of speech. Namely “one”=[ou][uh][n]. This utterance can be plotted as atmospheric pressure vs. time, see Fig.4.4. Now, each of the phonemes corresponds to a fairly easily distinguishable part of the waveform, namely [ou] occupies roughly time instants 800 to 2000, [uh] occupies 2000 to 3800 and [n] 3800 to 6000. Within these periods of time we can assume the signal to be a stationary stochastic process. So we try to model these segments of speech signal as the output of a DNM. In particular we will use two segments, a steady state [uh] segment depicted in Fig.4.5. and a steady state [n] segment depicted in Fig.4.6. However, this stochastic process is continuous valued. To place

Figure 4.5: A speech signal segment: [uh]

Figure 4.6: A speech signal segment: [n]

Figure 4.7: A quantized speech signal segment: [uh]

the experiment in the context of discrete valued stochastic processes, we quantize the signals of Figs.4.5, 4.6 to four levels. Then we get the following two waveforms appearing in Fig.4.7, Fig.4.8.

These two signals are the data we will use for the speech estimation experiments of this section.

#### 4.4.1 About Speech Recognition Systems

In this section we include a few considerations about *Speech Recognition*, which have played a role in our choice of speech models.

As already mentioned, HMMs form the core of the most successful Speech Recognition systems in use today (see [B+83, L+90] etc.). These systems essentially consist of a hierarchy of Markovian models, where the states of each level are full blown Markovian sub-models. At the lowest level we have a HMM for each phoneme. By decomposing each state at higher levels to the corresponding Markov process, we end up with a big HMM.

However these models are not fully HMM, in the following sense. The observables of the lowest

Figure 4.8: A quantized speech signal segment: [n]

level are not the raw signal, but instead some spectral information thereof. E.g. we can have LPC coefficients [L+90], or spectral coefficients [B+83] etc. When these coefficients are computed, it is possible that a lot of information (e.g. phase information) from the original signal is lost. Furthermore, the LPC etc. coefficients are not used as they are computed, but rather collapsed into classes of *Vector Quanta* (for a description of VQ see [L+90]). In this step we also lose information. The VQ process is done completely separately from the Hidden Markov Modelling; in essence we have two decoupled steps of modelling. We believe we would be much better off if we used the raw signal itself as the observable of our HMMs and performed just one-step modelling, within the HMM framework. This approach is attractive for two reasons:

1. First of all, we now have an integrated top-to-bottom HMM. Any decoding of the raw signal takes place within the framework of the HMM. In particular, we can compute the probability of the observable sequence of the raw signal ; this is not possible within the VQ framework.
2. Secondly, we believe that a sound principle of modelling is that the model should be good

enough to recreate the observed process. We can run our HMMs and produce a process that looks very much like the original one; this can not be done with the spectral / VQ models, because of the loss of phase information, vector quantization etc.

For these reasons we decide to build models of the raw signal. This is a rather drastic departure from standard Speech Recognition practice. The experiments we will report in the next few sections, can be seen as toy problems in Hidden Markov Modelling; the data we used just happen to be speech data. However, we believe we compute quite good models of speech, which have some value for the Speech Recognition problem as well. Our models will probably *synthesize* quite convincing speech signals. We believe good synthesis capabilities indicate good *recognition* capabilities as well. However, we recognize that the real value of our models for the Speech Recognition problem can only be judged when they are incorporated in a realistic speech recognizer.

#### 4.4.2 Amorphous Speech Models

In this section we experiment with the speech data of section 4.1. We use DNM's and the BF algorithm for estimation.

As we do not use any heuristic tricks in the selection of model topology, we will call the models of this section “amorphous”. In the next section we will deal with “customized” models, especially tailored to the reproduction of the original process.

We use a sequence of increasingly higher order models, hoping to approximate the original processes consistently in the limit. As the data is generated by a real world process we do not know matrices  $P$ ,  $Q$  (if such matrices exist at all). The evaluation of the model will be based on a subjective visual criterion: whether the model reproduces the characteristic visual texture of peaks and valleys that characterizes each phoneme.

1. In the first experiment we use the [uh] waveform. We start with a fourth order model with

Figure 4.9: Sample path from 4th order HMM estimate of [uh] process

$P_4$  initially chosen randomly and  $Q_4$  given by:

$$Q_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (4.13)$$

The final value is arrived at after 50 iterations (MATLAB time 20 minutes). The final matrix does not look very informative, so it is not reproduced here. At the level of near-scale interactions we do not have the actual  $P$ ,  $Q$  matrices so we cannot compute the cross entropy of the original signal. At any rate this would not be very informative. The big picture is in the long term signal correlations. The original signal has a quasiperiodic character which is easily visible in Fig.4.7. The model process does not appear to capture this visual texture (see Fig.4.9) so we proceed to higher order models.

**2.** In the second experiment we again use the [uh] waveform. We start with a 8th order model with  $P_8$  initially chosen randomly and  $Q_8$  given by:

$$Q_8 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4.14)$$

The final value is arrived at after 50 iterations (MATLAB time 30 minutes). The visual texture of the model signal is closer to the original (see Fig.4.10) than the 4th order model, but still not very good.

**3.** In the third experiment we again use the [uh] waveform; now we start with a 12th order

Figure 4.10: Sample path from 8th order HMM estimate of [uh] process

model with  $P_{12}$  initially chosen randomly and  $Q_{12}$  given by:

$$Q_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (4.15)$$

Figure 4.11: Sample path from 12th order HMM estimate of [uh] process

The final value is arrived at after 50 iterations (MATLAB time 45 minutes). The model process is plotted in Fig.4.11 and is quite close to the original.

The signal now looks a lot better. We perform a similar experiment with the [n] phoneme (Same starting values for  $P_{12}$ ,  $Q_{12}$ ) and we plot the results in Fig.4.12.

4. In the fourth experiment we use 16th order models with  $P_{16}$  initially chosen randomly and

Figure 4.12: Sample path from 12th order HMM estimate of  $[n]$  process

$Q_{16}$  given by:

$$Q_{16} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4.16)$$

Figure 4.13: Sample path from 16th order HMM estimate of [uh] process

The final value is arrived at after 50 iterations (MATLAB time 90 minutes) and is very close to the original. We repeat the same experiment for the [n] phoneme and plot the model process outputs in Figs. 4.13, 4.14.

In summary, the experiments indicate that the sequence of ML models of increasingly higher order provides increasingly better fits the original process.

### 4.4.3 Customized Speech Models

In the previous section we considered “amorphous” models, where we did not make special use of the properties of the original signal. Here we select the topology of the models with care so that we obtain “customized” models that are better for speech synthesis than the “amorphous” ones that we used in the previous section.

Careful observation of the original speech data reveals that it has a quasiperiodic character. For instance consider the [uh] waveform of Fig.4.5. It can be considered to consist of a sequence

Figure 4.14: Sample path from 16th order HMM estimate of [n] process

of distorted copies of the following prototype waveform in Fig.4.15. This follows a general idea of Grenander's (see Grenander in [Gre71] and Knoerr's work in [Kno88]).

Similarly, a prototype of [n] can be seen in Fig.4.16.

We will try to use this "deformed-prototype" phenomenon to our advantage. We proceed in the development of speech models in several steps.

1. This first experiment is not, strictly speaking, an experiment in estimation. We will just produce a (deterministic) time process which looks very similar to the original. Let us consider first the phoneme [uh]. We postulate a probability transition matrix which has 12 states, that is, as many as the time steps in the prototype of Fig.4.15. The state transition mechanism is deterministic. Namely every state goes to the next state with probability 1:  $p_{i,i+1} = 1$  for  $i = 1, \dots, 11$  and  $p_{12,1} = 1$ . Further, consider the prototype  $y_1, \dots, y_{12}$ . Then for the emission matrix we have  $q_{i,y_i} = 1$ . It should be obvious that this model produces exact copies of the prototype. For instance look at Fig.4.17.

Figure 4.15: A prototype component of [uh]

Figure 4.16: A prototype component of [n]

Figure 4.17: A deterministic model of [uh]

We build the exactly analogous model of phoneme [n], using the [n] prototype. Here the prototype has length 17, so the DNM has state process of order 17. The emission matrix  $Q$  is 17-by-4 and is given by  $q_{iy_i} = 1$ ,  $i = 1, \dots, 17$ . The typical output of this model is plotted in Fig.4.18.

These models look good with respect to the subjective visual criterion, but are weak models because any deformation of the prototype has zero probability. For instance, they would perform terribly for speech recognition, where they would assign zero probability to any speech segment that is not identical to Figs.4.17, 4.18. So we will develop more sophisticated models, using the deterministic models as starting points. For later reference, let us call the deterministic transition matrix we used in this experiment  $P_d$  and the emission matrix  $Q$ .

**2.** We now will construct phoneme models which use the same emission matrix  $Q$ , but a different transition probability matrix  $P_2$ . This new matrix has components of the form  $p_{ii} = \epsilon$ ,  $p_{i,i+1} = 1 - \epsilon$ , where  $\epsilon$  is chosen to be a small positive constant, in most cases .001. This assigns

Figure 4.18: A deterministic model of [n]

a small probability to small deformations of the prototype. That is, at any time the state process will move on to the next state with high probability, but there is a small chance that it will stay in the same state. The model  $(P_2, Q)$  is far more flexible than  $(P_1, Q)$ , because it allows for small local dilations and contractions of the time axis. A sample path from this model for [uh] is plotted in Fig.4.19, and the corresponding sample path from the model for [n] is plotted in Fig.4.20.

**3.** We are now going to use the heuristics of the previous models  $(P_1, Q)$  and  $(P_2, Q)$  to ML estimate some more refined models. In essence we are changing the structure of the space of  $P$  matrices in a way which makes it easier to search through it for the ML estimates. This is achieved by choosing the heuristic value of the  $Q$  matrix as in the previous cases **1**, **2**. Furthermore, we presume that a maximum (at least a local one) of the likelihood will exist for a  $P$  value close to the “customized” values  $P_1$  or  $P_2$ . We use this information to choose the initial value of  $P$ , call it  $P_{in}$ , to be a convex combination of  $P_1$  of the previous paragraphs and  $\bar{P}$  defined as follows:  $\bar{p}_{ij} = 1/12$  for all  $i, j$ . Now we use an initial  $P_{in}$  which is  $P_{in} = (1 - \epsilon)P_1 + \epsilon\bar{P}$ . In other words,

Figure 4.19: A quasideterministic model of [uh]

Figure 4.20: A quasideterministic model of [ɲ]

Figure 4.21: A probabilistic model of [uh]

we choose an initial matrix  $P_{in}$  that will allow (with small probability) any state transition, but will favor heavily the next state transitions. This  $P_{in}$  is “within  $\epsilon$ ” of  $P_1$  and we expect it to also be very close to a (at least locally) ML value of  $P$ . Indeed, after running the BF algorithm we get a final value for  $P$ , call it  $P_3$ , where most of the entries go down to zero- every state goes almost certainly to the next one and with a small probability to itself. Presumably the Likelihood of the estimated model is higher than that of  $P_1$ . The sample paths of the  $(P_3, Q)$  models for [uh] and [n] are plotted in Figs.4.21, 4.22 respectively.

4. In the previous example we chose carefully the initial value for  $P$  so as to force the final value of the model to be close to the desired  $P_1$ . However, it turns out that if we choose the appropriate  $Q$  (the same one as in the previous experiment), the initial value of  $P$  does not matter all that much. This is illustrated by the next experiment, where we use the BF algorithm with heuristic emission matrix  $Q$  and random initial  $P$ . The final value of the matrix, call it  $P_4$ , is very similar to the one of the previous experiment. However, convergence is slower. Sample

Figure 4.22: A probabilistic model of [n]

paths for [uh] and [n] are plotted in Fig.4.23, 4.24.

The convergence to a maximum close to  $P_1$ , independently of the initial values of  $P$  supports the conjecture that this must be very nearly the global maximum, or at least a very dominant local maximum.

The following interesting observation was made by S. Geman: note that any permutation of the model states and  $P$  matrix, accompanied by the corresponding permutation of the  $Q$  matrix leaves the output probabilities of the model unchanged. Therefore, rather than carefully customizing the  $Q$  matrix, we could just choose the number of states that map to output  $i$  ( $i = 1, 2, 3, 4$ ) to be equal to the mean occupation time of that output, as computed from the observations. This  $Q$  matrix is just a permutation of the one we have used here. Presumably then, the estimation algorithm would settle down to the  $P$  matrix values which correspond to the corresponding permutation of the  $P$  matrix. There seems to be an interesting connection to local time.

All of the estimation experiments took about 90 minutes of MATLAB time on a SUN/4.

Figure 4.23: A probabilistic model of [uh]

Figure 4.24: A probabilistic model of [n]

## Chapter 5

# Estimation Experiments with Hidden Gibbs Models

In this chapter we perform a sequence of numerical experiments. The general goal is to compute the Maximum Likelihood Hidden Gibbs Model of a given stochastic process. That is, we are given a class of HMMs  $\Phi$  and a finite sequence of measurements  $y_1, \dots, y_n$  from the **original** process  $Y$ . Now we seek a point  $\hat{\phi}$  that maximizes the Likelihood function

$$L(\phi; y_1 \dots y_n) \doteq Pr(Y_1^\phi = y_1, \dots, Y_n^\phi = y_n). \quad (5.1)$$

We consider one dimensional data (exactly the same as the ones used in Chapter 4) as well as two dimensional data (images). The chapter is organized as follows. In Section 5.1 we present the type of HGMs we will use for modelling of one dimensional data. In Section 5.2 we present the Likelihood maximization algorithm. In Section 5.3 we present experiments with real speech data (one dimensional). In Section 5.4 we present the type of HGMs we will use for modelling of one dimensional data. In Section 5.5 we present experiments with real image data (two dimensional).

## 5.1 Hidden Gibbs Models for One Dimensional Processes

In this and the next few sections, we use Hidden Gibbs Models to model one dimensional processes, that is, processes indexed by time. These are the only processes we have discussed so far, but certainly not the only existing ones; a type of *spatial* processes will be discussed in Sections 5.4 and 5.5.

All the estimation experiments performed in this chapter involve HGM.

A HGM is described in terms of the local conditional probabilities and the emission probabilities. The emission probabilities  $\{Q_{i,j}\}_{i,j=1}^{M,N}$  ( $M$  being the size of the state alphabet and  $N$  the size of the observable alphabet) will be chosen in advance and fixed for all the estimation experiments. Hence the HGM is completely described by the local conditionals, which in turn are in terms of the *energy coefficients*  $A = \{A_{i,j}\}_{i,j=1,\dots,M}$ :

$$Pr(X_t = i | X_{t+1} = j, X_{t-1} = k) = \frac{\exp(-A_{ij} - A_{ik})}{Z(A)}. \quad (5.2)$$

So the parameters we want to estimate are the energy coefficients  $A = [A_{ij}]_{i,j=1}^M$ . These determine the  $P$  matrix uniquely. Conversely,  $P$  determines  $A$  up to an additive constant.

Of course, every Hidden Gibbs Model is a Hidden Markov Model with positive state process (see Theorem 11) and there is a one-to-one relationship between the transition probability matrix  $P_{ij}$ ,  $i, j = 1, 2, \dots, N$  and the energy coefficients  $A_{i,j}$ ,  $i, j = 1, 2, \dots, N$ . In this sense there is nothing new in the Gibbs formulation. The reason we adopt the point of view of Hidden Gibbs Models because we want to extend our results to *multidimensional stochastic processes* (by this we mean processes of the form  $\{X_t\}_{t \in T}$ , where  $T$  is for instance the set of *pairs* of integers). This will be done in Sections 5.4 and 5.5.

From the Equivalence Theorem 9 and the Consistency Theorem 20 we expect to be able to find a consistent sequence of HGMs for most reasonable stochastic processes. Indeed in several of the experiments reported here we compute a sequence of Maximum Likelihood HGMs of increasingly high dimensionality and obtain successively better results; this can be taken as a practical

demonstration of the validity of our theorems.

## 5.2 Likelihood Maximization Algorithms

For Likelihood maximization of Hidden Gibbs Models we use the **Stochastic Relaxation** algorithm which works as follows. Essentially it is a steepest ascent algorithm. It should be emphasized that here the unknown parameters are the energy coefficients:

$$Pr(X_t = i | X_{t+1} = j, X_{t-1} = k) = \frac{\exp(-A_{ji} - A_{ik})}{Z(A)}. \quad (5.3)$$

$Z(A)$  is a normalizing constant which also depends on  $A$ . Now, the local conditionals in (5.3) are known to completely determine the probability function  $Pr(X_{t+1} \dots X_{t+n}; A)$  (here we indicate explicitly the  $A$  dependence). From the probability function and the emission probabilities we can compute the observable probability function  $Pr(Y_{t+1} = y_1, \dots, Y_{t+n} = y_n; A)$ , which is exactly the Likelihood function  $L(A; y_1 \dots y_n)$ . Therefore, to implement a steepest ascent method we need to compute  $\frac{\partial L(A; y_1 \dots y_n)}{\partial A_{ij}}$ . After considerable algebraic elaboration we arrive to the following expression:

$$\frac{\partial L(A; y_1 \dots y_n)}{\partial A_{ij}} = E_A \left( \sum_{t=1}^T \mathbf{1}_{ij}(X_t X_{t+1}) \right) - E_A \left( \sum_{t=1}^T \mathbf{1}_{ij}(X_t X_{t+1}) | Y_1 \dots Y_T = y_1 \dots y_T \right). \quad (5.4)$$

(The function  $\mathbf{1}_{ij}(xy)$  equals 1 iff  $x = i$ ,  $y = j$  and 0 otherwise.) The expectations in (5.4) can be explicitly calculated in terms of  $A_{ij}$ , but this is too cumbersome. Therefore, rather than an explicit calculation, we use the *relaxation* technique (this has the additional advantage of being generalizable to the 2-dimensional case). At the  $l$ -th step we assume  $A^l$  known and simulate a Gibbs process with energy coefficients  $A_{ij}^l$ ,  $i, j = 1, \dots, M$ . Then we count the number of occurrences (over an appropriately large sample) of  $ij$  pairs and use them in place of the expectations in (5.4). The simulation is done for two cases: first with the system running free, and secondly constraining the observations to be the same as the actual sample  $y_1 \dots y_T$ . It is well

known how to perform such simulations, see for instance [GG84]. Having computed the empirical averages of  $ij$  occurrences, both unconstrained and constrained under the actual sample, we substitute these for the ensemble expectations in (5.4) and use this value as  $\left[ \frac{\partial L(A; y_1 \dots y_n)}{\partial A_{ij}} \right]_{A_{ij}=A_{ij}^l}$ . Now we update the  $A$  parameters to the value  $A^{l+1}$  using

$$A_{ij}^{l+1} = A_{ij}^l + \delta \cdot \left[ \frac{\partial L(A; y_1 \dots y_n)}{\partial A_{ij}} \right]_{A_{ij}=A_{ij}^l}. \quad (5.5)$$

The step parameter  $\delta$  has to be chosen small enough so that we can be reasonably confident that we are actually taking an ascent step.

The computer programs that implement the stochastic relaxation algorithm have been written in the computer language FORTRAN, on a SUN/4.

### 5.3 Experiments with Hidden Gibbs Models and Speech

#### Data

The basic estimation problems are the same in this section as in Section 4.4. (the artificial data experiments were repeated with identical results as in Section 4.3, but we do not present the results here, as there is nothing new to report). We use the same speech waveforms as in Section 4.4; now we use HGMs rather than HMMs and maximize the Likelihood by Stochastic Relaxation rather than by the Backward - Forward algorithm. We present four models here, two for an [uh] process and two for an [n] process. We use the customized emission matrices of section 4.4.2 and present results for: (a) customized initial values of the state transition matrix  $P$  and (b) randomly selected initial value of  $P$ . From  $P$  we can easily compute initial values for the energy coefficients  $A$ .

In the following Figures 5.1 , 5.2 we present the results of customized initial values for [uh] and [n] values, respectively.

In the following Figures 5.3 , 5.4 we present the results of random initial values for [uh] and

Figure 5.1: A probabilistic model of [uh]

Figure 5.2: A probabilistic model of [n]

Figure 5.3: A probabilistic model of [uh]

[n] values, respectively.

The models we obtain with the HGM / Stochastic Relaxation combination are just as good as the ones we get with HMM / BF, but the computational effort is much larger. We programmed the Stochastic relaxation in FORTRAN, which is on the average 10-15 times faster than MATLAB and we had estimation times in the order of 60 - 90 minutes. Given that this is FORTRAN time, it equals about 900 - 1200 minutes of MATLAB times; in comparison, the MATLAB estimation time for speech problems in Section 4.4.2 was about 90 minutes.

## 5.4 Hidden Gibbs Models for Two Dimensional Processes

In this and the next section we extend our modelling and estimation methods to two-dimensional stochastic process. By this we mean processes of the form  $\{X_{s,t}\}_{s,t=0,\pm 1,\pm 2,\dots}$ . In other words, these processes are indexed by *pairs* of integers. A typical example of a real world two dimensional process is an *image* on a computer screen, which consists of a collection of *pixels* arranged on

Figure 5.4: A probabilistic model of [n]

a rectangular array (typically a 128-by-128 square). Such images have been modelled as two dimensional Gibbs processes, see for instance [GG84]. The one dimensional Gibbs process of Section 1.4 can be extended to two dimensions in a straight forward manner.

Such two dimensional processes can also be considered as Hidden Gibbs Models. Such a model is parametrized by the *energy coefficients*  $\{A_{i,j}^h, A_{k,l}^v\}_{i,j,k,l=1,\dots,M}$  (with superscripts  $h$  and  $v$  indicating horizontal and vertical systems of energy coefficients, respectively) and an emission matrix  $\{Q_{i,j}\}_{i=1,\dots,M,j=1,\dots,N}$ , where  $M$  and  $N$  are the size of the state and observable alphabet respectively. The local conditional probabilities are expressed in terms of the local energy coefficients:

$$Pr(X_{s,t} = i | X_{s-1,t} = j_1, X_{s+1,t} = j_2, X_{s,t-1} = j_3, X_{s,t+1} = j_4) = \frac{\exp(-A_{j_1 i}^h - A_{i j_2}^h - A_{j_3 i}^v - A_{i j_4}^v)}{Z(A^h, A^v)}. \quad (5.6)$$

In the experiments of the following section, we take  $Q$  to be fixed and the only parameters

we estimate are the energy coefficients  $\{A_{ij}^h\}_{i,j=1}^M$ ,  $\{A_{ij}^v\}_{i,j=1}^M$ . For Likelihood maximization we use the Stochastic Relaxation algorithm. All of these are straightforward extensions of the one dimensional case of Sections 5.2 and 5.3. However, note that, because of the two dimensional character of the process, we cannot use the Backward - Forward algorithm, which depends on a *linearly ordered* factorization of marginal probabilities. Such a factorization is not possible in the two dimensional case. This is one point where the Stochastic Relaxation algorithm is superior to the Backward - Forward algorithm.

It should also be pointed out that in the two dimensional case the local conditionals do *not* determine uniquely the marginals (unless additional conditions are imposed on them); this is the well known phenomenon of *phase transitions* and constitutes an important difference from the one dimensional case.

## 5.5 Experiments with Hidden Gibbs Models and Images

In this section we estimate energy coefficients for two dimensional HGMs of images. We use simple black - white images. Some of these images are artificially synthesized and some come from texture photographs in the Brodatz collection [Bro66]. The photographs were scanned in at 256 gray levels resolution and then thresholded to a two level resolution (black and white). We represent these pictures using stars (for black) and dots (for white). Consequently, the observable space is  $\{1, 2\}$ . In all the experiments the state space was of fixed size, namely the state process took values in  $\{1, 2, \dots, 10\}$ . We present three estimation experiments. The three pictures we used (an artificial stripes pattern, a thresholded pictures of paper and one of straw) appear in Figs.5.5, 5.7 and 5.9 and the three reconstructions (produced by performing stochastic relaxation on the ML estimated model until it reached equilibrium) appear in Figs.5.6, 5.8 and 5.10. The estimations run took between 15 and 20 hours of FORTRAN time on a SUN/4.

Looking at these pictures we realize that the computational load is much more than for the one dimensional case and the results not as good. We present these results as preliminary work;

Figure 5.5: A picture of stripes

Figure 5.6: A model of stripes

Figure 5.7: A picture of straw

Figure 5.8: A model of straw

Figure 5.9: A picture of paper

Figure 5.10: A model of paper

we believe the HGMs idea is very promising, but further work is needed before we can evaluate them in a more definitive way. On the other hand, the models in their current form may already be useful for image recognition (e.g. texture segmentation). Recognition is in general a task easier than synthesis.

## Chapter 6

# Conclusion

Let us now summarize the results of this dissertation and indicate some future research directions.

The dissertation consists of two fairly distinct parts. The first part, consisting of Chapters 1-3, is of a theoretical character. The second part, consisting of Chapters 4 and 5 is applications - oriented.

In the theoretical part we develop a mathematical framework to discuss the issues of approximation and estimation.

In particular, in Chapter 1 we define convergence of stochastic processes, enumerate all the different types of HMMs used in the literature (and we show they are all equivalent in their representation power), introduce Hidden Gibbs Models and review the theoretical HMM literature.

It is important to note that approximation results depend on the selection of a particular distance between stochastic processes. If a different definition of distance were used, our results would probably need to be modified, in particular they might hold for smaller or larger classes of stochastic processes. The investigation of different convergence definitions is one of the open research problems we will discuss shortly.

In Chapter 2, starting with our definitions of process approximation, we prove that every stationary ergodic stochastic process can be approximated by a sequence of HMM. The result is mostly of theoretical significance, because the approximating sequence does not consist of the

most “economical models”.

In Chapter 3 we turn to the selection of economical models, based on Maximum Likelihood Estimation, which is indeed the most often used method of Hidden Markov Modelling in applications. Our contribution is to show that consistent estimation is indeed possible, *provided we adapt the size of our models to the amount of observations available*. This is a formalization of a well known principle of statistics: *we should not overfit our data*. The same principle has been rediscovered in many disciplines that develop models of natural processes (Neural Nets Theory comes to mind).

Our approach in Chapters 1 - 3 is quite theoretical. Many practical issues are not resolved, for instance how to compute the globally Maximum Likelihood values of the model parameters, what is the actual amount of data required for a fixed order model and what is the best state transition topology for a particular problem. These are also open research topics.

The character of the second part of the dissertation, consisting of Chapters 4 and 5 is quite different. Here we deal with specific processes (either real world speech signals or real world images) and we develop specific models, restricting the space of all possible models based on our knowledge of the peculiarities of each problem. The choice of practical optimization algorithms, also depends on the problem at hand. We are able to develop some quite good models, but it must be noted that we take certain shortcuts and violate some of the conditions that are known to ensure consistency.

We feel justified in doing so, because the emphasis in this part of the dissertation is on model building. We believe both our speech model and image model (especially the former) are sufficiently novel and promising to warrant further investigation. In particular, let us note that time domain modelling of speech signals at the phonemic level has not been until now successfully integrated within the hierarchy of Hidden Markov Models which are used at all the other levels of the problem. We believe our approach will permit the use of a top - to - bottom Hidden Markov Model.

It is clear that there is a dichotomy between our theoretical and practical analysis. We do not

view this as a bad thing; on the contrary we believe the two approaches are complementary. In particular, the theoretical analysis, even if not directly applicable, offers valuable insight in the nature of the modelling problem, that can lead to successful practical design rules. For instance, the practical rule of HMM estimation, that indicates transition probabilities must be kept above a minimum strictly positive value, takes a new significance in view of our estimation results. The fact the consistent estimation can be proven only for a sufficiently slow rate of model size growth alerts us to the need of practical rules for correlating the number of data points available to the side of the model.

Further, we feel that specialized rules, that apply to the domain of specialized problems, are necessary for practically useful results. However, the unifying and guiding role of theory must not be overlooked.

To illustrate this last point, take a look at the history of HMMs. In the thirty five years that have elapsed since the “invention” of HMMs, they have been reinvented many times and in various forms; their theory has also been reinvented in various degrees of sophistication. We believe this is so because of the natural generalization of the Markov property to the Hidden Markov property, which allows for infinite (albeit fading) memory of the past.

Let us conclude with a (partial) list of future research directions.

1. We would like to experiment with different definitions of convergence of stochastic processes and see how far we can extend our results. For instance, convergence in total variation is probably too strong to obtain approximation (or consistency) results for *arbitrary* stochastic processes. On the other hand, using Ornstein’s  $d$ -bar distance, we believe we can extend his approximation results and also obtain consistency results.
2. We want to extend our representation, estimation and consistency results to the case of multidimensional processes. Computational efficiency and the issue of phase transitions are the main issues to be dealt with. S. Geman and H. Kuisch have already obtained some results in this direction.

3. We also want to extend our representation, estimation and consistency results to the case of continuous valued processes. This approach can become practical only if we discover efficient modelling algorithms for the continuous valued case; a EM-type of iterative procedure may prove useful.
4. A special case of time-space stochastic process that we want to investigate, is the state evolution of locally interacting particles, probabilistic cellular automata on an infinite lattice, stochastic neural networks etc. To be more concrete, here is an example of what we have in mind. Consider a graph (finite or infinite) where every node is a processor (particle, neuron etc.). The processor can be in one of a finite number of states (for concreteness, say it can either be on or off). At every time step each processor reads the state of his neighbors; then all processors update their state according to a conditional probability function, with the conditioning being on the state of their neighbors. We observe the state of each processor by a probabilistic mechanism. Such a system is obviously a HMM (possibly with an infinite state space); on the other hand it has obvious similarities with the HGMs of Chapter 1, in that it has a local neighbor structure. When we have an infinite state space, we cannot use the results of this thesis. We would like to obtain, for this type of system, representation and consistent estimation results and a practical ML algorithm (maybe an extension of the BF algorithm).
5. An application of the system discussed in the previous paragraph, would be the heuristic / intelligent design of HMMs and / or stochastic neural networks. Simple examples of such design are the customized emission matrices of our speech models. It would be interesting to find ways of customizing the topology of state transitions. A case of particular interest would be the incorporation of implicit reasoning *rules* in the transition mechanism. A similar technique has been already used to a small extent in the customized design of neural networks: certain neurons are identified with inhibiting or reinforcing behavior and inhibiting / reinforcing values of connection weights have been used in the network design. This

amounts to rewiring AND, OR etc. gates in the network. For the type of stochastic network we have in mind, the local conditional probabilities associated with each neuron can be seen as implementing a stochastic logic gate; and a network of such gates can be seen as implementing a *distributed probabilistic reasoning* mechanism.

6. Moving on to a more applications – oriented set of problems, we want to incorporate our Speech Model in a full scale Speech Recognizer. This recognizer would be an integrated hierarchical Hidden Markov Model, with the states of every level in the hierarchy being Markov processes. The lowest level process would be our speech phonemic model. We believe that we can obtain superior recognition results, because we do not have the problem of information loss inherent in Vector Quantization or Spectral methods.
7. Finally, we want to refine our Image Model, especially with the use of more general energy functions (e.g. polynomial, trigonometric etc. energy functions) and continuous valued states.

This is just an indication of possible directions to which this work can be extended. Some of the problems we indicated are more applied and some more theoretically directed.

# Bibliography

- [Aok87] M. Aoki. *State Space Modelling of Time Series*. Springer, Berlin, 1987.
- [Arb66] M. Arbib. Realization of stochastic systems. *Annals of Mathematical Statistics*, 37:927–933, 1966.
- [Ash65] R. Ash. *Information Theory*. Interscience, New York, 1965.
- [B<sup>+</sup>70] L.E. Baum et al. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. *Ann. of Math. Stat.*, 41(1):164–171, 1970.
- [B<sup>+</sup>83] L.R. Bahl et al. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. PAMI*, 5(2):179–190, March 1983.
- [BE67] L.E. Baum and J.A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov Processes. *Ann. of Math. Stat.*, 28:36–363, 1967.
- [Bil65] P. Billingsley. *Ergodic Theory and Information*. Wiley, New York, 1965.
- [Bil71] P. Billingsley. *Weak Convergence of Measures*. Wiley, New York, 1971.
- [Bil79] P. Billingsley. *Probability and Measure*. Wiley, New York, 1979.
- [BK57] D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite Markov Chains. *Ann of Math. Stat.*, 28:1011–1015, 1957.
- [Boo67] T. Booth. *Sequential Machines and Automata Theory*. Wiley, New York, 1967.

- [Bos75] K. Bosch. Notwendige und hinreichende bedingungen da fuer dass eine funktion einer homogener Markoffschen Ketten Markoffsch ist. *Z. Wahr. verw. Gebiete*, 31:199–202, 1975.
- [BP66] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov Chains. *Ann. of Math. Stat.*, 37:1554–1563, 1966.
- [Bro70] Ph. Brodatz. *Textures*. Dover, New York, 1966.
- [BR58] C.J. Burke and M. Rosenblatt. A Markovian function of a Markov chain. *Ann. of Math. St.*, 29:1112–1122, 1958.
- [Bri89] J. S. Bridle. Alpha-nets: A recurrent neural network architecture with a Hidden Markov Model interpretation. Technical Report SP4, Research Note 104, Royal Signals and Radar Establishment, October 1989.
- [Bro87] P. Brown. *The acoustic Modelling problem in automatic speech recognition*. PhD thesis, Carnegie Mellon Un., Pittsburgh, Pennsylvania, 1987.
- [BS68] L.E. Baum and G.R. Sell. Growth transformations for functions on manifolds. *Pac. J. of Math.*, 27(2):211–227, 1968.
- [BT77] D.R. Barr and M.U. Thomas. An eigenvector condition for Markov Chain lumpability. *Op. Res.*, 25(6):1028–1031, 1977.
- [BW88] H. Bourlard and C.J. Wellekens. Links between Markov models and multilayer perceptrons. *Phillips Research Lab*, 1988.
- [BW89] H. Bourlard and C.J. Wellekens. Speech dynamics and recurrent neural nets. In *Proc. of the ICASSP*. IEEE, 1989.
- [Dha63a] S.W. Dharmadikari. Functions of finite Markov Chains. *Ann. of Math. Stat.*, 34:1022–1032, 1963.

- [Dha63b] S.W. Dharmadikari. Sufficient conditions for a stationary process to be a function of a Markov Chain. *Ann. Math. Stat.*, 34:1033–1041, 1963.
- [Dha68] S.W. Dharmadikari. Splitting a single state of a stationary process. *Ann. of Math. Stat.*, 38(3):1069–1077, 1968.
- [DF37] W. Doeblin and R. Fortet. Sur des chaines a liaisons completes. *Bull. Soc. Math. France*, LXV:132-148, 1937.
- [Doo53] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.
- [D<sup>+</sup>77] A.P. Dempster et al. Maximum Likelihood estimation via the EM algorithm. *J. of the Stat. Roy. Soc. B*, 39(1):1–38, 1977.
- [Eri70] R.V. Ericson. Functions of Markov Chains. *Ann. of Math. Stat.*, 41(3):843–850, 1970.
- [FR85] D. R. Fredkin and J.A. Rice. On aggregated Markov Processes. *J. Appl. Prob.*, 23:208–214, 1985.
- [FR87] D.R. Fredkin and J.A. Rice. Correlation functions of a function of a finite state Markov Process with applications to kinetics. *Math. Biosciences*, 87:161–172, 1987.
- [Fri67] B.D. Fritchman. A binary channel characterization using partitioned Markov Chains. *IEEE Trans. IT*, 12(2):221–227, 1967.
- [GG84] S.Geman and D. Geman. Stochastic relaxation Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. PAMI*, 9:721–741, 1984.
- [GH79] S. Geman and Chii-Ruey Hwang. Nonparametric Maximum Likelihood estimation using the Method of Sieves. Brown University, Providence, Rhode Island, May 1979.
- [Gil59] E.J. Gilbert. On the identifiability problem for functions of finite Markov Chains. *Ann. Math. Stat.*, 30:688–697, 1959.

- [GM91] S. Geman and K. Manbeck. Machine Recognition of Human Coronary Arteries by Deformable Templates. Division of Applied Mathematics, Brown University. *In Preparation*.
- [Gre76] U. Grenander. *Lectures in Pattern Theory* Springer, New York, 1976.
- [Hac63] J. Hachigian. Collapsed Markov Chains and the Chapman-Kolmogorov equation. *Ann. of Math. Stat.*, 34:233-237, 1963.
- [Gre81] U. Grenander. *Abstract Inference*. Wiley, New York, 1981.
- [Hac63] J. Hachigian. Collapsed Markov Chains and the Chapman-Kolmogorov equation. *Ann. of Math. Stat.*, 34:233-237, 1963.
- [Hel65] A. Heller. On stochastic processes derived from Markov Chains. *Ann. Math. Stat.*, 36:1286-1291, 1965.
- [HL69a] C. G. Hilborn and D.G.Lainiotis. Optimal Estimation in the Presence of Unknown Parameters. *IEEE Trans. PAMI*, 1:38-43, 1969.
- [HL69b] C. G. Hilborn and D.G.Lainiotis. Unsupervised Learning Minimum Risk Pattern Classification for Dependent Hypotheses and Dependent Measurements. *IEEE Trans. PAMI*, 2:109-115, 1969.
- [JR85] B.H. Juang and L.R. Rabiner. Mixture autoregressive HMM for speech signals. *IEEE Trans. ASSP*, 33(6):1404-1413, December 1985.
- [Keh90a] A. Kehagias. The local Backward-Forward algorithm: Hidden Markov Models optimization for Connectionist Networks. Division of Applied Mathematics, Brown Un., Providence, Rhode Island, October 1990.
- [Keh90b] A. Kehagias. Reproducing infinite Boolean sequences: an application of Hidden Markov Models to connectionist learning. In *Proc. of IJCNN*, 1:291-294. Washington, DC. IEEE, 1990.

- [Keh90c] A. Kehagias. Optimal control for training: the missing link between Hidden Markov Models and Connectionist Networks. *J. of Math. and Comp. Mod.*, Vol.14, pp.284-289, 1990.
- [Keh91a] A. Kehagias. Stochastic Recurrent Networks: Prediction and Classification of Time Series. Division of Applied Mathematics, Brown Un., Providence, Rhode Island. February 1991.
- [Keh91b] A. Kehagias. Stochastic Recurrent Networks: Representation Properties. Division of Applied Mathematics, Brown Un., Providence, Rhode Island, *In preparation*.
- [Keh91c] A. Kehagias. Stochastic Recurrent Networks: Unification and Extension. Division of Applied Mathematics, Brown Un., Providence, Rhode Island, *In preparation*.
- [KH89a] S.Y. Kung and J.N. Hwang. A unified modelling of connectionist neural networks. *IEEE Trans. ASSP*, 32(1):1-10, 1989.
- [KH89b] S.Y. Kung and J.N. Hwang. A unified systolic architecture for artificial neural networks. In *Proc. of ICASSP'89*, Glasgow, England. IEEE. May 1989.
- [KH89c] S.Y. Kung and J.N. Hwang. A unifying viewpoint of multi-layer perceptrons and Hidden Markov Models. In *Int. Symposium on Circuits and Systems*, Portland, Oregon, 1989. IEEE.
- [Kie89] P. Kienker. Equivalence of aggregated Markov models of ion-channel gating. *Proc. of R.Soc. London, Ser. B*, 269-309, 1989.
- [KS60] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Van Nostrand, New York, 1960.
- [Kno88] A. P. Knoerr. *Global models of natural boundaries : theory and applications*. PhD Thesis, Brown University. Providence, R.I. 1988.
- [Ku59] S. Kullback. *Information theory and statistics*. Wiley. New York. 1959.

- [L<sup>+</sup>83] S.E. Levinson et al. An introduction to the application of the theory of probabilistic functions of a Markov Chain. *The Bell Sys. Tech. J.*, 62(4), April 1983.
- [L<sup>+</sup>90] K.F. Lee et al. An overview of the SPHINX speech recognition system. *IEEE Trans. ASSP*, 38(1):35–45, January 1990.
- [Ley67] F.W. Leysieffer. Functions of finite Markov Chains. *Ann. of Math. Stat.*, 38:206–211, 1967.
- [LS83] L. Ljung and T. Soderström. *Theory and Practice of Recursive Identification*. MIT, Cambridge, 1983.
- [May82] P.S. Maybeck. *Stochastic Models Estimation and Control*. Mathematics in Science and Engineering. Academic, New York, 1982.
- [MK90] W.D. Mao and S.Y. Kung. Shape recognition by ring HMM's. In *Proc. of IJCNN*, II:409-412. Washington, DC. IEEE. 1990.
- [OW90] D. Ornstein and B. Weiss. How Sampling reveals a process. *Ann. of Prob.*, 18(3):905–930, 1990.
- [Paz71] A. Paz. *An Introduction to Probabilistic Automata*. Academic, New York, 1971.
- [Pet69] T. Petrie. Probabilistic functions of finite state Markov Chains. *Ann. of Math. Stat.*, 40(1):97–115, 1969.
- [R<sup>+</sup>83] L.R. Rabiner et al. On the application of vector quantization and HMM to speaker independent isolated word recognition. *The Bell Sys. Tech. J.*, 62(4):1075–1105, April 1983.
- [R<sup>+</sup>85] L.R. Rabiner et al. Recognition of isolated digits using Hidden Markov Models with continuous mixture densities. *Bell System Tech. J.*, 64(6), July - August 1985.
- [Rab88] L.R. Rabiner. A tutorial on HMM and selected applications in speech recognition. *IEEE Proc.*, 257-285, 1988.

- [Ros59] M. Rosenblatt. Functions of a Markov process that are Markovian. *J. of Mathematics and Mechanics*, 8(4):585–596, 1959.
- [Ros71] M. Rosenblatt. *Markov Processes - Structure and Asymptotic Behavior*. Springer, Berlin, 1971.
- [Roy68] H.L. Royden. *Real Analysis*. MacMillan, New York, 1968.
- [RP81] L.C.G. Rogers and J.W. Pittman. Markov functions. *Ann. of Prob.*, 9(4):573–582, 1981.
- [SKK76] J.L. Snell, J.G. Kemeny and A.W. Knapp. *Denumerable Markov Chains*. Springer, New York, 2nd edition, 1976.
- [W<sup>+</sup>74] B.A. Wandell et al. Equivalence classes of functions of finite Markov Chains. *J. of Math. Psych.*, 11:391–403, 1974.
- [Wu83] C.F.J. Wu. On the convergence properties of the EM algorithm. *The Annals of Probability*, 11(1):95–103, 1983.