THESIS REPORT

Ph.D.

Consistent Estimation of the Order for Markov and Hidden Markov Chains

by L. Finesso
Advisor: J. Baras

Ph.D. 91-1

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Consistent Estimation of the Order for Markov and Hidden Markov Chains

by

Lorenzo Finesso

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1990

Advisory Committee:
Professor John Baras, Chairman/Advisor
Professor Carlos Berenstein
Associate Professor Armand Makowski
Associate Professor Prakash Narayan
Assistant Professor Adrianos Papamarcou

Abstract

Title of Dissertation: Consistent Estimation of the Order for Markov and Hidden Markov Chains

Lorenzo Finesso, Doctor of Philosphy, 1990

Dissertation Directed by: John S. Baras
Professor
Electrical Engineering Department

The structural parameters of many statistical models can be estimated maximizing a penalized version of the likelihood function. We use this idea to construct strongly consistent estimators of the order for Markov Chains and Hidden Markov Chain models. The specification of the penalty term requires precise information on the rate of growth of the maximized likelihood ratio. For Markov chain models we determine the rate using the Law of the Iterated Logarithm. For Hidden Markov chain models we find an upper bound to the rate using results from Information Theory. We give sufficient conditions on the penalty term to avoid overestimation and underestimation of the order. Examples of penalty terms that generate strongly consistent estimators are also given.

Acknowledgments

The idea of studying the problem of order determination for Hidden Markov Chains was suggested to me by my advisor, Dr. John Baras. He left me ample liberty of choice on the details of the methodology to follow, the probabilistic tools to employ and, in general, the literature to consult.

I decided on the methodology after perusing the booklet by Azencott and Dachuna-Castelle.

The probabilistic tools, I learned from the local guru, Dr. Eric Slud. His lectures have been the most stimulating intellectual experience I ever had.

My most sincere thanks go to Dr. Prakash Narayan for the numerous discussions we had on the problem of order determination for finite state systems. He has been an invaluable source of essential papers, good advice and encouragement. I could not have finished this thing without his help.

The completion of this project has only been possible thanks to the help that I received from many individuals and institutions. I gratefully acknowledge the financial support provided by Unisys Corp., the AFOSR and NSF.

A note of thanks to Mrs. Vigil. She has been the victim of my cuneiform and she has managed it without ever losing her good humor.

Contents

Chapter 0

Introduction

Let $\{Y_t, t\epsilon Z\}$ be a stationary finitely valued stochastic process that admits a representation of the form $Y_t = f(X_t)$ where $\{X_t, t\epsilon Z\}$ is a finite Markov chain and f is a many-to-one function. We call such a process a Hidden Markov Chain (HMC).

Under well known conditions on f a HMC inherits the Markov property of X_t and becomes a finite Markov chain itself, but this case is non-generic. In general a HMC need not be a Markov chain of any finite order and will therefore exhibit long-range dependencies of some kind. This fact means that the class of HMC's is a very rich one and it comes to no surprise that it is extensively present in many applications.

We can find HMC's appear under various disguises in such diverse fields as: engineering (stochastic automata, speech recognition), biosciences (in ethology to model the mating behavior of some species, in medicine to study neurotransmission), economics (stock market predictions), and many others.

On the theoretical side the same fact (lack of the Markov property) makes the class of HMC's difficult to work with. The general methods developed for the study of stationary processes apply but being non-specific they will not give the best results. Theoretical work on the specific class of HMC's has proceeded along two main lines.

The early contributions, inspired by the work of Blackwell and Koopman (1957) [5], concentrated on the probabilistic aspects. The basic question was the characterization of HMC's. More specifically the problem analyzed was: *among all finitely valued stationary processes Y_t characterize those that admit a HMC repre-*

sentation. This problem was solved by Heller [12] in 1965. To some extent Heller's result is not quite satisfactory since his methods are non-constructive. Even if Y_t is known to be representable as a HMC, no algorithm has been devised to produce a Markov chain X_t and a function f such that $Y_t = f(X_t)$ or at least $Y_t \sim f(X_t)$ (i.e. they have the same laws). In recent years the problem has attracted the attention of workers in the area of Stochastic Realization Theory, and while some of the issues have been clarified a constructive algorithm is still missing.

The first contributions dealing with statistical aspects were made in the late sixties. Baum and Petrie [4] studied maximum likelihood estimation of the parameters of a HMC proving consistency and asymptotic normality of the MLE. They also provided an algorithm for the numerical computation of the MLE (of course there is little hope for an explicit solution in a non-Markovian setting) basically inventing the EM algorithm that became popular only later thanks to the work of Dempster, Laird and Rubin [8]. After the mid seventies HMC's made only sporadic appearances in the statistical literature. In 1975 HMC's were proposed by Baker [2] as models for automatic speech recognition (ASR) and ever since they have been adopted as one of the models of choice in this field. Computational aspects became very important and much work was done on the implementation of Baum's algorithm. A good survey of this area of research is [16] which also includes an extensive bibliography.

Although much work has been dedicated to parameters estimation for HMC's only very recently the order estimation problem received some attention. The order of an HMC Y_t is the minimum integer q for which there exists a q-valued Markov chain X_t such that $Y_t = f(X_t)$ for some f. The knowledge of the order of an observed HMC Y_t allows the construction of the *most economical* representations $f(X_t)$ in the sense that the number of parameters (the transition probabilities of X_t) is minimized. The order cannot be estimated using the classical maximum likelihood because increasing the parameter q automatically increases the likelihood. This is the typical behavior of the likelihood function when the parameter is *structural* i.e. the parameter (usually integer valued) indexes the complexity of the model. As another example of structural parameter we mention the order of a

Markov chain i.e. the smallest integer m such that:

$$P(X_t \mid X_1^{t-1}) = P(X_t \mid X_{t-m}^{t-1}) \quad \forall t > m+1, \ \forall X_1^t.$$

Again the maximum likelihood technique fails when applied to the estimation of the parameter m.

In this thesis we study the problem of order estimation for Markov chains and hidden Markov chains. The technique we adopt is based on the compensation of the likelihood function. A penalty term, decreasing in q (or m), is added to the maximum likelihood and the resulting compensated likelihood is maximized with respect to q (or m). Proper choice of the penalty term allows the strongly consistent estimation of the structural parameter. Accurate information on the almost sure asymptotic behavior of the maximum likelihood is of critical importance for the correct choice of the penalty term and the Law of the Iterated Logarithm (LIL) is therefore the best tool for this study.

The technique that we have just (roughly) described and the same probabilistic tools have been used for the estimation of the structural parameters of ARMA processes (see e.g. [1], [11]), but we are not aware of any previous work that employs this approach for Markov chains or hidden Markov chains.

We conclude the introduction with a brief summary of the thesis. In Chapter 1 we formally define HMC's and collect some basic results that will be used in the sequel. Most of these results are available in the literature to which we refer the reader for a more detailed treatment. Chapter 2 is dedicated to the proof of the consistency of the maximum likelihood estimator (MLE) of the parameters of a HMC. The novelty with respect to Baum and Petrie [4], where consistency was first proved, is that we do not require the observed process to be a HMC. The main results of this chapter are new. In Chapter 3 we deal with the estimation of the order of a Markov chain. First we use the LIL to find delicate bounds on the asymptotic behavior of the maximum likelihood estimator and then use the bounds to construct strongly consistent estimators of the order. The main results of Chapter 3 are new, moreover the chapter is practically self-contained. The final Chapter 4 is dedicated to the estimation of the order of HMC's. The behavior of the maximum likelihood is difficult to evaluate because no explicit expressions

for the estimators are available. The LIL works for one special case, but we must use other methods to evaluate the asymptotics. We obtain some results with an approximation technique that uses Markov chains of increasing order m to approximate the given HMC. These results are too weak to solve the problem of order determination and we will have to resort to a result from Information Theory to get the necessary asymptotics of the maximum likelihood. Apart from this heavy dependence from Information Theory the rest of Chapter 4 is new.

Chapter 1

Hidden Markov Chains

In Section 1 we define formally HMC's adopting the elegant formalism originated in Stochastic Realization Theory, show its equivalence to the definition given in the introduction and present some useful formulas for the computation of probabilities of interest. In Section 2 we review some results from Realization Theory that demonstrate the elusive nature of the notion of minimality for HMC's. We will attempt (with modest success) to circumvent the difficulty introducing the class of regular HMC's. In Section 3 we give two results on the equivalence of representations. They are not the best available but will be enough for our purposes. In the final section we define parametric families of HMC's that will be used later and a sufficient condition for their identifiability is given. The results presented in this chapter are scattered in the literature, our main goal was to bring them together as coherently as possible.

1.1 HMC fundamentals

There are many equivalent ways of defining HMC's. We particularly like the definition that originated in Realization Theory [23] and we will borrow it.

Definition 1.1.1 *(SFSS)*
A pair $\{X_t, Y_t\ t \epsilon N\}$ *of stochastic processes defined on a probability space* $(\Omega, \mathcal{F}, \mathcal{P})$ *and taking values in the finite set* $\mathcal{X} \times \mathcal{Y}$ *is said to be a stationary finite stochastic system (SFSS) if the following conditions are met:*

i) (X_t, Y_t) *are jointly stationary*

ii) $\mathrm{P}(Y_{t+1} = y_{t+1}, X_{t+1} = x_{t+1} \mid Y_1^t = y_1^t, X_1^t = x_1^t)$
$= \mathrm{P}(Y_{t+1} = y_{t+1}, X_{t+1} = x_{t+1} \mid X_t = x_t)$

The processes X_t and Y_t are called respectively the state and the output of the SFSS. The cardinality of $\mathcal{X}$ will be called the size *of the SFSS.*

□

Definition 1.1.2 *(HMC)*

A stochastic process Y_t with value in the finite set $\mathcal{Y}$ is a Hidden Markov Chain (HMC) if it is equivalent to the output of a SFSS.

□

Recall that two stochastic processes are said to be equivalent if their laws coincide. Definition 1.1.2 has therefore to be interpreted as follows: the process Y_t is a HMC if its probability distribution function $P_Y(y_1^n) := Pr[Y_1^n = y_1^n]$ can be represented as $P_Y(y_1^n) = P(\tilde{Y}_1^n = y_1^n)$ where $\tilde{Y}_t$ takes value in $\mathcal{Y}$ and is the output of a SFSS. Observe that we do not require $\tilde{Y}_t$ to be defined on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$ as Y_t, they can be completely different objects but they are indistinguishable from observation. From now on when we refer to Y_t as a HMC we will actually refer to any process $\tilde{Y}_t$ in the same equivalence class. We will refer to any *SFSS* $(X_t, \tilde{Y}_t)$ with $\tilde{Y}_t$ equivalent to Y_t as a representation of the HMC Y_t.

Example (adapted from Ornstein [20])

A box contains a roulette wheel. We look at all possibilities for two consecutive spins of the wheel and divide these into two classes. Each time we spin the wheel (the spins are independent), we look at the last two spins and print out 0 if they fall in the first class and 1 if the fall the second class. The output of the box is a HMC. Observe that the probability of printing a 1 at time n cannot be determined

from the observation of the previous values (unless the two classes have trivial configurations), i.e., the output need not be Markov of any order m.

□

In the introduction we referred to HMC's as stationary processes of the form $Y_t = f(X_t)$ where X_t is a stationary Markov Chain, but this is equivalent to definition 1.1.2. Clearly if $Y_t = f(X_t)$ with X_t stationary Markov, the pair (X_t, Y_t) will be a SFSS and Y_t a HMC according to 1.1.2. Let now Y_t be a HMC according to 1.1.2 and X_t be the state process of a SFSS associated with Y_t. If we sum $1.1.1ii$ over y_{t+1} we get $P(X_{t+1} = x_{t+1} \mid X_1^t = x_1^t, Y_1^t = y_1^t) = \mathrm{P}(X_{t+1} = x_{t+1} \mid X_t = x_t)$ and after taking conditional expectations with respect to X_1^t we have $P(X_{t+1} = x_{t+1} \mid X_1^t = x_1^t) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$. Therefore X_t is a Markov Chain. As a direct consequence of $1.1.1ii$ we also have that the process $S_t = (X_t, Y_t)$ is a Markov Chain. Taking $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ to be the projection map on the second component i.e., $f(x, y) = y$ we get the representation $Y_t = f(S_t)$ as desired.

We insisted on the fact that in general HMC's do not have finite memory, nevertheless their laws are completely specified by a finite number of parameters. In fact to specify the laws of a SFSS it is sufficient to specify the finite set of matrices $\{M(y), y \epsilon \mathcal{Y}\}$ whose elements are: $m_{ij}(y) := P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i)$ $i, j = 1, 2, \cdots \mid \mathcal{X} \mid$. Observe that the matrix $A := \sum_y M(y)$ is the transition matrix of the Markov Chain X_t. If to the matrices $M(y)$ we add an initial distribution vector π such that $\pi = \pi A$ (stationarity) then we have a complete specification of the laws of the $SFSS$. Very often in the literature the following "factorization" hypothesis is made:

$$P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i) = P(Y_{t+1} = y \mid X_{t+1} = j) P(X_{t+1} = j \mid X_t = i)$$

Since the factorization hypothesis always holds for the process $S_t = (X_t, Y_t)$ we will assume it without loss of generality (later we will impose conditions on the values and therefore the assumption will become restrictive).

Let $b_{iy} := P(Y_t = y \mid X_t = i)$, B the $\mid \mathcal{X} \mid \times \mid \mathcal{Y} \mid$ matrix of the b_{iy}'s, and $B_y := \mathrm{diag}\{b_{1y}, b_{2y}, \cdots b_{cy}\}$ (where $c := \mid \mathcal{X} \mid$). The factorization hypothesis now gives: $M(y) = AB_y$.

We conclude this section with a collection of formulas. The formulas for the probability of some cylinders in terms of the parameters are well known and will be used later. The other formulas are obtained by elementary algebraic manipulations from 1.1.1*ii*. We present them here because they shed some light on the nature of the dependencies between the parameters of $SFSS$.

In the sequel when notationally convenient and not misleading we will identify random variables with their values e.g. $P(y_t)$ will mean $P(Y_t = y_t)$.

Lemma 1.1.1

$$\begin{aligned} P(y_{t+1}^n \mid X_t = i) &= e_i^T M(y_{t+1}) \cdots M(y_n) e \quad \forall n \geq t+1 \\ P(y_1^t, X_t = i) &= \pi M(y_1) \cdots M(y_t) e_i \\ P(y_1^n) &= \pi M(y_1) \cdots M(y_n) e \end{aligned}$$

where e_i is the i-th standard basis vector of $\mathcal{R}^c$ and $e = \sum e_i$

Proof: everything follows directly from 1.1.1*ii*.

□

For future reference let us introduce the following definitions. For any word $v = y_1 y_2 \cdots y_k$ define:

$$\begin{aligned} M(v) &:= M(y_1)M(y_2)\cdots M(y_k) \quad \textit{a square matrix of size } |\mathcal{X}|. \\ g(v) &:= \pi M(v) \quad \textit{a row vector of size } |\mathcal{X}|. \\ h(v) &:= M(v)e \quad \textit{a column vector of size } |\mathcal{X}|. \end{aligned}$$

With ϕ we denote the null word and define $g(\phi) := \pi, h(\phi) := e$.

The formula for the output probability can be written is scalar form as follows:

$$P(y_1^n) = \sum_{x^n} P(y_1^n, x_1^n) = \sum_{x^n} P(y_1^n \mid x_1^n) P(x_1^n)$$

Given x_1^n, the sequence y_1^n is independent in fact:

$$\begin{aligned} P(y_1^n \mid x_1^n) &= \frac{P(y_1^n, x_1^n)}{P(x_1^n)} \\ &= \frac{P(y_n, x_n \mid x_{n-1}) P(x_1^{n-1}, y_1^{n-1})}{P(x_n \mid x_{n-1}) P(x_1^{n-1})} \\ &= P(y_n \mid x_n) P(y_1^{n-1} \mid x_1^{n-1}) \end{aligned}$$

and proceeding by induction we get $P(y_1^n \mid x_1^n) = \prod_{k=1}^n P(y_k \mid x_k)$

Therefore:

$$\begin{aligned} P(y_1^n) &= \sum_{x^n} P(x_1)P(y_1 \mid x_1) \prod_{k=2}^{n} P(x_{k+1} \mid x_k)P(y_{k+1} \mid x_{k+1}) \\ &= \sum_{x^n} \pi_{x_1}\, b_{x_1 y_1}\, a_{x_1 x_2}\, b_{x_2 y_2}\, a_{x_2 x_3} \cdots a_{x_{n-1} x_n}\, b_{x_n y_n} \end{aligned}$$

For the final part of the section we make the extra assumption that all probabilities are strictly positive (it is actually enough to assume $M(y) > 0$).

Lemma 1.1.2 *The following conditions are equivalent* $(1 \leq t \leq n)$*:*

i) $P(X_{t+1}, Y_{t+1} \mid X_1^t, Y_1^t) = P(X_{t+1}, Y_{t+1} \mid X_t)$

ii) $P(X_{t-1}, Y_{t-1} \mid X_t^n, Y_t^n) = P(X_{t-1}, Y_{t-1} \mid X_t)$.

Proof: Since the process (X_t, Y_t) is Markov it is sufficient to check that $P(X_{t-1},\ Y_{t-1}\ \mid\ X_t,\ Y_t) = P(X_{t-1},\ Y_{t-1}\ \mid X_t)$.

But

$$\begin{aligned} P(X_{t-1}, Y_{t-1} \mid X_t, Y_t) &= \frac{P(X_t, Y_t \mid X_{t-1}, Y_{t-1})P(X_{t-1}, Y_{t-1})}{P(X_t, Y_t)} \\ &= \frac{P(X_t \mid X_{t-1})P(X_{t-1}, Y_{t-1})}{P(X_t)} \\ &= P(Y_{t-1} \mid X_{t-1})P(X_{t-1} \mid X_t). \end{aligned}$$

This last expression is independent from Y_t and therefore equals $P(X_{t-1},\ Y_{t-1}\ \mid\ X_t)$.

□

Lemma 1.1.3 *For* $(1 \leq t \leq n)$ *:*

i) $P(X_t \mid X_1^{t-1}, Y_1^n) = P(X_t \mid X_{t-1}, Y_t^n)$

ii) $P(X_t \mid X_{t+1}^n, Y_1^n) = P(X_t \mid X_{t+1}, Y_1^t)$

Proof: First we will prove i.

The LHS is:

$$\begin{aligned}\frac{P(X_1^t, Y_1^n)}{P(X_1^{t-1}, Y_1^n)} &= \frac{P(Y_{t+1}^n \mid X_1^t, Y_1^t)P(X_1^t, Y_1^t)}{P(Y_t^n \mid X_1^{t-1}, Y_1^{t-1})P(X_1^{t-1}, Y_1^{t-1})} \\ &= \frac{P(Y_{t+1}^n \mid X_t)}{P(Y_t^n \mid X_{t-1})}, P(X_t, Y_t \mid X_{t-1})\end{aligned}$$

since the last expression does not depend on X_1^{t-2} it must coincide with $P(X_t \mid X_{t-1}, Y_1^n)$. Now we must prove that $P(X_t \mid X_{t-1}, Y_1^n) = P(X_t \mid X_{t-1}, Y_1^n)$. But

$$\begin{aligned}P(X_t \mid X_{t-1}, Y_1^n) &= \frac{P(Y_1^n, X_t, X_{t-1})}{P(Y_1^n, X_{t-1})} \\ &= \frac{P(Y_t^n, X_t \mid Y_1^{t-1}, X_{t-1})P(Y_1^{t-1}, X_{t-1})}{P(Y_t^n \mid X_{t-1}, Y_1^{t-1})P(Y_1^{t-1}, X_{t-1})} \\ &= \frac{P(Y_t^n, X_t \mid X_{t-1})}{P(Y_t^n \mid X_{t-1})} \\ &= P(X_t \mid X_{t-1}, Y_t^n)\end{aligned}$$

This proves *i)*. For *ii)* use the same technique and Lemma 1.1.2.

□

We next use Lemma 1.1.3 to find an expression for the posterior probability of the state process X_t as a function of the filter and the one-step predictor.

Lemma 1.1.4

$$P(X_1^n \mid Y_1^n) = \prod_{t=1}^{n} \frac{P(X_t \mid Y_1^t)}{P(X_{t+1} \mid Y_1^t)} \quad P(X_{t+1} \mid X_t)$$

Proof: Requires a little manipulation using $1.1.3ii$ and the easily proved fact that $P(Y_1^t \mid X_t, X_{t+1}) = P(Y_1^t \mid X_t)$.

□

Again by manipulating the formulas and with the help of Lemma 1.1.3 we find that:

$$P(Y_t, X_t \mid Y_1^{t-1}, X_1^{t-1}, X_{t+1}^n, Y_{t+1}^n) = P(Y_t \mid X_t), P(X_t \mid X_{t-1}, X_{t+1})$$

which gives us the structure of the neighborhood system for the Markov random field (X_t, Y_t).

1.2 Results from Realization Theory

In [12] Heller characterized the finite valued stationary processes Y_t that are HMC's. We need some preliminaries to present his results. $\mathcal{Y}$ will denote a finite set, $\mathcal{Y}^*$ the set of finite words from $\mathcal{Y}$, and $\mathcal{C}^*$ the set of probability distributions on $\mathcal{Y}^*$. $\mathcal{C}^*$ is convex. A convex subset $\mathcal{C} \subset \mathcal{C}^*$ is polyhedral if $\mathcal{C} = \text{conv } \{q_1(\cdot), \cdots, q_c(\cdot)\}$ i.e. $\mathcal{C}$ is generated by finitely many distributions $q_i(\cdot) \epsilon \mathcal{C}^*$. A convex polyhedral subset $\mathcal{C} \subset \mathcal{C}^*$ is stable if $\mathcal{C} = \text{conv } \{q_1(\cdot), \cdots, q_c(\cdot)\}$ and for $1 \leq i \leq c$ and $\forall y \epsilon \mathcal{Y}$ the conditional distributions $q_i(\cdot \mid y) := \frac{q_i(y \cdot)}{q_i(y)} \epsilon \mathcal{C}$. We are now ready to enunciate the main result.

Theorem 1.2.1 *(Heller [12])*

$P_Y(\cdot)$ is the pdf of a HMC iff the set

$\mathcal{C}_Y := conv\{P_Y(\cdot \mid u) \quad u \epsilon \mathcal{Y}*\}$

is contained in a polyhedral stable subset of $\mathcal{C}^*$.

□

For an elementary and insightful proof of Heller's theorem, see Picci [23] which we followed for the presentation of the result. Suppose that a given $P_Y(\cdot)$ satisfies the conditions of Heller's theorem and that the Y_t process is therefore a HMC. Two questions now arise naturally. We called any $SFSS$ $(X_t, \tilde{Y}_t)$ with $\tilde{Y}_t$ equivalent to Y_t a representation of the HMC Y_t and showed that the distributions of a $SFSS$ are completely specified by the set of parameters $\mathcal{M} := \{c, M(y), \pi\}$ where $c = \mid \mathcal{X} \mid$. It is therefore natural to identify a representation of the HMC Y_t with the set $\mathcal{M}$. When clear from the context we will omit c from the list of parameters.

The first question is: can the parameters of a representation be determined directly from $P_Y(\cdot)$?

Such a representation of Y_t is inherently non-unique and we would like to find the "simplest" one. Take $\mid \mathcal{X} \mid$ as a measure of complexity, and for a given HMC Y_t define its **order** as the minimum of $\mid \mathcal{X} \mid$ among all representations. A representation for which $\mid \mathcal{X} \mid$ equals the order is said to be a minimal representation.

The second question is: can the order be determined?

At present the answer to both questions is in the negative, but there are a few clues. Unless otherwise noted the following is derived from the works of Gilbert [10], Carlyle [6] and Paz [21]. Let $p(\cdot)$ be an arbitrary *pdf* (not necessarily HMC), and $v_1 \cdots v_n \; v'_1 \cdots v'_n$ $2n$ arbitrary words from $\mathcal{Y}^*$. The compound sequence matrix (c.s.m.) $P(v_1 \cdots v_n, \; v'_1 \cdots v'_n)$ is the $n \times n$ matrix with i, j element $p(v_i v'_j)$. The rank of $p(\cdot)$ is defined as the maximum of the ranks of all possible c.s.m. if such maximum exists or $+\infty$ otherwise. Suppose now that $p(\cdot)$ is the *pdf* of a HMC which admits a representation $\mathcal{M} := \{c, M(\cdot), \pi\}$ of size c. Using the definitions following Lemma 1.1.1 we have that: $P(v_1 \cdots v_n v'_1 \cdots v'_n) = G(v_1 \cdots v_n) H(v'_1 \cdots v'_n)$ where G, H are $n \times c$ and $c \times n$ matrices respectively. The i-th row of G is $g(v_i)$ and the j-th column of H is $h(v'_j)$ in fact: $p(v_j v'_j) = \pi M(v_i v'_j) e = \pi M(v_i) M(v'_j) e$ $= g(v_i) h(v'_j)$.

It clearly follows that the rank of a HMC cannot exceed the size of any of its representations and therefore in particular:

Lemma 1.2.1 *The rank of a HMC is a lower bound to its order.*

□

Remark: The concept of rank of a *pdf* is only loosely related to the HMC property because there are examples of *pdf's* with finite rank that are not HMC. Also there are examples of HMC's whose order is strictly greater than their rank.

A representation $\mathcal{M} = \{c, M(\cdot), \pi\}$ of size c is **regular** if the rank of the corresponding *pdf* equals c. Regular representations are minimal as a direct consequence of Lemma 1.2.1. As explained in the remark not all HMC's admit regular representations, but the following two results will justify our interest in them. The first result states that it is "easy" to check regularity.

Lemma 1.2.2 *A finite number of operations is sufficient to determine the regularity of a given representation* $\mathcal{M} = \{c, M(\cdot), \pi\}$.

Proof: Let c be the size of $\mathcal{M}$. We must check that there exist $2c$ words $v_1 \cdots v_c \, v'_1 \cdots v'_c$ such that the c.s.m. of size c: $G(v_1 \cdots v_c) \; H(v'_1 \cdots v'_c)$ is invertible. This is equivalent to checking the invertibility of both G and H. To complete the

proof it is sufficient to show that rank $G(v_1 \cdots v_c)$ attains its maximum for at least one set of words $(v_1 \cdots v_c)$ with $\mid v_i \mid \leq c\ (1 \leq i \leq c)$ (and similarly for $H(v'_1 \cdots v'_c)$).

Let $L_k := span\{g(v)\ v \epsilon \mathcal{Y}^*, \mid v \mid = k\}$, L_k is a linear subspace of the vector space $\mathcal{R}^c$. We will show that for $k \geq c$ all subspaces L_k are identical. Since $g(v) = \sum_y g(vy)$ we have that $L_k \subset L_{k+1}$, and since $g(vy) = g(v)M(y)$ we have that $L_k = L_{k+1} \Rightarrow L_{k+m} = L_k\ \forall m \geq 1$. Let J be the first integer for which $L_J = L_{J+1}$ then $\dim L_o + J \leq \dim L_J \leq c$ and we conclude that $J \leq c$.

The proof for H is analogous.

□

The second result states that almost all representations are regular. Let Γ be the set of all $\mathcal{M} := \{c, M(\cdot), \pi\}$ of size c. Γ is a compact set in $\mathcal{R}^k$ for some k depending on c.

Lemma 1.2.3 *The non-regular elements of Γ are a closed subset of $\mathcal{R}^k$-Lebesgue measure zero.*

Proof: The non-regular points of Γ are characterized by the vanishing of the determinant of a finite number of matrices.

□

We conclude this section with an observation on the structure of the *pdf* of HMC's.

Lemma 1.2.4 *If Y_t is a HMC known to admit a representation of size c then its pdf $p(\cdot)$ is completely determined by the values $\{p(v), \mid v \mid \leq 2c\}$.*

Proof: The existence of a representation of size c implies that the rank of $p(\cdot)$ cannot exceed c. Let r be the rank of $p(\cdot)$ and P an $r \times r$ c.s.m. of rank r. Such a P exists, moreover the proof of Lemma 1.2.2 shows that the words $v_1 \cdots v_r, v'_1 \cdots v'_r$ defining P can be chosen of length $\leq c$. Let $w \epsilon \mathcal{Y}^*$ arbitrary, since P is invertible we have that: $[p(wv'_1) \cdots p(wv'_r)] = a(w)P$ for some row vector $a(w)$ which only depends on w and P. Construct the c.s.m. $\tilde{P}(v_1, \cdots v_r, w, v'_1 \cdots v'_r, \phi)$.

Since P is of maximal rank, the rank of $\tilde{P}$ must also be r and therefore:

$$\begin{aligned} p(w) &= [p(wv_1') \cdots p(wv_r')]P^{-1}[p(v_1) \cdots p(v_r)]^T \\ &= a(w)[p(v_1) \cdots p(v_r)]^T \end{aligned}$$

□

We gave the result in this form since it can easily be proved from what we already presented. The best possible bound on the length of the words determining $p(\cdot)$ completely is $2c$-*1* (see Paz[21]). In Carlyle [6] a **recursive algorithm** is given for the computation of $p(\cdot)$ of long strings starting from $\{p(v), \mid v \mid \leq 2c-1\}$.

Remark: It is always possible to construct a c.s.m. P of maximum rank taking $v_1 = v_1' = \phi$ this can be seen taking $w = \phi$ in the proof of Lemma 1.2.4. Expanding the determinant of $\tilde{P}$ along the last row and along the last column and comparing we get the result.

1.3 Equivalent representations

In this section we study conditions for the equivalence of representations. Our results are not the best possible (see [13]) but they are relatively straightforward.

The following is a sufficient condition for equivalence.

Lemma 1.3.1 *Let* $\mathcal{M} := \{c, M(\cdot), \pi\}$ *and* $\hat{\mathcal{M}} := \{\hat{c}, \hat{M}(\cdot), \hat{\pi}\}$. *If* X, Y *are* $c \times \hat{c}$ *and* $\hat{c} \times c$ *matrices respectively such that:*

$$\begin{aligned} \hat{M}(y) &= YM(y)X \quad \forall y \epsilon \mathcal{Y} \\ \hat{\pi} &= \pi X \\ \hat{e} &= Ye \\ XY &= I_c \end{aligned}$$

then $\mathcal{M}$ *and* $\hat{\mathcal{M}}$ *are equivalent.*

Proof: It is sufficient to verify that for an arbitrary word $w\epsilon\mathcal{Y}^*$ $\hat{\pi}\hat{M}(w)\hat{e} = \pi M(w)e$. This follows immediately by substitution.

□

Remark: The condition $XY = I_c$ implies rank $X \geq c$ and rank $Y \geq c$, therefore the lemma is non-trivial only for $\hat{c} \geq c$.

If one of the representations is regular we can give a necessary condition for equivalence as follows.

Lemma 1.3.2 *Let $\mathcal{M} := \{c, \mathcal{M}(\cdot), \pi\}$ and $\hat{\mathcal{M}} := \{\hat{c}, \hat{\mathcal{M}}(\cdot), \hat{\pi}\}$ and assume $\mathcal{M}$ to be regular. If $\mathcal{M}$ is equivalent to $\hat{\mathcal{M}}$ then $\hat{c} \geq c$ and there exist X, Y of dimensions $c \times \hat{c}$, $\hat{c} \times c$ respectively such that:*

$$\begin{aligned} M(y) &= X\hat{M}(y)Y \quad \forall y\epsilon\mathcal{Y} \\ \pi &= \hat{\pi}Y \\ e &= X\hat{e} \\ XY &= I_c \end{aligned}$$

Proof: The necessity of $\hat{c} \geq c$ follows from the minimality of regular representations. We will exhibit a pair of matrices X, Y satisfying the conditions.

Since $\mathcal{M}$ is regular there exists an invertible c.s.m. $P(v_1 \cdots v_c, v_1' \cdots v_c')$. By the last remark of Section 1.2 we can always select $v_1 = v_1' = \phi$. Therefore $P(\phi, v_2 \cdots v_c, \phi, v_2' \cdots v_c') = G(\phi, v_2 \cdots v_c)H(\phi, v_2' \cdots v_c')$ where both G and H are invertible. Observe that the first row of G is π and the first column of H is e. Since $\hat{\mathcal{M}}$ and $\mathcal{M}$ are equivalent they have the same c.s.m. In particular this means that:

a) $\hat{G}(\phi, v_2 \cdots v_c)\hat{H}(\phi, v_2' \cdots v_c') = G(\phi, v_2 \cdots v_c)H(\phi, v_2' \cdots v_c')$

b) $\hat{G}(\phi, v_2 \cdots v_c)\hat{M}(y)\hat{H}(\phi, v_2' \cdots v_c') = G(\phi, v_2 \cdots v_c)M(y)H(\phi, v_2' \cdots v_c') \quad (\forall y\epsilon\mathcal{Y})$

Observe that $\hat{G}$, and $\hat{H}$ are $c \times \hat{c}$ and $\hat{c} \times c$ respectively and each of rank c since their product is of rank c.

Define $X := G^{-1}\hat{G}$ and $Y := \hat{H}H^{-1}$. Then by b) $M(y) = X\hat{M}(y)Y$. To verify that $\pi = \hat{\pi}Y$ observe that $\hat{\pi}$ is the first row of $\hat{G}$ therefore $\hat{\pi}Y$ = (first row of $\hat{G})\hat{H}H^{-1}$ = first row of $(\hat{G}\hat{H})H^{-1}$= first row of $(GH)H^{-1} = \pi$. Analogously it can be proved that $e = X\hat{e}$. Finally $XY = G^{-1}\hat{G}\hat{H}H^{-1} = G^{-1}GHH^{-1} = I_c$.

□

1.4 Families of HMC's

In this section we introduce the families of HMC's that will be used as model classes in the following chapters.

From now on $\mathcal{Y}$ will be a fixed finite set with $\mid \mathcal{Y} \mid = r$. The family Θ of all HMC's of all orders (taking values in $\mathcal{Y}$) can be identified with the family of all $\theta := \{c_\theta, M_\theta(y), \pi_\theta\}$ with $c \epsilon N$. For $\theta \epsilon \Theta$ define $P_\theta(y_1^n) := \pi_\theta M_\theta(y_1^n)e_{c_\theta}$ (we will often drop the subscripts in the RHS and simply write $P_\theta(y_1^n) = \pi M(y_1^n)e$). Define $\Theta_q := \{\theta \epsilon \Theta; c_\theta = q\}$.

Lemma 1.4.1
$\forall\, q \forall \theta \epsilon \Theta_q\ \exists \bar{\theta} \epsilon \Theta_{q+1}$ *such that* $P_{\bar{\theta}}(\cdot) = P_\theta(\cdot)$ *or, abusing the notation,* $\Theta_q \subset \Theta_{q+1}$.

Proof: Let $\theta = \{q, M(y), \pi\}$ and construct $\bar{\theta}$ as follows:

$$\bar{q} = q + 1, \quad \textit{and} \quad \bar{M}(y) := diag\{M(y), \bar{m}(y)\},$$

where $\bar{m}(y) \epsilon [0,1]$ and $\sum_y \bar{M}(y) = \bar{A}$ is a stochastic matrix, $\bar{\pi} = (\pi, 0)$. It is immediate to verify that $P_{\bar{\theta}}(\cdot) = P_\theta(\cdot)$.

□

Statisticians refer to families satisfying Lemma 1.4.1 as *nested families*. A few considerations about the identifiability of Θ are now in order. A point $\theta \epsilon \Theta_q$ is **identifiable in** Θ_q if for any $\theta' \neq \theta (\theta' \epsilon \Theta_q) P_\theta(\cdot) \neq P_{\theta\prime}(\cdot)$ i.e. for at least one word $w, P_\theta(w) \neq P_{\theta\prime}(w)$. This definition is too strong and it would give no identifiable points in any Θ_q. In fact for a given θ at least the (finitely many) points θ' obtained by permutations of the rows and columns of $M(y)$ and π give $P_{\theta\prime}(\cdot) = P_\theta(\cdot)$. We

will say that $\theta \epsilon \Theta_q$ is identifiable modulo permutations (i.m.p.) if the only points $\theta' \epsilon \Theta_q$ with $P_\theta(\cdot) = P_{\theta\prime}(\cdot)$ are obtained by permutation as described above. Regular points $\theta \epsilon \Theta_q$ (i.e. points for which rank $P_\theta = q$) are good candidates for being i.m.p. but a few (mild) extra conditions must be added. In [22] Petrie proves a theorem on identifiability that we will adapt to our case.

Definition 1.4.1
$\theta = \{q, M(y), \pi\}$ *is a Petrie point if*
θ *is regular*
$M(y)$ *is invertible* $\forall y$
$\exists y \epsilon \mathcal{Y}$ *such that* b_{iy}, $(i = 1, 2, \cdots q)$, *are distinct.*

Theorem 1.4.1 *(Petrie [22] adapted)*
The Petrie points of Θ_q *are identifiable modulo permutation.*

Proof: Let $\theta \epsilon \Theta_q$ be a Petrie point. We show that if $\bar{\theta} \epsilon \Theta_q$ and $P_\theta(\cdot) = P_{\bar{\theta}}(\cdot)$ then $\bar{\theta}$ and θ differ by a permutation. By regularity there exists $v_1 \cdots v_q \;\; v_1' \cdots v_q'$ such that $P(v_1 \cdots v_q \, v_1' \cdots v_q' = G(v_1 \cdots v_q) H(v_1' \cdots v_q')$ is invertible together with $G(v_1 \cdots v_q)$ and $H(v_1' \cdots v_q')$. Since $P_{\bar{\theta}}(\cdot) = P_\theta(\cdot)$ their c.s.m.'s are identical and therefore:

(1) $\bar{G}(v_1 \cdots v_q)\bar{H}(v_1' \cdots v_q') = G(v_1 \cdots v_q)H(v_1' \cdots v_q')$.

We conclude that $\bar{G}$ and $\bar{H}$ are also invertible. (For convenience we dropped $v_1 \cdots v_q, v_1' \cdots v_q'$). Analogously $\forall y \epsilon \mathcal{Y}$ $\bar{G}\bar{M}(y)\bar{H} = GM(y)H$ (since these too are c.s.m.'s), from which we have: $\bar{M}(y) = \bar{G}^{-1}GM(y)H\bar{H}^{-1}$. From (1) $H\bar{H}^{-1} = G^{-1}\bar{G}$. Let $X := G^{-1}\bar{G}$ (invertible) then:

(2) $\bar{M}(y) = X^{-1}M(y)X$.

To conclude it is enough to prove that X is a premutation matrix. Toward this end sum (2) over $\mathcal{Y}$ and get $\bar{A} = X^{-1}AX$, substituting into (2) we have $\bar{A}\bar{B}_y = X^{-1}AX\bar{B}y = X^{-1}AB_yX$. Since $M(y) = AB_y$ is invertible (θ is a Petrie point) so must be A. We finally have: $B_yX = X\bar{B}_y$. Let $y_o \epsilon \mathcal{Y}$ be such that b_{iy_o} $i = 1 \cdots q$ are distinct. From $B_{y_o}x_j = X\bar{B}_j$ we see that the j-th column of X satisfies: $B_{y_o}x_j = b_{jy_o}x_j$. Since B_{y_o} is diagonal with distinct elements, x_j can only be one

of the standard basis vectors $e_1, e_2 \cdots e_q$. This means that there is at most a 1 in each column of X. Observing that $Xe = e$ concludes the proof.

□

Lemma 1.4.2

The set of Petrie points is open and of full Lebesgue measure in Θ_q.

Proof: We already proved this for regular points in Lemma 1.2.3. The added hypotheses can be dealt with in the same way.

□

It will often be convenient to somewhat restrict the family Θ in order to simplify statistical considerations.

Definition 1.4.2

For $0 < \delta < \frac{1}{q}$ *define:*

$\Theta_q^\delta := \{\theta \epsilon \Theta_q;\ a_{ij} \geq \delta,\ b_{jy} \geq \delta, \quad \forall i, j, y\}$.

□

Remark: If $\theta \epsilon \Theta_q^\delta$ the stochastic matrix A_θ is certainly irreducible and aperiodic and its invariant vector π_θ is uniquely determined. In this case θ is completely specified by $\{M(y)\}$ or by $\{A, B\}$.

The following lemma is simple but it will be essential later. With the abuse of notation introduced in Lemma 1.4.1 we have:

Lemma 1.4.3

$$\Theta_q^\delta \subset \Theta_q^{\delta/2}$$

Proof: Let $\theta \epsilon \Theta_q^\delta$. Define X, Y matrices $q \times (q+1)$ and $(q+1) \times q$ respectively as follows:

$$X = \left[\begin{array}{c|cc} I_{q-1} & 0 & 0 \\ \hline 0 & \frac{1}{2} & \frac{1}{2} \end{array}\right], Y = \left[\begin{array}{c} I_q \\ \hline e_q^T \end{array}\right]$$

and $\bar{M}(y) = YM(y)X \quad \forall y \epsilon \mathcal{Y}$.

It is easily checked that the conditions for the validity of Lemma 1.3.1 are satisfied and therefore $\bar{\theta} = \{\bar{M}(y)\}$ is equivalent to $\bar{\theta}$. The definitions of X and Y also guarantee that $\theta \epsilon \Theta_q^{\delta/2}$.

□

Chapter 2

HMC's as Models of Stationary Processes

The consistency of the Maximum Likelihood Estimator (MLE) for HMC's was established in [4] under the assumption that the true distribution of the observations comes from a HMC. Here we show that, if Y_t is stationary and ergodic, the MLE taken on a class of HMC's converges to the model closest to the true distribution in the divergence sense. The result in [4] is therefore a special case of ours. In Section 2.1 we briefly review the misspecified model approach that we followed here. In Section 2.2 we present our version of the consistency of the MLE. In Section 2.3 we prove a slightly generalized version of the Shannon-McMillan-Breiman theorem which is related to the consistency results of the previous section. The final Section 2.4 settles a technical problem.

2.1 The misspecified model approach to parameter estimation

Suppose a given series of observations $\{y_1, y_2 \cdots y_n\}$ is to be modeled for some specific reason. For example we might want to predict y_{n+1} or compress $\{y_1 \cdots y_n\}$ for storage. Confronted with this problem a statistician would most likely set up a related parameter estimation problem as follows. He or she would first assume that the sample is generated by some unknown stochastic mechanism, let us say $y_k = g_k(\omega), 1 \leq k \leq n$. The observed data sample is now interpreted as

the initial segment of a realization of an unknown stochastic process. Based on prior information, insight, and mathematical tractability, a class of models would then be selected. The models in the class will be denoted $\{f_k(\cdot,\ \theta), \theta \epsilon \Theta\}$ where $\{f_k(\cdot,\theta)\}_{k\geq 1}$ is a stochastic process whose probability law is completely specified by the parameter θ. The modeling problem is now reduced to an estimation problem. According to some specified criterion of optimality the statistician selects a model, i.e. estimates the θ, that best fits the data. Let us call the estimator based on n observations $\hat{\theta}_n$.

How are we to judge the quality of $\hat{\theta}_n$? Ideally we should compare $f_k(\cdot,\hat{\theta}_n)$ to $g_k(\cdot)$ but the latter is unknown. There are two possible solutions. The classical one is to assume that the unknown process g_k is actually a member of the selected class i.e. $g_k(\cdot) = f_k(\cdot,\theta_0)$ for some true (but unknown) θ_0. The estimator $\hat{\theta}_n$ is then judged to be good if it behaves well, uniformly with respect to $\theta_0 \epsilon \Theta$. Based on this idea a great deal of statistical theory has been developed on the asymptotic properties of various estimators.

The second approach (which we prefer) does not rely on the existence of a true parameter θ_0 in Θ. After all the class of models was chosen more or less arbitrarily, why should g_k belong to it? The problem is transformed into one of best approximation. A distance $d(\cdot,\cdot)$ between probability measures is introduced and θ_* is defined as $d(P_g, P_{\theta_*}) = \min_\theta d(P_g, P_\theta)$. The estimator $\hat{\theta}_n$ is judged to be good if it is close to θ_*. In the statistical literature this is known as the misspecified model approach. We learned about misspecified models from Nishii [19] which treats the *iid* case.

2.2 HMC's as misspecified models of stationary processes

In this section we introduce our first statistical result involving HMC's. We observe the process Y_t with values in the finite set $\mathcal{Y}$. **The only assumptions on Y_t are stationarity and ergodicity.** Denote by Q the probability distribution on $\mathcal{Y}^*$ induced by Y_t. The class of models for Y_t will be $\Psi := \Theta_q^\delta$ with q and δ fixed (see Section 1.4 for the notations). Notice that we do not assume a priori that $Q = P_{\theta_0}$

for some $\theta_0 \epsilon \Psi$. Instead we are adopting the misspecified model approach described in the previous section.

Our goal is to prove the analog of the consistency of the maximum likelihood estimator in this set up. Toward this end define:

$$h_n(\theta, Y) := \frac{1}{n} \log P_\theta(Y_1^n) \tag{2.1}$$

Following the terminology from [19] we define the **quasi-maximum likelihood estimator** (q.m.l.e.) $\hat{\theta}(n)$ as:

$$\hat{\theta}(n) := \{\tilde{\theta} \epsilon \Psi;\ h_n(\tilde{\theta}, Y) = \sup_{\theta \epsilon \Psi} h_n(\theta, Y)\} \tag{2.2}$$

Remark: $\hat{\theta}(n)$ is defined as a set because no unicity is guaranteed for this class of models. It is easy to see that in (2.2) the sup can be replaced by a max.

We need a notion of "distance" between Q and the P_θ's. A reasonable choice justified by its widespread use in statistics and engineering would be the divergence rate:

$$D(Q \parallel P_\theta) := \lim_{n \to \infty} \frac{1}{n} E_Q \left[\log \frac{Q(Y_1^n)}{P_\theta(Y_1^n)} \right] \tag{2.3}$$

Clearly we must prove the existence of the limit, for this distance to be well defined. Refering to the proof given later we state here that indeed 2.3 is a legitimate definition.

It will also be proved later that:

$$D(Q \parallel P_\theta) = H_Q - H_Q(\theta) \geq 0 \tag{2.4}$$

where $H_Q := E_Q[\log Q(Y_0 \mid Y_{-\infty}^{-1})]$ is minus the entropy of Y_t under Q, and $H_Q(\theta) := E_Q[\log P_\theta(Y_0 \mid Y_{-\infty}^{-1})]$ is a well-defined and continuous function of $\theta \epsilon \Psi$.

Next define the **quasi-true parameter** set as:

$$\mathcal{N} := \{\tilde{\theta} \epsilon \Psi; \quad D(Q \parallel P_{\tilde{\theta}}) = \min_{\theta \epsilon \Psi} D(Q \parallel P_\theta)\}$$

A direct consequence of (2.4) is that:

$$\mathcal{N} = \{\tilde{\theta} \epsilon \Psi;\ H_Q(\tilde{\theta}) = \max_{\theta \epsilon \Psi} H_Q(\theta)\} \tag{2.5}$$

For the proof of Theorem 2.2.1 we need the following result, proved in Section 2.4.

$$h_n(\theta, Y) \to H_Q(\theta) \quad a.s.\ Q,\ uniformly\ in\ \theta. \tag{2.6}$$

Remark: We recall the notion of a.s. set convergence that will be used. For any subset $\mathcal{E} \subset \Psi$ define the ε-**fattened set** $\mathcal{E}_\varepsilon := \{\theta \epsilon \Psi;\ \rho(\theta, \mathcal{E}) < \varepsilon\}$, where ρ is the euclidean distance. Then $\hat{\theta}(n) \to \mathcal{N}$ *a.s.* Q if $\forall \varepsilon > 0\ \exists\ N(\varepsilon, \omega)$ such that $\forall n \geq N(\varepsilon, \omega),\ \hat{\theta}(n) \subset \mathcal{N}_\varepsilon$.

We are now ready to state our result:

Theorem 2.2.1

$$\hat{\theta}(n) \to \mathcal{N} \quad a.s.\ Q$$

Proof: Recall that $\Psi := \Theta_q^\delta$ is compact. Fix $\varepsilon > 0$. $\mathcal{N}_\varepsilon$ being open, the complement $\mathcal{N}_\varepsilon^c$ is compact. Cover $\mathcal{N}_\varepsilon^c$ with euclidean open balls $B(\theta, \lambda_\theta)$ centered at θ, and of radius λ_θ. The radii λ_θ can be chosen so that $\forall \theta,\ B(\theta, \lambda_\theta) \subset \mathcal{N}_{\varepsilon/2}^c$ strictly. Let $\bar{B}(\theta, \lambda_\theta)$ be the closure of $B(\theta, \lambda_\theta)$. The following chain of inclusions is easily verified:

$$\mathcal{N}_\varepsilon^c \subset \bigcup_{\theta \epsilon \mathcal{N}_\varepsilon^c} B(\theta, \lambda_\theta) \subset \bigcup_{\theta \epsilon \mathcal{N}_\varepsilon^c} \bar{B}(\theta, \lambda_\theta) \subset \mathcal{N}_{\varepsilon/2}^c$$

By the compactness of $\mathcal{N}_\varepsilon^c$ there exists a finite subcovering:

$$\mathcal{N}_\varepsilon^c \subset \bigcup_{i=1}^{I} B_i \subset \bigcup_{i=1}^{I} \bar{B}_i \subset \mathcal{N}_{\varepsilon/2}^c \quad (where\ B_i := B(\theta_i, \lambda_{\theta_i}))$$

Let $H_Q^* := H_Q(\theta)\ |_{\theta \epsilon \mathcal{N}}$ (i.e. the maximum value attained by $H_Q(\cdot)$). By the uniform convergence of $h_n(\theta, Y)$:

$$\max_{\theta \epsilon \bar{B}_i} h_n(\theta, Y) \to \max_{\theta \epsilon \bar{B}_i} H_Q(\theta) = H_Q^* - \alpha_i \quad a.s.\ Q$$

for some $\alpha_i > 0$. Therefore for n large enough:

$$\max_{\theta \epsilon \bar{B}_i} h_n(\theta, Y) < H_Q^* - \frac{\alpha_i}{2}$$

Piecing together the I balls B_i and letting $\alpha := \min_i \alpha_i$ we have:

$$\max_{\theta \epsilon \mathcal{N}_\varepsilon^c} h_n(\theta, Y) < H_Q^* - \frac{\alpha}{2} \tag{2.7}$$

On the other hand the uniform convergence of h_n also implies that:

$$\sup_{\theta \epsilon \mathcal{N}_\epsilon} h_n(\theta, Y) \rightarrow \sup_{\theta \epsilon \mathcal{N}_\epsilon} H(\theta) = H_Q^* \quad a.s.\ Q$$

and therefore for n large enough:

$$\sup_{\theta \epsilon \mathcal{N}_\epsilon} h_n(\theta, Y) > H_Q^* - \frac{\alpha}{2} \tag{2.8}$$

Comparing (2.7) and (2.8) we conclude that $\hat{\theta}(n) \subset \mathcal{N}_\epsilon$.

□

This proof is even simpler than the one given by Baum and Petrie [4] for the case of perfect modeling (i.e. $Q = P_{\theta_0}$ for some $\theta_0 \epsilon \Psi$) because it uses the uniform convergence of $h_n(\theta,\ Y)$.

2.3 A generalization of the Shannon-McMillan-Breiman theorem

In this section we present a slightly generalized version of the Shannon-McMillan-Breiman (SMB) theorem and prove, *en passant*, (2.3) and (2.4). The SMB theorem, first introduced by Shannon in 1948, has already a rich history of extensions and generalizations vestiges of which are found in its very name. The classic version of the theorem is the following:

Theorem 2.3.1
Let Y_t be a finitely valued stationary ergodic process with probability distribution $Q(\cdot)$. Then:

$$\frac{1}{n} \log Q(Y_1^n) \rightarrow E_Q[\log Q(Y_0 \mid Y_{-\infty}^{-1})] \quad a.e.\ and\ in\ L_1$$

□

In this form the theorem has direct application in Information Theory because it allows the estimation of the entropy rate of a finite alphabet stationary ergodic source. Generalizations of theorem 2.3.1 have appeared for the case of real valued processes. Barron [3] gives the following version.

Theorem 2.3.2

Let $(\Omega, \mathcal{F})$ be a sequence space i.e. $\Omega = \mathcal{R}_0^\infty$ and $\mathcal{F} = \mathcal{B}_0^\infty$ where $\mathcal{R}$ is a standard Borel space and $\mathcal{B}$ its Borel σ-algebra. Let M be a finite order, stationary, Markov measure on $(\Omega, \mathcal{F})$ and Y_t a stationary ergodic process with values in $\mathcal{R}$ and distribution Q. Assume absolute continuity of the n-th order marginal of Q with respect to the n-th order marginal of M and denote the corresponding density by $q(y_0^{n-1})$. Define the divergence rate of the true distribution Q with respect to the reference measure M as:

$$D_1(Q \parallel M) := \lim_k E_Q[\log q(Y_k \mid Y_0^{k-1})]$$

Then $D_1(Q \parallel M)$ is well defined and moreover:

$$\frac{1}{n} \log q(y_1^n) \to D_1(Q \parallel M) \quad a.e. \text{ and in } L_1$$

□

In this form the theorem becomes very useful in statistics, (see [3] for some applications).

The requirement that M be Markov for 2.3.2 to obtain seems to be almost necessary (see Kieffer [14]). Our result generalizes Theorem 2.3.2 to reference measures M of the HMC type but it applies only to finitely valued processes.

Theorem 2.3.3

Let Y_t be a process with values in the finite set $\mathcal{Y}$. Assume Y_t to be stationary ergodic under the probability distribution Q and a HMC under the alternative distribution $P \epsilon \Theta_q^\delta$ for some fixed q and δ. Let $q(Y_1^k) = \frac{Q(Y_1^k)}{P(Y_1^k)}$ and define:

$$D_1(Q \parallel P) := \lim_k E_Q[\log q(Y_k \mid Y_0^{k-1})]$$

Then D_1 is well defined and moreover:

$$\frac{1}{n} \log \frac{Q(Y_1^n)}{P(Y_1^n)} \to D_1(Q \parallel P) \quad a.e.\ Q \tag{2.9}$$

□

The proof will be given through a series of lemmas. First we will prove, with the help of Lemma 2.3.1, the existence and finiteness of D_1. Lemmas 2.3.3 and 2.3.4 will allow us to find a more explicit expression for D_1. Using this new expression it will be easy to complete the proof, i.e. show (2.9), applying the ergodic theorem and the basic inequality for HMC's (A.3).

Lemma 2.3.1

$$\delta \leq P(Y_k \mid Y_0^{k-1}) \leq 1-\delta \qquad \forall k,\ \forall Y,\ \forall P \epsilon \Theta_q^\delta$$

Proof:

$$\begin{aligned} P(Y_k \mid Y_0^{k-1}) &= \sum_i P(Y_k, X_k = i \mid Y_0^{k-1}) \\ &= \sum_i P(Y_k \mid X_k = i)\ P(X_k = i \mid Y_0^{k-1}) \end{aligned}$$

Therefore $\forall Y_0^{k-1}$

$$\min_i\ P(Y_k \mid X_k = i)\ \leq\ P(Y_k \mid Y_0^{k-1})\ \leq\ \max_i P(Y_k \mid X_k = i)$$

and under the assumption that $P \epsilon \Theta_q^\delta$ we have:

$$\forall\ P,\ k,\ Y_0^k \quad :\quad \delta\ \leq\ P(Y_k \mid Y_0^{k-1})\ \leq\ 1-\delta$$

□

Lemma 2.3.2 *D_1 exists and is finite*

Proof: Define $R(Y_0^k) := Q(Y_0^{k-1})P(Y_k \mid Y_0^{k-1})$. It is easily verified that R is a probability measure on $\sigma(Y_0^k)$ and that $q(Y_k \mid Y_0^{k-1}) = \frac{Q(Y_0^k)}{R(Y_0^k)}$. Being a likelihood ratio $\{q(Y_k \mid Y_0^{k-1}),\ \sigma(Y_0^k)\}$ is an R-martingale and from the convexity of the function $x \log x$ it follows that $\{q(Y_k \mid Y_0^{k-1}) \log q(Y_k \mid Y_0^{k-1}),\ \sigma(Y_0^k)\}$ is an R-submartingale. All of this is trivially verified and it implies that $\{\log q(Y_k \mid Y_0^{k-1}),\ \sigma(Y_0^k)\}$ is a Q-submartingale. This can be seen as follows:

$$\begin{aligned} & E_R[q(Y_k \mid Y_0^{k-1}) \log q(Y_k \mid Y_0^{k-1}) \mid Y_0^{k-1}] \\ =\ & \sum_{y_k} R(y_k \mid Y_0^{k-1}) q(y_k \mid Y_0^{k-1}) \log q(y^k \mid Y_0^{k-1}) \\ =\ & \sum_{y_k} \frac{Q(Y_0^{k-1})P(y_k \mid Y_0^{k-1})}{Q(Y_0^{k-2})P(Y_{k-1} \mid Y_0^{k-2})} \cdot \frac{Q(y_k \mid Y_0^{k-1})}{P(y_k \mid Y_0^{k-1})} \log q(y_k \mid Y_0^{k-1}) \\ =\ & q(Y_{k-1} \mid Y_0^{k-2}) E_Q[\log q(Y_k \mid Y_0^{k-1}) \mid Y_0^{k-1}] \end{aligned}$$

Therefore, since $q \log q$ is an R-submartingale:

$$\begin{aligned} & E_Q[\log q(Y_k \mid Y_0^{k-1}) \mid Y_0^{k-1}] \\ = & \frac{1}{q(Y_{k-1} \mid Y_0^{k-2})} E_R[q(Y_k \mid Y_0^{k-1}) \log q(Y_k \mid Y_0^{k-1}) \mid Y_0^{k-1}] \\ \geq & \frac{q(Y_{k-1} \mid Y_0^{k-2}) \log q(Y_{k-1} \mid Y_0^{k-2})}{q(Y_{k-1} \mid Y_0^{k-2})} = \log q(Y_{k-1} \mid Y_0^{k-2}) \end{aligned}$$

which proves the assertion.

From the Q-submtg property of $\log q(Y_k \mid Y_0^{k-1})$ we immediately conclude that D_1 exists since the expectations $E_Q[\log q(Y_k \mid Y_0^{k-1})]$ increase in k and therefore have limit (possibly $+\infty$). The finiteness of D_1 is obtained as follows. For a fixed k we have:

$$\begin{aligned} E_Q[\log q(Y_k \mid Y_0^{k-1})] &= E_Q[\log Q(Y_k \mid Y_0^{k-1})] \\ &- E_Q[\log P(Y_k \mid Y_0^{k-1})] \end{aligned}$$

The first term on the RHS equals $E_Q[\log Q(Y_0 \mid Y_{-k}^{-1})]$ (by stationarity) and this converges to minus the entropy of Y_t which is bounded by $\log \mid \mathcal{Y} \mid$. The second term on the RHS is bounded because of Lemma 2.3.4. This concludes the proof of the lemma.

□

Remark: By stationarity we have that:

$$D_1(Q \parallel P) = \lim_{k \to \infty} E_Q \left[\log q(Y_0 \mid Y_{-k}^{-1})\right] \tag{2.10}$$

The following Lemmas 2.3.3 and 2.3.4 will allow us to find a more explicit expression for D_1.

Lemma 2.3.3

$$\log q(Y_0 \mid Y_{-k}^{-1}) \to Z \quad \textit{a.e. } Q \textit{ and in } L_1, \textit{ for some r.v. } Z \textit{ in } L_1$$

Proof: It follows from Barron [3] (equation 2.7 on page 1295) that:

$$E_Q \left[\sup_k \mid \log q(Y_0 \mid Y_{-k}^{-1}) \mid\right] \leq e\, D_1(Q \parallel P) + (e+3)$$

Since in our case $D_1 < \infty$ this proves that the random variables $\log q(Y_0 \mid Y_{-k}^{-1})$ are dominated by a Q-integrable r.v. (and are therefore uniformly integrable). We also have:

$$\begin{aligned} & \sup_k E_Q[\log^+ q(Y_0 \mid Y_{-k}^{-1})] \\ \leq & \sup_k E_Q[\mid \log q(Y_0 \mid Y_{-k}^{-1}) \mid \\ \leq & E_Q[\sup_k \mid \log q(Y_0 \mid Y_{-k}^{-1}) \mid] < \infty \end{aligned}$$

Since $\{\log q(Y_0 \mid Y_{-k}^{-1}),\ \sigma(Y_{-k}^0)\}$ is a Q-submtg it follows from standard theory that $\log q(Y_0 \mid Y_{-k}^{-1}) \to Z$ *a.e.* Q. The Q-integrability of Z and the convergence in L_1 follow from uniform integrability.

□

Lemma 2.3.4 *$P(Y_0 \mid Y_{-k}^{-1})$ converges to a limit $P(Y_0 \mid Y_{-\infty}^{-1})$* uniformly in $P \epsilon \Theta_q^\delta$ and $Y(\omega)$

Proof: Let $f_k := P(Y_0 \mid Y_{-k}^{-1})$ then f_k converges *iff* $\sum_{j=1}^k f_{j+1} - f_j$ converges. From the application of inequality A.0.3 it is easily seen that the last series converges absolutely and therefore it converges. The convergence is uniform in P and $Y(\omega)$ because ρ in A.0.3 is independent from them.

□

What makes Lemma 2.3.4 remarkable is the fact that the convergence of $P(Y_0 \mid Y_{-k}^{-1})$ is uniform in $Y(\omega)$. A similar, but weaker, result holds for the stationary measure Q. In particular $Q(y_0 \mid Y_{-k}^{-1})$ is a bounded Q-martingale $\forall y_0 \epsilon \mathcal{Y}$ and therefore it converges *a.e.* Q to the limit $Q(y_0 \mid Y_{-\infty}^{-1})$. Since $\mathcal{Y}$ is finite it also follows that $Q(Y_0 \mid Y_{-k}^{-1}) \to Q(Y_0 \mid Y_{-\infty}^{-1})$ *a.e.* Q.

Since $Q(Y_0 \mid Y_{-k}^{-1}) \to Q(Y_0 \mid Y_{-\infty}^{-1})$ *a.e.* Q. and $P(Y_0 \mid Y_{-k}^{-1}) \to P(Y_0 \mid Y_{-\infty}^{-1})$ for all $Y(\omega)$ we have:

$$\log q(Y_0 \mid Y_{-k}^{-1}) \to \log \frac{Q(Y_0 \mid Y_{-\infty}^{-1})}{P(Y_0 \mid Y_{-\infty}^{-1})} \quad a.e.\ Q.$$

From Lemma 2.3.3 and the unicity of the limit we can identify the $r.v. Z$ as:

$$Z = \log \frac{Q(Y_0 \mid Y_{-\infty}^{-1})}{P(Y_0 \mid Y_{-\infty}^{-1})} \quad a.e.\ Q$$

From the fact that $\log q(Y_0 \mid Y_{-k}^{-1}) \to Z$ in L_1 and (2.10) we finally conclude that:

$$D_1(Q \parallel P) = E_Q \left[\log \frac{Q(Y_0 \mid Y_{-\infty}^{-1})}{P(Y_0 \mid Y_{-\infty}^{-1})} \right]$$

To complete the proof of Theorem 2.3.3 it is now sufficient to show (2.9) i.e., using the last expression for D_1, that:

$$\lim_{n-\infty} \frac{1}{n} \log \frac{Q(Y_0^{n-1})}{P(Y_0^{n-1})} = E_Q \left[\log \frac{Q(Y_0 \mid Y_{-\infty}^{-1})}{P(Y_0 \mid Y_{-\infty}^{-1})} \right]. \tag{2.11}$$

This can be obtained from the ergodic theorem with the help of A.0.3. Define:

$$\begin{aligned} g_k(Y) &:= \log q(Y_0 \mid Y_{-k}^{-1}) \\ g(Y) &:= \log q(Y_0 \mid Y_{-\infty}^{-1}) \end{aligned}$$

and denote by T the shift operator. Equation (2.11) now becomes:

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} g_k\,(T^k Y) = E_Q\,[g(Y)] \quad a.e.\ Q$$

while the ergodic theorem gives:

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} g(T^k Y) = E_Q\,[g(Y)] \quad a.e.\ Q$$

To finish observe that:

$$\begin{aligned} &| \frac{1}{n} \sum_{k=0}^{n-1} g_k(T^k Y) - \frac{1}{n} \sum_{k=0}^{n-1} g(T^k Y) \mid \leq \\ &| \frac{1}{n} \sum_{k=0}^{n-1} \log Q(Y_k \mid Y_0^{k-1}) - \frac{1}{n} \sum_{k=0}^{n-1} \log Q(Y_k \mid Y_{-\infty}^{k-1}) \mid + \\ &| \frac{1}{n} \sum_{k=0}^{n-1} \log P(Y_k \mid Y_0^{k-1}) - \frac{1}{n} \sum_{k=0}^{n-1} \log P(Y_k \mid Y_{-\infty}^{k-1}) \mid \end{aligned}$$

The first term on the RHS equals:

$$| \frac{1}{n} \log Q(Y_0^{n-1}) - \frac{1}{n} \log Q(Y_0^{n-1} \mid Y_{-\infty}^{-1}) \mid \to \quad 0 \quad (a.e.\ Q)$$

Since both $\frac{1}{n}\log Q(Y_0^{n-1})$ and $\frac{1}{n}\log Q(Y_0^{n-1} \mid Y_{-\infty}^{-1})$ converge $(a.e.\ Q)$ to minus the entropy.

For the second term on the RHS we have:

$$|\frac{1}{n}\sum_{k=0}^{n-1}\log P(Y_k \mid Y_0^{k-1}) - \frac{1}{n}\log P(Y_k \mid Y_{-\infty}^{k-1})| \leq$$

$$\frac{1}{n}\sum_{k=0}^{n-1}|\log P(Y_k \mid Y_0^{k-1}) - \log P(Y_k \mid Y_{-\infty}^{k-1})| \to 0 \quad \textit{everywhere.}$$

The convergence to zero follows from the application of A.0.3 and Lemma 2.3.1. This completes the proof of Theorem 2.3.3.

Aside:

We prove here (2.3) and (2.4) from Section 2.2. In (2.3) we defined the divergence rate as: $D(Q \parallel P) := \lim \frac{1}{n} E_Q[\log \frac{Q(Y_0^{n-1})}{P(Y_0^{n-1})}]$.

Clearly $D(Q \parallel P) = \lim \frac{1}{n}\sum_{k=0}^{n-1} E_Q[\log q(Y_k \mid Y_0^{k-1})]$. From the definition of $D_1(Q \parallel P)$ given in Theorem 2.3.3 and the Cesaro convergence theorem it follows that if D_1 exists then D exists and $D = D_1$. From Lemma 2.3.2 we therefore conclude that D is indeed well defined. From what we have just proved we have:

$$D(Q \parallel P_\theta) = E_Q\left[\log \frac{Q(Y_0 \mid Y_{-\infty}^{-1})}{P_\theta(Y_0 \mid Y_{-\infty}^{-1})}\right]$$

Define:

$$H_Q(\theta) := E_Q\left[\log P_\theta(Y_0 \mid Y_{-\infty}^{-1})\right]$$

From Lemma 2.3.4 we know that $P_\theta(Y_0 \mid Y_{-\infty}^{-1})$ is the uniform limit of the sequence of continuous functions $P_\theta(Y_0 \mid Y_{-k}^{-1})$ and is therefore continuous in $\theta \epsilon \Theta_q^\delta$. Continuity of $H_Q(\theta)$ follows.

Defining $H_Q := E_Q[\log Q(Y_0 \mid Y_{-\infty}^{-1})]$ we get $D(Q \parallel P_\theta) = H_Q - H_Q(\theta)$ which is the decomposition claimed in (2.4).

□

2.4 Uniform convergence of $h_n(\theta, Y)$

We prove here (2.6) of Section 2.2, i.e. that:

$$h_n(\theta, Y) \to H_Q(\theta) \quad a.e.\ Q,\ \textit{uniformly in}\ \theta \epsilon \Psi$$

In (2.1) we defined $h_n(\theta, Y) := \frac{1}{n} \log P_\theta(Y_1^n)$ and from the results of Section 2.3 we already have that:

$$\frac{1}{n} \log P_\theta(Y_1^n) \to H_Q(\theta) \quad a.e.\ Q;\ pointwise\ in\ \theta \epsilon \Psi$$

Since Ψ is compact, to conclude that the convergence is uniform it is enough to show the equicontinuity of the functions $h_n(\theta, Y)$.

Lemma 2.4.1 *$h_n(\theta, Y)$ is an equicontinuous sequence*

Proof: We will show that $\forall \epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that:

$$\forall n \quad \mid h_n(\theta, Y) - h_n(\theta', Y) \mid \leq \epsilon \quad if \mid \theta - \theta' \mid < \delta(\epsilon)$$

This can easily be seen working directly with the Markov process. $S_t = (X_t, Y_t)$ which has state space $\mathcal{T} = \mathcal{X} \times \mathcal{Y}$. If $\theta = \{q, A, B\}, s = (i, y), \bar{s} = (j, \bar{y})$ then the transition matrix T of S_t has elements $t_{s\bar{s}} := P_\theta(S_{t+1} = \bar{s} \mid S_t = s) = a_{ij} b_{j\bar{y}}$. Since $\theta \epsilon \Theta_q^\delta$ the matrix T is strictly positive and admits a unique invariant vector τ. We have that:

$$P_\theta(S_1^n = s_1^n) = \tau_{s_1} \prod_{j=1}^{n} t_{s_j s_{j+1}} = \tau_{s_1} \prod_{(s,\bar{s})} t_{s\bar{s}}^{n_{s\bar{s}}}$$

where

$$n_{s\bar{s}} := \sum_{t=1}^{n} 1(S_t = s, S_{t+1} = \bar{s})$$

Let now $\theta' := \{q, A', B'\}$ be another point in Θ_q^δ. We have:

$$\mid h_n(\theta, S) - h_n(\theta', S) \mid \leq \mid \frac{1}{n} \log \tau_{s_1} - \frac{1}{n} \log \tau'_{s_1} \mid + \mid \frac{1}{n} \sum_{s,\bar{s}} n_{s\bar{s}} \log t_{s\bar{s}} - \frac{1}{n} \sum_{s,\bar{s}} n_{s\bar{s}} \log t'_{s\bar{s}} \mid$$

Since $\frac{n_{s\bar{s}}}{n} \leq 1$:

$$\leq \frac{1}{n} \mid \log \tau_{s_1} - \log \tau'_{s_1} \mid + \sum_{s,\bar{s}} \mid \log t_{s\bar{s}} - \log t'_{s\bar{s}} \mid$$

Define

$$\Delta(\theta, \theta') := \sum_{s\bar{s}} \mid \log t_{s\bar{s}} - \log t'_{s\bar{s}} \mid$$

From the expression for $t_{s\bar{s}}$ given above we have that:

$$\Delta(\theta,\theta') \leq \left[\sum_{ij} \mid \log a_{ij} - \log a'_{ij} \mid + \sum_{jy} \mid \log b_{jy} - \log b_{j\bar{y}} \mid\right]$$

Since all parameters are $\geq \delta$ this shows that for some $C > O$

$$\Delta(\theta,\theta') \leq C \parallel \theta - \theta' \parallel_1$$

Since τ is an eigenvalue of T of geometric multiplicity 1 its components are continuous functions of the components of T and therefore:

$$\mid \frac{1}{n}\log \tau_{s_1} - \frac{1}{n}\log \tau'_{s_1} \mid \leq \frac{C}{n} \parallel \theta - \theta' \parallel_1$$

The final estimate is:

$$\mid h_n(\theta,S) - h_n(\theta',S) \mid \leq 2C \parallel \theta - \theta' \parallel_1$$

Therefore $\forall \epsilon > 0$ there exist $N(\epsilon)$ and $\delta(\epsilon)$ such that for $n \geq N(\epsilon)$

$$\mid h_n(\theta,S) - h_n(\theta',S) \mid \leq \epsilon \quad if \quad \parallel \theta - \theta' \parallel \leq \delta(\epsilon) \tag{2.12}$$

To go back to the Y process observe that (2.12) can be written as:

$$\mid \frac{1}{n}\log \frac{P_\theta(S_1^n)}{P_{\theta'}(S_1^n)} \mid \leq \epsilon.$$

Therefore from

$$P_{\theta'}(S_1^n) \leq \exp\{n\epsilon\} P_\theta(S_1^n)$$

we get:

$$\begin{aligned} P_{\theta'}(Y_1^n) &= \sum_{x_1^n} P_{\theta'}(X_1^n, Y_1^n) \\ &= \sum_{x_1^n} P_{\theta'}(S_1^n) \leq \exp\{n\epsilon\} \sum_{x_1^n} P_\theta(X_1^n, Y_1^n) \\ &= \exp\{n\epsilon\} P_\theta(Y_1^n) \end{aligned}$$

and similarly exchanging the roles of θ and θ'.

□

Remark: The idea of working directly with the process S_t was suggested by Chuangchun Liu.

Chapter 3

Estimation of the Order of a Markov Chain

As originally planned this should have been a short review chapter on the applications of the Law of the Iterated Logarithm (LIL) in estimation problems. For the reasons explained in the introduction to Section 3.4 below, we decided to show the LIL in action on a real problem: the estimation of the order of a finite order Markov chain. We will later make use of these results in the context of HMC's, where finite order Markov chains will be useful to approximate the distribution of the HMC.

In Section 3.1 we briefly present a version of the LIL for square integrable martingales following Neveu [18]. Section 3.2 shows an application to Markov chains and gives a result on the estimation of the stationary vector. In Section 3.3 we introduce the notion of finite order Markov chain, give sufficient conditions for its ergodicity and study the asymptotics of the Maximum Likelihood Estimator (MLE) of the transition matrix. The delicate rate estimate given in Theorem 3.3.2 is the key to Section 3.4.

We start Section 3.4 clarifying the notion of order as "minimal memory" of the finite order Markov chain, and then use our asymptotic results to construct an estimator of the order. The basic idea of Section 3.4 occured to us while reading the beatiful booklet by Azencott and Dacunha-Castelle [1] on the estimation of the order of ARMA processes. Another useful source of ideas, especially for Section 3.3, has been Nishii [19] where the *iid* case is treated.

3.1 The LIL for Square Integrable Martingales

We sketchily present here the version of the LIL that is more convenient for our purposes. Everything is standard and can be found in full detail in e.g. Neveu [18]. Readers familiar with the result announced in the title are advised to only browse through this section.

The classical version of the Strong Law of Large Numbers (SLLN) asserts that if X_t is a sequence of *iid* random variables with $E \mid X_1 \mid < \infty$ and $EX_1 = \mu$ then $\frac{1}{n}S_n \to \mu$ a.e. where $S_n := \sum_{t=1}^{n} X_t$. Under the additional hypothesis that $EX_1^2 < \infty$ the variance $\sigma^2 := E(X_1 - \mu)^2$ is finite and the rate of convergence of $\frac{1}{n}S_n$ can be evaluated as follows:

$$\mid \frac{1}{n}S_n - \mu \mid = Z_n \sqrt{2\ \sigma^2 \frac{\log\log n}{n}}$$

where $\overline{\lim}\ Z_n = 1$ *a.s.*

This is the classical Kolmogorov's LIL. Many versions of the LIL have been developed to extend Kolmogorov's result to the non *iid* case. We will be content with the version for square integrable martingales as given in Neveu [18] pg. 147-156. The statement of the theorem is followed by a brief comment on its conditions and implications.

Theorem 3.1.1
Let $(X_n, n\epsilon N)$ *be a square integrable martingale such that*
$\sup_n \mid X_{n+1} - X_n \mid \leq c$ *a.e. for some finite constant* c.
If A_n *denotes the increasing process associated to the submartingale* $(X_n^2, n\epsilon N)$ *then:*

$$\overline{\lim} \frac{X_n}{\sqrt{2\ A_n \log\log A_n}} = 1 \quad a.s.\ on\ [A_\infty = \infty]$$
$$\underline{\lim} \frac{X_n}{\sqrt{2\ A_n \log\log A_n}} = -1 \quad a.s.\ on\ [A_\infty = \infty]$$

□

Comments 3.1.2

a) The sequence of r.v.'s X_n is a square integrable martingale if:

i) $EX_n^2 < \infty \quad \forall n \epsilon N$; and

ii) $E[X_n \mid F_{n-1}] = X_{n-1}$ where F_{n-1} is the σ-field generated by X_1^{n-1}.

b) The increasing process A_n (associated to the Doob decomposition of X_n^2).

is given by:

$$A_{n+1} - A_n = E[X_{n+1}^2 \mid F_n] - X_n^2$$

But for every r.v. Y and sub-sigma-field $\mathcal{B}$:

$$E[(Y - E(Y \mid \mathcal{B}))^2 \mid \mathcal{B}] = E[Y^2 \mid \mathcal{B}] - (E[Y \mid \mathcal{B}])^2$$

and therefore we obtain:

$$A_{n+1} - A_n = E[(X_{n+1} - X_n)^2 \mid F_n]$$

or:

$$A_n = \sum_{k=1}^{n} E[(X_k - X_{k-1})^2 \mid F_{k-1}]$$

with the convention that $X_0 = 0$.

c) A weaker form of the result is:

$$\overline{\lim} \frac{\mid X_n \mid}{\sqrt{2\ A_n \log\log A_n}} = 1 \quad a.s. \tag{3.1}$$

This can be immediately inferred from Theorem (3.1.1) and will be used very often for our results.

d) **Definition** $(O_{a.s})$

Let Z_n be a sequence of r.v.'s and $\alpha_n > 0$ a sequence of positive reals. We say that $Z_n = O_{a.s}(\alpha_n)$ if there exists a positive random variable C almost surely finite such that: $\mid Z_n \mid \leq C\alpha_n \quad \forall n$.

e) We will often be able to substitute A_n in (3.1) with n and get

$$\overline{\lim_n} \frac{\mid X_n \mid}{\sqrt{n \log\log n}} = \beta \quad a.s.$$

for some constant β. Which easily gives $X_n = O_{a.s}(\sqrt{n \log\log n})$

3.2 Application to Markov Chains

As a first example of application we will use the LIL to find the rate of convergence of the maximum likelihood estimators of the parameters of a Markov chain. The derivation of the results is only sketched because it will be given in full detail, for a more general case, in Section 3.3. Trivial as it might seem, Theorem 3.2.1 is a little puzzling because it cannot be derived directly.

Let $(X_t, t\epsilon N)$ be a finite Markov chain with state space $\mathcal{X} = \{1, 2, \cdots, q\}$ and assume that the transition matrix A^* of X_t is strictly positive. This is equivalent to the existence of a $\delta > 0$ such that $a_{ij} \geq \delta$ $(\forall i, j)$. To A^* there corresponds a unique invariant vector π^* whose components $\pi_j^* \geq \delta$ $(\forall j)$. We want to estimate A^* from the trajectory X_1^n. It is convenient to take as parameters $\theta := \{a_{ij}\ i = 1, \cdots q; j = 1, \cdots q-1\}$. The maximum likelihood estimator of θ based on n observations is given by:

$$\hat{a}_{ij}(n) = \frac{N(i,j,n)}{N(i,n)}$$

where

$$N(i,j,n) := \sum_{t=1}^{n-1} 1(X_t = i, X_{t+1} = j)$$

and

$$N(i,n) := \sum_{t=1}^{n} 1(X_t = i)$$

By the SLLN:

$$\hat{a}_{ij}(n) \to a_{ij}^* \quad a.s.$$

What is the rate of convergence? Observe that:

$$\hat{a}_{ij}(n) - a_{ij}^* = \frac{N(i,j,n) - a_{ij}^* N(i,n)}{N(i,n)} \tag{3.2}$$

Define:

$$M_{ij}(n) := N(i,j,n) - a_{ij}^* N(i,n)$$

It is easily seen that $M_{ij}(n)$ is a square integrable martingale satisfying the conditions of Theorem 3.1.1. The corresponding A_n process is given by:

$$A_n = N(i,n)\ a^*_{ij}(1 - a^*_{ij})$$

By the SLLN $\frac{N(i,n)}{n} \rightarrow \pi^*_i$ $a.s.$ and therefore:

$$\frac{A_n}{n} \rightarrow \pi^*_i\ a^*_{ij}(1 - a^*_{ij}) \quad a.s.$$

Since by hypothesis the RHS is strictly positive we have:

$$A_n \rightarrow \infty \quad a.s.$$

We also have that (defining $\beta_{ij} := \pi^*_i\ a^*_{ij}(1 - a^*_{ij})$):

$$\lim_n \frac{A_n \log\log A_n}{n \log\log n} = \beta_{ij} \quad a.s.$$

Substitution into (3.1) gives:

$$\overline{\lim_n} \frac{\mid M_{ij}(n) \mid}{\sqrt{n \log\log n}} = \sqrt{2\beta_{ij}} \quad a.s.$$

Therefore $M_{ij}(n) = O_{a.s.}(\sqrt{n \log\log n})$

Dividing numerator and denominator in (3.2) by n we get:

$$\hat{a}_{ij}(n) - a^*_{ij} = O_{a.s.}(\sqrt{\frac{\log\log n}{n}}) \tag{3.3}$$

Theorem 3.2.1

$$\frac{N(i,n)}{n} - \pi^*_i = O_{a.s.}(\sqrt{\frac{\log\log n}{n}})$$

Proof: First we observe that $N(i,n) - n\ \pi^*_i$ is not a martingale anymore and therefore the previous method for the determination of the rate cannot be applied. The idea we use is the following. Let $N^c(i,j,n)$ and $N^c(i,n)$ denote the counts taken with the circular convention (i.e. considering X_1 as the successor of X_n). From the asymptotic point of view nothing changes since: $N^c(i,j,n) = N(i,j,n) \pm 1$

and $N^c(i,n) = N(i,n) \pm 1$ but now it is easily verified that defining the vector $\hat{\pi}_n$ and the matrix $\hat{A}_n$ via:

$$\begin{aligned}(\hat{\pi}_n)_i &:= \frac{N^c(i,n)}{n} \\ (\hat{A}_n)_{ij} &:= \frac{N^c(i,j,n)}{N^c(i,n)}\end{aligned}$$

we have:

$$\hat{\pi}_n \, \hat{A}_n = \hat{\pi}_n$$

From the previous analysis we know that with $\alpha_n := \frac{\log\log n}{n}$

$$\hat{\pi}_n = \hat{\pi}_n(A^* + O_{a.s.}(\alpha_n)) \quad (component\ wise)$$

Subtract $\pi^* = \pi^* A^*$ from both sides. Then

$$\hat{\pi}_n - \pi^* = (\hat{\pi}_n - \pi^*)A^* + \hat{\pi}_n \, O_{a.s.}\,(\alpha_n)$$

which we can rewrite as:

$$(\hat{\pi}_n - \pi^*)(I - A^*) = O_{a.s.}\,(\alpha_n)$$

But $A^* > 0$ by hypothesis and therefore rank $(I - A^*) = q - 1$ (because the eigenvector π^* of A^* has geometric multiplicity 1). Let Δ be a minor of order $q-1$ and rank $q-1$ of A^* and denote by $(\hat{\pi}_n - \pi^*)_\Delta$ the corresponding $(q-1)$-subvector of $\hat{\pi}_n - \pi^*$. Then $(\hat{\pi}_n - \pi^*)_\Delta \, \Delta = O_{a.s.}\,(\alpha_n)$ and since Δ is invertible $(\hat{\pi}_n - \pi^*)_\Delta = O_{a.s.}\,(\alpha_n)$

Let $\bar{j}$ the index of the component of $\hat{\pi}_n - \pi^*$ not contained in $(\hat{\pi}_n - \pi^*)_\Delta$. Clearly:

$$(\hat{\pi}_n)_{\bar{j}} = 1 - \sum_{j \in \Delta}(\hat{\pi}_n)_j = 1 - \sum_{j \in \Delta}(\pi_j + O_{a.s.}(\alpha_n)) = \pi_{\bar{j}} + O_{a.s.}\,(\alpha_n)$$

and the conclusion is that:

$$\hat{\pi}_n - \pi^* = O_{a.s.}(\sqrt{\frac{\log\log n}{n}}) \quad (component\ wise)$$

This is in perfect agreement with (3.3).

□

3.3 Rates of convergence of the MLE

The results collected here will be essential in the next section where we solve the problem of the estimation of the order of a Markov chain X_t from the observations $\{X_1^n\}$.

Definition 3.3.1
The stationary, finitely valued, process $(X_t,\ t \geq 1)$ *is a finite order Markov chain if for some integer* $m \geq 0$

$$P(X_t = j \mid X_1^{t-1}) = P(X_t = j \mid X_{t-m}^{t-1}) \quad a.e. \quad \forall\, t \geq m+1, \quad \forall\, j \epsilon \mathcal{X}$$

where $\mathcal{X} = \{1, 2, \cdots, q\}$ *is the state space of* X_t.

□

This is the classical definition (see Doob [9] pg 89) and it is somewhat unsatisfactory because it does not uniquely specify what the order is. For the time being we will say that a chain has order less than or equal to m if m is any integer satisfying Definition 3.3.1. A precise definition of the order will be given in Section 3.4 below.

The case $m = 0$ corresponds to an *iid* process, and $m = 1$ to a Markov chain. The probability distribution of a finite order stationary Markov chain of order $\leq m$ is completely specified by the set of transition probabilities (t.p.):

$$a(i^m, j) := P(X_t = j \mid X_{t-m}^{t-1} = i^m), \quad i^m \epsilon \mathcal{X}^m,\ j \epsilon \mathcal{X}$$

and by the initial probabilities (i.p.):

$$v(i^m) := P(X_1^m = i^m), \quad i^m \epsilon \mathcal{X}^m$$

Observe that X_t is stationary only if $v(i^m)$ is an invariant measure.

The probability of the cylinder $\{X_1^n = x_1^n\}$, for $n \geq m+1$, in terms of t.p., and i.p. is:

$$
\begin{aligned}
P(X_1^n = x_1^n) &= P(X_1^m = x_1^m) \prod_{t=m+1}^{n} P(X_t = x_t \mid X_{t-m}^{t-1} = x_{t-m}^{t-1}) \\
&= v(x_1^m) \prod_{t=m+1}^{n} a(x_{t-m}^{t-1}, x_t) \\
&= v(x_1^m) \prod_{i^m, j} a(i^m, j)^{N(i^m, j, n)} \qquad (3.4)
\end{aligned}
$$

where:

$$
N(i^m, j, n) := \sum_{t=m+1}^{n} 1(X_{t-m}^{t-1} = i^m, X_t = j) \qquad (3.5)
$$

Later in this section we will need the SLLN for functionals of X_t. The stationary of X_t allows us to use the *a.e.* version of the ergodic theorem to get the required SLLN, but to keep matters as simple as possible we will assume ergodicity.

Remarks: If we also assume that the observations consist of the initial segment of one trajectory then at no extra cost we may assume that there is only one ergodic class.

Conditions for the ergodicity of X_t can easily be given in terms of the t.p. $a(i^m, j)$, but to express them nicely we need to introduce a new process.

Definition 3.3.2

Let X_t be a stationary process. The m-th derived process Y_t is defined as: $Y_t = (X_t, X_{t+1} \cdots X_{t+m-1}), \quad t \geq 1$

□

If X_t is a finite order Markov chain of order $\leq m$ then it is immediately seen that Y_t is a Markov chain.

The transition probability matrix T of Y_t is of size $q^m \times q^m$ with at most q^{m+1} elements different from zero since obviously:

$$
\begin{aligned}
&t_{i^m, j^m} = 0 \ \textit{unless} \ j_1 = i_2, \cdots, j_{m-1} = i_m \\
&t_{i^m, j^m} = a(i^m, j_m) \ \textit{when} \ j_1 = i_2, \cdots, j_{m-1} = i_m.
\end{aligned}
$$

The following lemma will give a sufficient condition for the ergodicity of X_t in terms of T.

Lemma 3.3.1

Let X_t be any stationary process and Y_t its m-th derived process. If Y_t is ergodic then X_t is ergodic.

Proof: Ergodicity of a stationary finitely valued process Z_t is equivalent to (Walters [24] pg 41): $\forall\ u \geq 1,\ v \geq 1,\ z_1^u,\ z_1^v$

$$\frac{1}{n}\sum_{k=0}^{n-1} P(Z_1^u = z_1^u,\ Z_{k+1}^{k+v} = z_1^v) \to P(Z_1^u = z_1^u)\ P(Z_1^v = z_1^v) \quad a.s. \tag{3.6}$$

Write (3.6) for $Y_t = (X_t, X_{t+1} \cdots X_{t+m-1})$ then we have $\forall\ u' \geq 1,\ \forall\ v' \geq 1$:

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} P(X_1^{u'+m-1} = x_1^{u'+m-1},\ X_{k+1}^{k+v'+m-1} = x_1^{v'+m-1}) \\ = \ & P(X_1^{u'+m-1} = x_1^{u'+m-1})\ P(X_1^{v'+m-1} = x_1^{v'+m-1}) \quad a.s. \end{aligned}$$

This shows that X_t satisfies (3.6) $\forall\ u \geq m,\ v \geq m$. For the case $u < m, v < m$ just observe that:

$$\begin{aligned} &P(X_1^u = x_1^u,\ X_{k+1}^{k+v} = x_1^v) = \\ \sum_{x_{u+1}^m,\ x_{v+1}^m} \ &P(X_1^u = x_1^u,\ X_{u+1}^m = x_{u+1}^m,\ X_{k+1}^{k+v} = x_1^v,\ X_{k+v+1}^{k+m} = x_{v+1}^m) \end{aligned}$$

Condition (3.6) is verified for each term on the RHS and therefore is satisfied also by the LHS. For the other cases, $(u < m,\ v \geq m)$ and $(u \geq m,\ v < m)$ the proof is analogous.

□

Remark: The converse of Lemma 3.3.1 is false.

It is well known that for a (finitely valued) Markov chain Y_t with t.p.m. T the following conditions are equivalent:

i) Y_t is ergodic

ii) T is irreducible

iii) there exists a unique invariant vector t $(t = tT)$

iv) all elements of t are strictly positive

Applying Lemma 3.3.1 we now have:

Corollary

A finite order Markov chain of order $\leq m$ is ergodic if the t.p.m. T of the m-th derived process $Y_t = (X_t, \cdots X_{t+m-1})$ is irreducible

□

Since the initial probabilities of the Y_t process are in one-to-one correspondence with the initial probabilities of the X_t process, we conclude that when T is irreducible there is a unique set of strictly positive initial probabilities $\{\pi(i^m), i^m \epsilon \mathcal{X}^m\}$ corresponding to the t.p. $a(i^m, j)$. The $\pi(i^m)'s$ can be found solving the equation $tT = t$. When T is irreducible a set of parameters that completely specifies the probability distribution of the corresponding X_t chain is:

$$\theta := \{a(i^m, j)\ i^m \epsilon \mathcal{X}^m,\ j = 1, 2, \cdots q-1\}$$

We now study the Maximum Likelihood Estimator (MLE) of θ.

Theorem 3.3.1

Let $X_t \epsilon \mathcal{X} = \{1, 2, \cdots q\}$ be a finite order Markov chain of order $\leq m$ and assume that the m-th derived process Y_t is ergodic.

Let $\theta^* := \{a^*(i^m, j), i^m \epsilon X^m, j \leq q-1\}$ be the true parameter of X_t:

The MLE of θ^* is given by:

$$\hat{a}(i^m, j, n) = \frac{N(i^m, j, n)}{N(i^m, n)}$$

where:

$$\begin{aligned} N(i^m, j, n) &= \sum_{t=m+1}^{n} 1(X_{t-m}^{t-1} = i^m,\ X_t = j) \\ N(i^m, n) &= \sum_{t=m+1}^{n+1} 1(X_{t-m}^{t-1} = i^m) \end{aligned}$$

The MLE converge as to θ_* with rate:

$$\hat{a}(i^m, j, n) = a^*(i^m j) + O_{a.s.}(\sqrt{\frac{\log \log n}{n}})$$

The exact asymptotics for $a^*(i^m j) \neq 0$ are:

$$\overline{\lim}\left(\frac{\log\log n}{n}\right)^{-1/2} \left(\hat{a}(i^m,j,n) - a^*(i^m,j)\right) = \sqrt{\frac{2\ a^*(i^m,j)(1-a^*(i^m,j))}{\pi^*(i^m)}}\ a.s.$$

$$\underline{\lim}\left(\frac{\log\log n}{n}\right)^{-1/2} \left(\hat{a}(i^m,j,n) - a^*(i^m,j)\right) = -\sqrt{\frac{2\ a^*(i^m,j)(1-a^*(i^m,j))}{\pi^*(i^m)}}\ a.s.$$

Proof: The MLE is obtained by direct computation. The asymptotics follow from Theorem 3.1.1. First we deal with the trivial cases. If for some (i^m,j) the t.p. $a^*(i^m,j) = 0$ then $N(i^m,j,n) = 0$ with probability 1 for every n and therefore the rate condition is trivially satisfied.

We now assume $a^*(i^m,j) > 0$.

Observe that:

$$\hat{a}(i^m,j,n) - a^*(i^m,j) = \frac{N(i^m,j,n) - a^*(i^m,j)\ N(i^m,n)}{N(i^m,n)} \tag{3.7}$$

The numerator in (3.7) is a martingale, indexed by n, with respect to $\sigma\{X_1^n\}$. To see this define for given (i^m,j) and $t \geq m+1$:

$$u(t) := 1(X_{t-m}^{t-1} = i^m, X_t = j) - E[1(X_{t-m}^{t-1} = i^m, X_t = j) \mid X_1^{t-1}] \tag{3.8}$$

The process $u(t)$ is centered at the conditional expectation given $\sigma\{X_1^{t-1}\}$ and is therefore automatically a martingale difference. The expectation in (3.8) can be computed explicity:

$$E[1(X_{t-m}^{t-1} = i^m, X_t = j) \mid X_1^{t-1}] = 1(X_{t-m}^{t-1} = i^m)\ a^*(i^m,j)$$

And substituting in (3.8):

$$u(t) = 1(X_{t-m}^{t-1} = i^m,\ X_t = j) - a^*(i^m,j)\ 1(X_{t-m}^{t-1} = i^m)$$

Since $u(t)$ is a mtg difference the process $M(n)$ defined for $n \geq m+1$ by:

$$M(n) := \sum_{t=m+1}^{n} u(t)$$

is a martingale with respect to $\sigma\{X_1^n\}$. $M(n)$ coincides with the numerator of (3.7) thus proving the claim.

$M(n)$ is square integrable because it is bounded, moreover

$$\mid M(n+1) - M(n) \mid = \mid u(n) \mid \leq 2$$

thus verifying the technical conditions of Theorem 3.1.1.

The increasing process A_n is given by (see comment 3.1.2b)

$$A_n = \sum_{t=m+1}^{n} E[u^2(t) \mid X_1^{t-1}]$$

The t-th term is:

$$E[u^2(t) \mid X_1^{t-1}] = a^*(i^m, j)(1 - a^*(i^m, j))\ 1(X_{t-m}^{t-1} = i^m)$$

and therefore:

$$A_n = a^*(i^m, j)(1 - a^*(i^m, j))\ N(i^m, n)$$

Dividing both sides by n:

$$\frac{A_n}{n} = a^*(i^m, j)(1 - a^*(i^m, j))\ \frac{N(i^m, n)}{n}$$

By the SLLN

$$\frac{N(i^m, n)}{n} \to \pi^*(i^m) \quad a.s.$$

Define:

$$\beta(i^m, j) := a^*(i^m, j)(1 - a^*(i^m, j))\ \pi^*(i^m)$$

Under our hypotheses $\beta(i^m, j) > 0$ and therefore:

$$\begin{aligned} \lim_{n\to\infty} A_n &= +\infty \quad a.s. \\ \lim_{n\to\infty} \frac{A_n \log\log A_n}{n \log\log n} &= \beta(i^m, j) \quad a.s. \end{aligned}$$

Theorem 3.1.1 now gives

$$\begin{aligned} \overline{\lim} \frac{M(n)}{\sqrt{n \log\log n}} &= \sqrt{2\beta(i^m, j)} \quad a.s. \\ \underline{\lim} \frac{M(n)}{\sqrt{n \log\log n}} &= -\sqrt{2\beta(i^m, j)} \quad a.s. \end{aligned}$$

From here minor algebraic computations give the exact asymptotics. The rate of convergence follows immediately.

□

The following theorem will play a central role in the next section (for the order estimation problem).

Theorem 3.3.2

Let X_t be as in theorem 3.3.1. Then:

$$\frac{1}{n}\log P_{\hat{\theta}n}(X_1^n) = \frac{1}{n}\log P_{\theta*}(X_1^n) + O_{a.s.}(\frac{\log\log n}{n}) \tag{3.9}$$

Moreover the following bounds on the asymptotics hold:

$$\overline{\lim}(\log\log n)^{-1}\ (\log P_{\hat{\theta}n}(X_1^n) - \log P_{\theta*}(X_1^n)) \leq C_\eta\ q^m\ (q-1) \tag{3.10}$$

$$\underline{\lim}(\log\log n)^{-1}\ (\log P_{\hat{\theta}n}(X_1^n) - \log P_{\theta*}(X_1^n)) \geq 0 \tag{3.11}$$

where $C_\eta := 2\frac{q}{\eta}$ *and* $\eta := \min_{i^m,j}\{a^*(i^m,j)\}$

Proof: Expanding in Taylor's series:

$$\frac{1}{n}\log P_{\hat{\theta}_n} - \frac{1}{n}\log P_{\theta*} = \frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta\ |_{\theta*}\ (\hat{\theta}_n - \theta_*)$$
$$+\frac{1}{2}\ (\hat{\theta}_n - \theta_*)^T\frac{1}{n}\frac{\partial^2 log P_\theta}{\partial\theta^2}\ |_{\theta*}\ (\hat{\theta}_n - \theta_*) + r_n$$

where $r_n = o(\|\ \hat{\theta}_n - \theta_*\ \|^2) = o(\frac{\log\log n}{n})$

The derivative wrt $a(i^m,j)$ is (see (3.4)):

$$\frac{\partial}{\partial a(i^m,j)}\log P_\theta\ |_{\theta*} = \frac{\partial}{\partial a(i^m,j)}\log\pi(X_1^m)\ |_{\theta*} + [\frac{N(i^m,j,n)}{a^*(i^m,j)} - \frac{N(i^m,q,n)}{a^*(i^m,q)}]$$

Asymptotically, after normalization by $\frac{1}{n}$, the first term is negligible. Therefore:

$$\frac{\partial}{\partial a(i^m,j)}\log P_\theta\ |_{\theta*} = \frac{N(i^m,j,n)}{a^*(i^m,j)} - \frac{N(i^m,q,n)}{a^*(i^m,q)}$$

The scalar product of the derivative and $\hat{\theta}_n - \theta_*$ is given by (we drop the dependence from n in the N's):

$$\sum_{i^m, j \le q-1} \frac{N(i^m,j) - a^*(i^m,j)N(i^m)}{N(i^m)}$$
$$\times \left[\frac{N(i^m)}{a^*(i^m,j)} \left(\frac{N(i^m,j) - a^*(i^m,j)N(i^m)}{N(i^m)} \right) - \frac{N(i^m)}{a^*(i^m,q)} \left(\frac{N(i^m,q) - a^*(i^m,q)N(i^m)}{N(i^m)} \right) \right]$$
$$= \sum_{i^m, j \le q-1} \frac{N(i^m)}{a^*(i^m,j)} \left(\frac{N(i^m,j) - a^*(i^m,j)N(i^m)}{N(i^m)} \right)^2$$
$$- \sum_{i^m, j \le q-1} \frac{N(i^m)}{a^*(i^m,q)} \left(\frac{N(i^m,j) - a^*(i^m,j)N(i^m)}{N(i^m)} \right) \left(\frac{N(i^m,q) - a^*(i^m,q)N(i^m)}{N(i^m)} \right)$$

Where the sums must be taken over all $i^m \epsilon \mathcal{X}^m$ and $j \le q-1$. Sum over j the second $\sum$ and add to the first to get:

$$= \sum_{i^m, j} \frac{N(i^m)}{a^*(i^m,j)} \left(\frac{N(i^m,j) - a^*(i^m,j)N(i^m)}{N(i^m)} \right)^2$$

where the sum is now extended over all $(i^m, j) \epsilon \mathcal{X}^{m+1}$.

From Theorem 3.3.1 and the trivial inequality $\overline{\lim}(\sum) \le \sum \overline{\lim}$ we have:

$$\overline{\lim} (\log \log n)^{-1} \frac{\partial}{\partial \theta} \log P_\theta \mid_{\theta *} (\hat{\theta}_n - \theta_*)$$
$$\le \sum_{i^m, j} \overline{\lim} \left(\frac{\log \log n}{n} \right)^{-1} \frac{\hat{\pi}(i^m)}{a^*(i^m,j)} (\hat{a}(i^m,j) - a^*(i^m,j))^2 \qquad (*)$$

Theorem 3.3.1 applies directly to the terms with index $j \le q-1$, but terms with $j = q$ must be dealt with separately. We do the latter first.

$$\hat{a}(i^m,q) - a^*(i^m,q) = - \sum_{j \le q-1} (\hat{a}(i^m,j) - a^*(i^m,j))$$

and therefore:

$$(\hat{a}(i^m,q) - a^*(i^m,q))^2 = \left[\sum_{j \le q-1} (\hat{a}(i^m,j) - a^*(i^m,j)) \right]^2$$
$$\le (q-1) \sum_{j \le q-1} (\hat{a}(i^m,j) - a^*(i^m,j))^2$$

The total contribution to $(*)$ from the terms with $j = q$ is therefore:

$$\overline{\lim} \sum_{i^m} \left(\frac{\log \log n}{n} \right)^{-1} \frac{\hat{\pi}(i^m)}{a^*(i^m,q)} (\hat{a}(i^m,q) - a^*(i^m,q))^2$$

$$\leq (q-1)\sum_{i^m}\sum_{j\leq q-1}\overline{\lim}\left(\frac{\log\log n}{n}\right)^{-1}\frac{\hat{\pi}(i^m)}{a^*(i^m,q)}\,(\hat{a}(i^m,j)-a^*(i^m,j))^2$$

$$\leq (q-1)\sum_{i^m,j\leq q-1}\frac{\pi^*(i^m)}{a^*(i^m,q)}\cdot 2\frac{a^*(i^m,j)(1-a^*(i^m,j))}{\pi^*(i^m)} \tag{3.12}$$

On the other hand the contribution to $(*)$ from the terms with $j \leq q-1$ is upper bounded by (from Theorem 3.3.1 directly):

$$\sum_{i^m,j\leq q-1}\frac{\pi^*(i^m)}{a^*(i^m,j)}\,2\frac{a^*(i^m,j)(1-a^*(i^m,j))}{\pi^*(i^m)} = \sum_{i^m,j\leq q-1}2(1-a^*(i^m,j)) \tag{3.13}$$

Adding together (3.12) and (3.13) we get the final upper bound:

$$2\,(q-1)\;\sum_{i^m,j\leq q-1}\;\frac{a^*(i^m,j)(1-a^*(i^m,j))}{a*(i^m,q)}+2\sum_{i^m,j\leq q-1}(1-a^*(i^m,j))$$

$$\leq\;\sum_{i^m,j\leq q-1}\;(\frac{2(q-1)}{a^*(i^m,q)}+2)$$

With $\eta := \min\{a^*(i^m,q)\}$ and $\tilde{C}_\eta := 2(\frac{q-1}{\eta}+1)$ we get the best possible bound i.e $\tilde{C}_\eta\; q^m(q-1)$. Taking $C_\eta := 2\frac{q}{\eta} \geq \tilde{C}_\eta$ we get the looser but simpler bound given in the statement of the Theorem.

We now proceed to the direct evaluation of the quadratic term in Taylor's expansion. For the sake of readability notation will be kept to the bare essential. In particular we will drop the dependence from m on the first index i.e. $a(i^m,j)$ will be denoted $a_{i,j}$, etc.

We start by observing that:

$$\frac{1}{n}\,\frac{\partial^2}{\partial\theta^2}\log P_\theta\;|_{\theta*} = \;\frac{\partial^2}{\partial\theta^2}E_{\theta*}(\log P_\theta)\;|_{\theta*}+o(1)$$

Where:

$$E_{\theta*}(\log P_\theta)=\sum_i\left(\sum_{j\leq q-1}\pi_i^*\;a_{i,j}^*\log a_{i,j}+\pi_i^*\;a_{i,q}^*\log a_{i,q}\right)$$

The first derivatives are:

$$\frac{\partial}{\partial a_{h,k}}\;E_{\theta*}(\log P_\theta)=a_{h,k}^*\;\pi_h^*\;\frac{1}{a_{h,k}}-a_{h,q}^*\;\pi_h^*\;\frac{1}{a_{h,q}}$$

The second derivatives:

$$\frac{\partial}{\partial a_{h,k}\,\partial a_{e,m}} E_{\theta_*}(\log P_\theta) = -\pi_h^*\, a_{h,k}^* \,\frac{1}{a_{h,k}^2}\,\delta_{(h,k)(e,m)} - \pi_h^*\, a_{h,q}^* \,\frac{1}{a_{h,q}^2}\,\delta_{h,e}$$

The δ's are Kroneker symbols.

$$\frac{\partial}{\partial a_{h,k}\,\partial a_{e,m}} E_{\theta_*}(\log P_\theta)\mid_{\theta_*} = -\pi_h^* \,\frac{1}{a_{h,k}^*}\,\delta_{(h,k)(e,m)} - \pi_h^* \,\frac{1}{a_{h,q}^*}\,\delta_{h,e}$$

The quadratic form is now:

$$\begin{aligned}&(\hat\theta_n - \theta_*)^T \,\frac{\partial^2}{\partial\theta^2}\, E_\theta(\log P_\theta)\mid_{\theta_*} (\hat\theta_n - \theta_*)\\ = &-\sum_{(h,k),k\le q-1}\ \sum_{(e,m),m\le q-1} (\hat a_{hk} - a_{hk}^*)(\hat a_{em} - a_{em}^*)\left[\frac{\pi_h^*}{a_{h,k}^*}\delta_{(h,k)(e,m)} + \frac{\pi_h^*}{a_{hq}^*}\delta_{h,e}\right]\end{aligned}$$

Some cumbersome algebraic manipulations give the final form:

$$= -\sum_{(h,k)} (\hat a_{hk} - a_{hk}^*)^2\,\frac{\pi_h^*}{a_{hk}^*}$$

The quadratic form is negative definite, as was to be expected.

Obviously:

$$\overline{\lim}\ (\log\log n)^{-1}\ (\hat\theta_n - \theta_*)^T \frac{\partial^2 \log P_\theta}{\partial\theta}\mid_{\theta_*} (\hat\theta_n - \theta_*)\ \le 0$$

and therefore this term does not influence the global upper bound. For the lower bound (3.11) observe that by definition:

$$P_{\hat\theta_n}\,(X_1^n) \ge P_{\theta_*}\,(X_1^n)$$

from this (3.11) follows trivially.

□

3.4 Estimation of the order

After the publication of [17] we thought the results of this section would loose some of their interest, but after careful studying we decided to present them for two reasons. We have been unable to convince ourselves of the validity of some of the

arguments given in [17]. Moreover our results do not intersect those given in [17] and are obtained by a totally different method. The problem can be roughly posed as follows. We observe the process X_t which is known to be a finite Markov chain of order m^*, the transition probabilities of X_t and the order m^* are unknown. Our goal: construct a consistent estimator of m^*. To formulate the problem correctly we must first define the order of a finite order chain.

Definition 3.4.1

i) The **order** *of a finite order Markov chain X_t is the minimum m satisfying definition 3.3.1*

ii) A **representation** *of a finite order Markov chain of order m is any set of transition probabilities and stationary initial probabilities:*

$\{a(i^{m'}, j),\ \pi(i^{m'}) \quad (i^{m'}, j) \epsilon X^{m'+1}\}$

with $m' \geq m$ that generate the probability distribution of X_t. m' will be called the **memory** *of the representation.*

iii) A **minimal representation** *is a representation whose memory equals the order.*

Remark: The notions of order and minimality introduced here do not coincide with those of System Theory. Roughly said, in System Theory the order is the cardinality of the smallest state space that allows a description of the process. If X_t is a **stationary** (standard) Markov chain then its system theoretical order would be the cardinality of the set of ergodic states, because no transient state could ever be observed from any trajectory of the stationary chain (the invariant probabilities associated to transient states are all zero). In our definition the state space $\mathcal{X}$ is fixed in advance.

Theorem 3.4.1

Let X_t be an m-th order Markov chain,

$\mathcal{M} := \{a(i^m, j), \pi(i^m)\}$

a minimal representation of X_t, and Y_t the m-th derived process:

i) There is only one minimal representation if and only if Y_t is ergodic.

ii) For all $m' \geq m$ one representation is given by:

$$\begin{aligned} & \mathcal{M}' := \{a'(i^m, j), \pi'(i^{m'})\} \\ \textit{where} \quad & a'(i^{m'}, j) = a(i_{m'-m+1}^{m'}, j) \\ & \pi'(i^{m'}) = \pi(i^m)\ a(i^m, i_{m+1}) \cdots a(i_{m'-m}^{m'-1}, i_{m'}) \end{aligned}$$

iii) The m'-th derived process Y_t is ergodic for all $m' > m$ if and only if $a(i^m, j) > 0$ for all (i^m, j).

iv) For any $m' > m$ there is only one representation with memory m' if and only if $a(i^m, j) > 0$ for all (i^m, j).

Proof: i) If Y_t is ergodic then from the general fact that two distinct stationary ergodic processes are mutually singular we conclude that $\mathcal{M}$ is unique. On the other hand if Y_t is not ergodic at least one of the stationary probabilities $\pi(i^m) = 0$ and therefore the corresponding "row" of t.p.'s $a(i^m, j)$ can be changed without altering the probability distribution and we may therefore construct infinitely many representations equivalent to $\mathcal{M}$.

ii) For $m' > m$ we have:

$$\begin{aligned} a'(i^{m'}, j) &:= P(X_{m'+1} = j \mid X_1^{m'} = i^{m'}) \\ &= P(X_{m'+1} = j \mid X_{m'-m+1}^{m'} = i_{m'-m+1}^{m'}) = a(i_{m'-m+1}^{m'}, j) \\ \pi'(i^{m'}) &:= P(X_1^{m'} = i^{m'}) = P(X_{m+1}^{m'} = i_{m+1}^{m'} \mid X_1^m = i^m)\ P(X_1^m = i^m) \\ &= \pi(i^m)\ a(i^m, i_{m+1})\ a(i_2^{m+1}, i_{m+2}) \cdots a(i_{m'-m}^{m'-1}, i_{m'}) \end{aligned}$$

It follows from the definition that these are the t.p.'s associated to the m'-th derived process $Y_t = (X_t \cdots X_{t+m-1})$ when X_t is generated by $\mathcal{M}$.

iii) If $a(i^m, j) > 0\ \forall i^m, j$ then for all $m' \geq m$ the chain can move between any two sequences of states, $X_1^{m'}$ and $\bar{X}_1^{m'}$, in at most $m' + 1$ steps. This is more than required for the ergodicity of the m'-th derived process for all $m' \geq m$.

On the other hand if $a(i^m, j) = 0$ for some (i^m, j) then the t.p. matrix of the $(m+1)$-th derived process has a "column" of zeros and is therefore not irreducible. It follows that the $(m+1)$-th derived process is not ergodic.

iv) If $a(i^m, j) > 0\ \forall i^m, j$ then for any $m' \geq m$ the m'-th derived process Y_t is an ergodic Markov chain of order 1. By i) Y_t has a unique minimal representation

which must therefore coincide with $\mathcal{M}'$ constructed in ii). If on the other hand $a(i^m, j) = 0$ for some (i^m, j) then the $(m+1)$-th derived process is non ergodic as proved in iii) and therefore its representation of memory $m+1$ is non-unique again by i).

□

For the proof of the consistency of the order estimation procedure we will need the ergodicity of the m'-th derived processes for all $m' \geq m$ so that the SLLN will be valid for all $m' \geq m$. This fact, in view of Theorem 3.4.1, justifies the following assumption:

Assumption SP:
The observed process is a finite Markov chain, taking value in $\mathcal{X} = \{1, 2 \cdots, q\}$, of unknown order m^*, and unknown **strictly positive** transition probabilities $\{a^*(i^{m^*}, j)\}$.

And now we can formulate our:

Problem:
Let X_t be a process satisfying assumption SP. From the observation of an arbitrarily large initial segment of one trajectory of X_t construct a strongly consistent estimator of the unknown order m^*.

The most natural parametric model for the process X_t is given by the following:

Definition 3.4.2

$$\Theta_m := \{\textit{all possible } a(i^m, j) > 0 \quad i^m \epsilon \mathcal{X}^m, j = 1, 2, \cdots q-1\}$$
$$\Theta := \bigcup_{m \geq 0} \Theta_m$$

The results of Theorem 3.4.1 now tell us that Θ_m contains no representation of X_t if $m < m^*$ and only one representation of X_t for any $m \geq m^*$.

Definition 3.4.3

The compensated maximum log*-likelihood is defined as:*

$$C(m, n) := -L_n(\hat{\theta}_m(n)) + \delta_n(m)$$

Where:

$\hat{\theta}_m(n)$ *is the ML estimator of* $\theta \epsilon \Theta_m$ *based on* n *observations*

$L_n(\hat{\theta}_m(n)) := \frac{1}{n} \log P_{\hat{\theta}_m(n)}(X_1^n)$

$\delta_n(m)$ *is a positive, increasing function of* m *to be specified.*

Definition 3.4.4

$$\hat{m}(n) := \min\{\arg\min_{m \geq 0} C(m,n)\}$$

We will show how to choose the functions $\delta_n(m)$ to guarantee the strong consistency of $\hat{m}(n)$.

The idea of studying the compensated likelihood and the technique for the choice of the $\delta_n(m)$ functions follow Azencott, Dacunha-Castelle [1]. Obviously our main result (Theorem 3.3.2) and its technique of proof are totally different.

Theorem 3.4.2 *Compensators Avoiding Underestimation*

If $\lim_n \delta_n(m) = 0 \quad \forall m$ *then:*

$\underline{\lim}\, \hat{m}(n) \geq m^* \quad P_{\theta^*} - a.s.$

Proof: Define $L_n(\theta) := \frac{1}{n} \log P_\theta(X_1^n)$. From Lemma 2.4.1 we have:

$L_n(\theta^*) - L_n(\theta) \to D(P_{\theta^*} \parallel P_\theta)$ a.s. and uniformly in $\theta \epsilon \Theta_m \ \forall m$.

Therefore for $m < m^*$:

$$\inf_{\theta \epsilon \Theta_m} [L_n(\theta^*) - L_n(\theta)] = L_n(\theta^*) - L_n(\hat{\theta}_m(n)) \to \inf_{\theta \epsilon \Theta_m} D(P_{\theta^*} \parallel P_\theta) := \gamma > 0 \quad (3.14)$$

($\gamma > 0$ since no point in Θ_m is equivalent to X_t for $m < m^*$).

Theorem 3.3.1 shows that $\hat{\theta}_{m^*}(n) \to \theta^*$ a.s. which implies:

$$\lim_n [L_n(\theta^*) - L_n(\hat{\theta}_{m^*}(n))] = 0 \quad a.s. \qquad (3.15)$$

From (3.14) and (3.15)

$$\lim_n [L_n(\hat{\theta}_{m^*}(n)) - L_n(\hat{\theta}_m(n))] = \gamma \ > \ 0 \qquad (3.16)$$

By the definition of $C(m,n)$ and the condition $\lim_n \delta_n(m) = 0$, (3.16) gives:

$$\lim_n [C(m,n) - C(m^*,n)] = \gamma > 0$$

for any $m < m^*$. On the other hand $[C(\hat{m}(n), n) - C(m^*, n)] \leq 0$ by definition of $\hat{m}(n)$ and therefore we conclude that all the limiting values of $\hat{m}(n)$ must be $\geq m^*$. i.e. $\underline{\lim}\, \hat{m}(n) \geq m^*$ as claimed.

□

We now study the conditions to be imposed on the function $\delta_n(m)$ to avoid overestimation. Theorem 3.3.2 will be our main tool.

Lemma 3.4.1

For any $m > m^*$ we have:

$$\overline{\lim}(\frac{\log\log n}{n})^{-1}\ [L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n))] \leq C_\eta q^m (q-1)$$

Proof:

$$\begin{aligned}
& \overline{\lim}(\frac{\log\log n}{n})^{-1}\ [L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n))] \\
\leq\ & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ [L_n(\hat{\theta}_m(n)) - L_n(\theta_*)] \\
-\ & \underline{\lim}(\frac{\log\log n}{n})^{-1}\ [L_n(\hat{\theta}_{m^*}(n)) - L_n(\theta_*))] \\
\leq\ & C_\eta\ q^m\ (q-1)
\end{aligned}$$

To bound $\overline{\lim}$ and $\underline{\lim}$ we used theorem 3.3.2 which is valid for any $m > m^*$

□

Theorem 3.4.3 *Compensators Avoiding Overestimation*

If the compensator is of the form

$\delta_n(m) := \varphi(n)\ h(m)$

where the function φ satisfies:

$\underline{\lim}(\frac{\log\log n}{n})^{-1}\ \varphi(n) > 1$

and the function h satifies:

$h(m') - h(m) \geq C_\eta\ q^{m'}\ (q-1)$ *for all* $m' > m \geq 0$.

Then:

$\overline{\lim}\ \hat{m}(n) \leq m^*\ P_{\theta_*}$ *a.s.*

Proof: We now assume $m > m^*$, and from the form of $\delta_n(m)$ we have:

$$C(m^*,n) - C(m,n) = L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n)) + \varphi(n)\ [h(m^*) - h(m)]$$

Therefore:

$$\begin{aligned} & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ (C(m^*,n) - C(m,n)) \\ \leq\ & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ [L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n))] \\ +\ & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ \varphi(n)[h(m^*) - h(m)] \end{aligned}$$

The first term is bounded by (Lemma 3.4.1) $C_\eta\ q^m\ (q-1)$.

For the second term observe that by hypothesis:

$$h(m^*) - h(m) \leq -C_\eta\ q^m\ (q-1) \quad (since\ m > m^*)$$

Substitution gives:

$$\begin{aligned} & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ \varphi(n)\ [h(m^*) - h(m)] \\ \leq\ & \overline{\lim}(\frac{\log\log n}{n})^{-1}\ \varphi(n)(-C_\eta\ q^m\ (q-1)) \end{aligned}$$

And therefore:

$$\begin{aligned} & \overline{\lim}\left(\frac{\log\log n}{n}\right)^{-1}\ (C(m^*,n) - C(m,n)) \\ \leq\ & C_\eta\ q^m\ (q-1)\ [1 - \underline{\lim}\left(\frac{\log\log n}{n}\right)^{-1}\ \varphi(n)] \end{aligned}$$

The hypothesis on φ makes the bracket strictly negative. We conclude that $\forall\ m > m^*$:

$$\overline{\lim}(\frac{\log\log n}{n})^{-1}(C(m^*,n) - C(m,n)) < 0$$

On the other hand, by definition of $\hat{m}(n)$, $C(m^*,n) - C(\hat{m}(n),n) \geq 0$ thus the conclusion $\overline{\lim}\ \hat{m}(n) \leq m^*$.

□

We must now show that functions h and φ satisfying the conditions imposed by Theorems 3.4.2 and 3.4.3 do exist. The h function is substantially different from the h function of [1].

Example of h Function

$$h(m) := C_\eta(q-1)\ \frac{(q+1)^{m+1}}{q}$$

We must check that for all $0 \leq m < m'$

$$h(m') - h(m) \geq C_\eta\ (q-1)\ q^{m'}$$

But:

$$h(m') - h(m) = C_\eta\ (q-1) \left[\frac{(q+1)^{m'+1}}{q} - \frac{(q+1)^{m+1}}{q} \right]$$

The condition is satisfied if:

$$\frac{(q+1)^{m'+1}}{q} - \frac{(q+1)^{m+1}}{q} \geq q^{m'}$$

for all $0 \leq m < m'$

Dividing both sides by $q^{m'}$ we get:

$$(\frac{q+1}{q})^{m'+1} - (\frac{q+1}{q})^{m'+1} \frac{1}{(q+1)^{m'-m}} \geq 1$$

$$(\frac{q+1}{q})^{m'+1}[1 - \frac{1}{(q+1)^{m'-m}}] \geq 1$$

Since $m' > m$ the term in brackets is greater than $(1 - \frac{1}{q+1})$. The inequality is therefore satisfied if:

$$(\frac{q+1}{q})^{m'+1}\ (\frac{q}{q+1}) \geq 1$$

i.e.

$$(\frac{q+1}{q})^{m'} \geq 1 \quad \forall\, m' \geq 0$$

which is trivially satisfied.

□

Example of φ Function

$\varphi(n)$ satisfies both Theorems 3.4.2 and 3.4.3 if it is taken as:

$$\varphi(n) = \frac{\log\log n}{n}(1+\epsilon) \quad for\ some\ \epsilon > 0$$

□

The reader that patiently followed us may object that Theorem 3.4.3 is useless from a practical point of view, since C_η depends on the true distribution. Theorem 3.4.4 will reassure him or her on the practicability of our approach, but first let us observe that in the ARMA case the h function does not depend on the true distribution (see [1]). We are investigating the reasons of this discrepancy.

Theorem 3.4.4 *Consistent Estimators*

The compensator:

$$\delta_n(m) := h(m)\ \frac{\log n}{n}$$

where $h(m)$ is any strictly increasing function of m, produces a strongly consistent estimator of m^*.

Proof: $\delta_n(m) \to 0$ for all m and therefore it avoids underparametrization as proved in Theorem 3.4.2.

For the case of overparametrization we reason as in the proof of Theorem 3.4.3. Assume $m > m^*$.

$$\begin{aligned} &\overline{\lim}(\frac{\log n}{n})^{-1}\ [C(m^*,n) - C(m,n)] \\ \leq\ &\overline{\lim}(\frac{\log n}{n})^{-1}\ [L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n))] + (h(m^*) - h(m)) \end{aligned}$$

Since the difference $L_n(\hat{\theta}_m(n)) - L_n(\hat{\theta}_{m^*}(n)) = O_{a.s.}(\frac{\log\log n}{n})$, and by hypothesis $h(m^*) - h(m) < 0$, we get for $m > m^*$

$$\overline{\lim}(\frac{\log n}{n})^{-1}\ [C(m^*,n) - C(m,n)] \leq h(m^*) - h(m) < 0.$$

On the other hand $C(m^*,n) - C(\hat{m}(n),n) \geq 0$, by definition of $\hat{m}(n)$, and therefore $\overline{\lim}\ \hat{m}(n) \leq m^*$ i.e. $\hat{m}$ avoids overparametrization too.

□

Chapter 4

Estimation of the Order of a Hidden Markov Chain

The technique that was employed in Chapter 3 for the estimation of the order of a Markov chain will now be adapted to the estimation of the order of a HMC. As we have seen in the Markov case, the crucial step is the evaluation of the rate of growth of the maximized likelihood ratio (MLR). For Markov chains we evaluated this rate to be $O_{a.s.}(\log\log n)$ (Theorem 3.3.2) and we also had very precise results for the $\overline{\lim}$ and the $\underline{\lim}$ of the MLR. For HMC's we will be able to get the rate $O_{a.s.}(\log\log n)$ only in special cases. For the general case we get $O_{a.s.}(\log n)$.

At first the problem of estimating the rate of the MLR for HMC seems easy to solve. For any y_1^n write: $P_\theta(y_1^n) = \sum_{x_1^n} P_\theta(y_1^n, x_1^n) = \sum_{x_1^n} P_\theta(s_1^n)$ where the process $S_t = (X_t, Y_t)$ is a Markov chain.

Clearly $\max_\theta P_\theta(y_1^n) \leq \sum_{x_1^n} \max_\theta P_\theta(s_1^n)$.

Since S_t is a Markov chain we know from Theorem 3.3.2 that:

$$\frac{\max_\theta P_\theta(s_1^n)}{P_{\theta_0}(s_1^n)} = e^{\alpha_n}$$

where $\alpha_n = O_{a.s.}(\log\log n)$

Substituting in the previous inequality we find:

$$\max_\theta P_\theta(y_1^n) \leq \sum_{x_1^n} e^{\alpha_n} P_{\theta_0}(s_1^n) = e^{\alpha_n} \sum_{x_1^n} P_{\theta_0}(s_1^n) = e^{\alpha_n} P_{\theta_0}(y_1^n)$$

From this we immediately get the desired rate:

$$\log \frac{\max_\theta P_\theta(y_1^n)}{P_{\theta_0}(y_1^n)} = O_{a.s.}(\log\log n)$$

This idea, or variations of it, has appeared in the literature, but unfortunately it is wrong. The problem is that Theorem 3.3.2 does *not* say that $\alpha_n = O_{a.s.}(\log\log n)$ *uniformly with respect to the realization* ω.

In Section 4.1 we pose the problem of the order estimation for HMC's and prove the analog of Theorem 3.4.2 on estimators that avoid underestimation. In Section 4.2 the MLR is studied for the true order. In this case we get the $O_{a.s.}(\log\log n)$ rate of growth. In Section 4.3 we approximate the MLR using Markov chains of finite memory and get a rather weak bound on the rate of growth. This result is more of theoretical than practical interest because the bound depends strongly from the true distribution. In Section 4.4 we state a result from Information Theory and use it to find the $O_{a.s.}(\log n)$ bound on the rate of the MLR. The final Section 4.5 is dedicated to the construction of strongly consistent estimators of the order.

4.1 Preliminaries

In Section 1.2 we defined the order of a HMC Y_t as the minimum integer q for which there exists a representation of Y_t with $\mid \mathcal{X} \mid = q$. In analogy with Section 3.4 we would like to construct a consistent estimator of the order based on the compensated maximum likelihood. The HMC case is complicated by the fact that our knowledge of the set of equivalent representations is only partial (see Sections 1.3 and 1.4). To cope with this difficulty we have to impose restrictions on the observed process Y_t thus limiting the applicability of the results. Fortunately all of the assumptions are satisfied by a generic HMC and therefore the results are still widely applicable.

Assumption SP':
The observed process Y_t is a HMC taking values in $\{1, 2, \cdots r\}$, of unknown order q_0. One representation of Y_t is given by $\theta_0 = \{q_0, A_0, B_0\}$ where θ_0 is a Petrie point of $\Theta^{\delta}_{q_0}$ for some $\delta > 0$.

(Petrie's points are defined in 1.4.1 and $\Theta^{\delta}_{q_0}$ in 1.4.2).

The class of parametric models that will be used is

$$\Theta := \cup_{q \geq 1} \Theta^{\delta}_{q}.$$

The results of Sections 1.2, 1.3 and 1.4 guarantee that Θ_q^δ contains no point equivalent to θ_0 if $q < q_0$ and a finite number of points equivalent to θ_0 if $q = q_0$. For $q > q_0$ there are infinitely many points in Θ_q^δ equivalent to θ_0, as can easily be seen applying Lemma 1.3.1. In analogy with Section 3.4 the compensated maximum log-likelihood is defined as:

$$C(q,n) := -L_n(\hat{\theta}_q(n)) + \delta_n(q)$$

where:

$\hat{\theta}_q(n)$ is the MLE of $\theta \epsilon \Theta_q^\delta$ based on n observations

$L_n(\hat{\theta}_q(n)) := \frac{1}{n} \log P_{\hat{\theta}_q(n)}(Y_1^n)$

$\delta_n(q)$ is a positive increasing function of q and n to be determined.

The estimator of the order is defined by:

$$\hat{q}(n) := \min\{\arg\min_{q \geq 1} C(q,n)\}$$

The problem of order estimation can now be posed as follows.

Problem:
The HMC Y_t satisfying assumption SP' is observed. Find a compensator sequence $\delta_n(q)$ such that the estimator $\hat{q}(n)$ is strongly consistent i.e. $\hat{q} \rightarrow q_0$ *a.s.* P_{θ_0}.

The analog of Theorem 3.4.2 is valid and we can easily give a sufficient condition on $\delta_n(q)$ that avoids underestimation.

Theorem 4.1.1 *Compensators avoiding underestimation*

Let Y_t be a process satisfying conditions SP'.

If $\lim_{n\to\infty} \delta_n(q) = 0 \quad (\forall\, q)$

Then $\underline{\lim}_{n\to\infty} \hat{q}(n) \geq q_0 \quad P_{\theta_0} - a.s.$

Proof: The proof is completely analogous to the proof of Theorem 3.4.2 and based on the essential fact that for $q < q_0$ there is no point in Θ_q^δ equivalent to θ_0. This last fact follows easily as a consequence of Lemma 1.3.2. Since $\theta_0 = (q_0, A, B)$ is regular any $\theta = (q, A, B)$ equivalent to it must have $q \geq q_0$.

□

Sufficient conditions on $\delta_n(q)$ to avoid overestimation are much more difficult to find. The crucial problem is to determine the *a.s.* rate of growth of:

$$\log \frac{P_{\hat{\theta}_q(n)}(Y_1^n)}{P_{\theta_0}(Y_1^n)}$$

In Section 3.4 we used the LIL to study the *a.s.* asymptotic behaviour of the ratio but, as it will become apparent in the next section, for HMC's this technique works only for $q = q_0$. Theorem 4.5.1 will give sufficient conditions on $\delta_n(q)$ to avoid overestimation.

4.2 Rate of Convergence in Θ_{q0}^{δ}

We study here the rate of growth of the maximized log-likelihood ratio (MLR)

$$\log \frac{P_{\hat{\theta}_{q_0}(n)}(y_1^n)}{P_{\theta_0}(y_1^n)}$$

Since q_0 is fixed, in this Section $\hat{\theta}_{q_0}(n)$ will be denoted $\hat{\theta}_n$. We need one extra assumption on the HMC Y_t which will be in force through this section.

Assumption PH:

$$-\frac{\partial^2}{\partial\theta^2} H_{\theta_0}(\theta) \mid_{\theta_0} > 0$$

Recall that: $H_{\theta_0}(\theta) := E_{\theta_0}[\log P_\theta(Y_0 \mid Y_{-\infty}^{-1})]$.

After giving two preliminary results we will prove that the MLR is $0_{as}(\log\log n)$. Remember that (Section 2.2):

$$\hat{\theta}_n = \{\hat{\theta} \epsilon \Theta_{q_0}^{\delta} \; ; \; P_{\hat{\theta}}(y_1^n) = \max_{\theta} P_\theta(y_1^n)\}$$

and that in general $\hat{\theta}_n$ is not a singleton. Our first result shows that it is always possible to choose a convergent sequence $\hat{\theta}_n^{\cdot} \epsilon \hat{\theta}_n$.

Lemma 4.2.1 *There exists $N > 0$ finite and a sequence $\hat{\theta}_n^{\cdot} \epsilon \hat{\theta}_n \quad (\forall\, n \geq N)$ such that $\hat{\theta}_n^{\cdot} \to \theta_0$*

Proof: From Theorem 2.2.1 we know that $\hat{\theta}_n \to \mathcal{N} = \{\theta \epsilon \Theta_{q_0}^{\delta} \; ; \; H_{\theta_0}(\theta) = H_{\theta_0}(\theta_0)\}$
a.s. P_{θ_0}. From assumption SP' and Theorem 1.4.1 it follows that $\mathcal{N}$ is a finite subset of $\Theta_{q_0}^{\delta}$. In particular $\mathcal{N} = \{\theta \epsilon \Theta_{q_0}^{\delta} \; ; \; \theta = \sigma(\theta_0)\}$ where $\sigma(\theta_0)$ denotes the permutation of the matrices (A_0, B_0) induced by the permutation σ of the state space $\mathcal{X}$. As a consequence of these two facts we have that for any $\varepsilon > 0$ there exists an integer N_ε such that $\hat{\theta}_n \subset \mathcal{N}_\varepsilon$ for all $n \geq N_\varepsilon$. (Here $\mathcal{N}_\varepsilon$ denotes the fattened set). Since $\mathcal{N}$ is finite its points are isolated. Choose ε small enough for $\mathcal{N}_\varepsilon$ to be the union of disjoint balls of radius ε centered at the points of $\mathcal{N}$ i.e. $\mathcal{N}_\varepsilon = \cup_\sigma \mathcal{B}(\sigma(\theta_0))$. Let $N = N_\varepsilon$.

To complete the proof it is enough to show that for all $n \geq N$ there exists a point $\hat{\theta}_{n^{\cdot}} \epsilon \mathcal{B}_\varepsilon(\theta_0)$ such that $P_{\hat{\theta}_n^{\cdot}}(Y_1^n) = P_{\hat{\theta}_n}(Y_1^n)$. Let $\tilde{\theta} \epsilon \hat{\theta}_n$ and $n \geq N$. Then for some σ we have $\tilde{\theta} \epsilon \mathcal{B}_\varepsilon(\sigma(\theta_0))$ and from the identity $\sigma(\mathcal{B}_\varepsilon(\theta)) = \mathcal{B}_\varepsilon(\sigma(\theta))$ we conclude that $\sigma^{-1}(\tilde{\theta}) \epsilon \mathcal{B}_\varepsilon(\theta_0)$. We can define $\hat{\theta}_n^{\cdot} := \sigma^{-1}(\tilde{\theta})$. From $\hat{\theta}_n \to \mathcal{N}$ we now conclude that $\hat{\theta}_n^{\cdot} \to \theta_0$.

□

From now on we will suppose that the choice of a sequence of points $\hat{\theta}_n^{\cdot} \to \theta_0$ has already been made and with slight abuse of notation we will denote by $\hat{\theta}_n$ the point $\hat{\theta}_n^{\cdot}$ itself.

The lemma below will be needed for the application of the LIL.

Lemma 4.2.2 *For some finite C, $\forall k$, $\forall l$, $\forall \theta$:*

$$\left| \frac{\partial}{\partial \theta_l} \log P_\theta(y_k \mid y_1^{k-1}) \right| \leq C \quad a.s.\ P_{\theta_0}$$

Proof: In the statement θ_l can be any element of $\theta = (A, B)$. First we prove the case $\theta_l := a_{ij}$ from some (i, j) with $1 \leq i \leq q_0$ and $1 \leq j \leq q_0 - 1$. A direct

computation of the derivative gives:

$$\begin{aligned}
&\frac{\partial}{\partial a_{ij}} \log P_\theta(y_k \mid y_1^{k-1}) = \frac{\partial}{\partial a_{ij}} \log P_\theta(y_1^k) - \frac{\partial}{\partial a_{ij}} \log P_\theta(y_1^{k-1}) \\
= &\frac{1}{P_\theta(y_1^k)} \frac{\partial}{\partial a_{ij}} \sum_{x_1^k} P_\theta(y_1^k, x_1^k) - \frac{1}{P_\theta(y_1^{k-1})} \frac{\partial}{\partial a_{ij}} \sum_{x_1^k} P_\theta(y_1^{k-1}, x_1^{k-1}) \\
= &\sum_{x_1^k} \left(\frac{N_{ij}(x_1^k)}{a_{ij}} - \frac{N_{iq_0}(x_1^k)}{a_{iq_0}} \right) P_\theta(x_1^k \mid y_1^k) \\
- &\sum_{x_1^{k-1}} \left(\frac{N_{ij}(x_1^{k-1})}{a_{ij}} - \frac{N_{iq_0}(x_1^{k-1})}{a_{iq_0}} \right) P_\theta(x_1^{k-1} \mid y_1^{k-1}) \\
= &\sum_{t=1}^{k-1} \left(\frac{P_\theta(x_t = i, x_{t+1} = j \mid y_1^k)}{a_{ij}} - \frac{P_\theta(x_t = i, x_{t+1} = q_0 \mid y_1^k)}{a_{iq_0}} \right) \\
- &\sum_{t=1}^{k-2} \left(\frac{P_\theta(x_t = i, x_{t+1} = j \mid y_1^{k-1})}{a_{ij}} - \frac{P_\theta(x_t = i, x_{t+1} = q_0 \mid y_1^{k-1})}{a_{iq_0}} \right) \\
= &\frac{1}{a_{ij}} \sum_{t=1}^{k-2} P_\theta(x_t = i, x_{t+1} = j \mid y_1^k) - P_\theta(x_t = i, x_{t+1} = j \mid y_1^{k-1}) \\
- &\frac{1}{a_{iq_0}} \sum_{t=1}^{k-2} P_\theta(x_t = i, x_{t+1} = q_0 \mid y_1^k) - P_\theta(x_t = i, x_{t+1} = q_0 \mid y_1^{k-1}) \\
+ &\frac{P_\theta(x_{k-1} = i, x_k = j \mid y_1^k)}{a_{ij}} - \frac{P_\theta(x_{k-1} = i, x_k = q_0 \mid y_1^k)}{a_{iq_0}}
\end{aligned}$$

From Lemma A.0.3 we get form some $\rho < 1$:
$\mid P_\theta(x_t = i_1\ x_{t+1} = i_2 \mid y_1^k) - P_\theta(x_t = i_1\ x_{t+1} = i_2 \mid y_1^{k-1}) \mid \leq \rho^{k-t-1}$.
From the triangle inequality and the bounds $\delta \leq a_{ij} \leq 1 - \delta$ we have:

$$\begin{aligned}
\mid \frac{\partial}{\partial a_{ij}} \log P_\theta(y_k \mid y_1^{k-1}) \mid \ &\leq \ \frac{1}{a_{ij}} \sum_{t=1}^{k-2} \rho^{k-t-1} + \frac{1}{a_{iq_0}} \sum_{t=1}^{k-2} \rho^{k-t-1} + \frac{2}{\delta} \\
&\leq \ \frac{2}{\delta} \left(\frac{1}{1-\rho} + 1 \right) := C.
\end{aligned}$$

The proof for $\theta_l = b_{jy}$ for $1 \leq j \leq q_0$, $1 \leq y \leq r-1$ is very similar and will be omitted.

□

We are now ready to study the rate of convergence.

Theorem 4.2.1

$$\frac{1}{n}\log P_{\hat{\theta}_n}(y_1^n) = \frac{1}{n}\log P_{\theta_0}(y_1^n) + O_{a.s.}\left(\frac{\log\log n}{n}\right)$$

Proof: The proof is divided into three steps. In step 1 we use the LIL for martingales to prove that $\frac{\partial}{\partial\theta}\log P_\theta(y_1^n) \mid_{\theta_0} = O_{a.s.}(\sqrt{n\log\log n})$. In step 2 we use a Taylor's expansion and step 1 to prove that $\hat{\theta}_n - \theta_0 = O_{a.s.}(\sqrt{\frac{\log\log n}{n}})$. Finally in step 3 we use another Taylor's expansion and steps 1, 2 to conclude.

Step 1:

We first prove that $\frac{\partial}{\partial\theta}\log P_\theta(y_1^n) \mid_{\theta_0} = O_{a.s.}(\sqrt{n\log\log n})$ componentwise. For the generic component θ_l we have:

$$\frac{\partial}{\partial\theta_l}\log P_\theta(y_1^n) \mid_{\theta_0} = \sum_{k=1}^{n} u_k$$

where

$$u_k := \frac{\partial}{\partial\theta_l}\log P_\theta(y_k \mid y_1^{k-1}) \mid_{\theta_0} .$$

Computing the conditional expectation:

$$E_{\theta_0}(u_k \mid y_1^{k-1}) = \sum_y P_{\theta_0}(Y_k = y \mid y_1^{k-1})\frac{\partial}{\partial\theta_l}\log P_\theta(y_k = y \mid y_1^{k-1}) \mid_{\theta_0} = 0.$$

We see that $\{\frac{\partial}{\partial\theta_l}\log P_\theta(y_1^n) \mid_{\theta_0}, \sigma(y_1^n)\}$ is a martingale.

Since Y_t takes values in a finite set it follows trivially that the martingale is square integrable. Lemma 4.2.2 guarantees that $\mid u_k \mid \leq C$ *a.s.* P_{θ_0} for some constant C and therefore we can apply the LIL for martingales as given in Theorem 3.1.1. In particular

$$\overline{\lim}_{n\to\infty}\frac{\mid \frac{\partial}{\partial\theta_l}\log P_\theta(y_1^n) \mid_{\theta_0}\mid}{\sqrt{2\ A_n\log\log A_n}} = 1 \quad a.s.\ on\ [A_n \to \infty].$$

The definition of A_n gives (see comment 3.1.2b)

$$A_n := \sum_{k=1}^{n} E_{\theta_0}(u_k^2 \mid y_1^{k-1})$$

Observe that:

$$E_{\theta_0}(u_k^2 \mid y_1^{k-1}) = -E_{\theta_0}(\frac{\partial^2}{\partial\theta_l^2}\log P_\theta(y_k \mid y_1^{k-1}) \mid_{\theta_0} \mid y_1^{k-1})$$

can be proved by direct computation.

To complete the first part of the proof we show that:

$$\lim_{n\to\infty} \frac{A_n}{n} = \beta^2$$

where $\beta^2 := -\frac{\partial^2}{\partial\theta_l^2} H_{\theta_0}(\theta) \mid_{\theta_0} > 0$ by hypothesis PH.

This is a consequence of the ergodic theorem and of the following bound. (C is a finite constant and $0 < \rho < 1$):

$$\left| E_{\theta_0}\left(\frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y_k \mid y_{-\infty}^{k-1}) \mid_{\theta_0} \mid y_{-\infty}^{k-1}\right) - E_{\theta_0}(u_k^2 \mid y_1^{k-1}) \right| \leq C\ \rho^k$$

The bound is proved as follows:

$$\begin{aligned}
&\left| E_{\theta_0}\left(\frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y_k \mid y_{-\infty}^{k-1}) \mid_{\theta_0} \mid y_{-\infty}^{k-1}\right) - E_{\theta_0}(u_k^2 \mid y_1^{k-1}) \right| \\
= &\left| \sum_y P_{\theta_0}(y \mid y_{-\infty}^{k-1}) \frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y \mid y_{-\infty}^{k-1}) \mid_{\theta_0} - P_{\theta_0}(y \mid y_1^{k-1}) \frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y \mid y_1^{k-1}) \mid_{\theta_0} \right| \\
\leq &\sum_y P_{\theta_0}(y \mid y_{-\infty}^{k-1}) \left| \frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y \mid y_{-\infty}^{k-1}) \mid_{\theta_0} - \frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y \mid y_1^{k-1}) \mid_{\theta_0} \right| \\
+ &\sum_y \left| \frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y \mid y_{-\infty}^{k-1}) \mid_{\theta_0} \right| \left| P_{\theta_0}(y \mid y_{-\infty}^{k-1}) - P_{\theta_0}(y \mid y_1^{k-1}) \right|
\end{aligned}$$

From Lemma 4.1 of Baum-Petrie [4] we get the desired bound. For $\frac{A_n}{n}$ we therefore have:

$$\left| \frac{A_n}{n} - \frac{1}{n} \sum_{k=1}^{n} E_{\theta_0}\left(-\frac{\partial^2}{\partial\theta_l^2} \log P_\theta(y_k \mid y_{-\infty}^{k-1}) \mid_{\theta_0} \mid y_{-\infty}^{k-1} \right) \right| \leq \frac{1}{n} \sum_{k=1}^{n} \rho^k$$

taking the limit for $n \to \infty$ and applying the ergodic theorem we conclude that $\frac{A_n}{n} \to \beta^2$ and the result of Theorem 3.1.1 becomes

$$\overline{\lim}_{n\to\infty} \frac{\mid \frac{\partial}{\partial\theta_l} \log P_\theta(y_1^n) \mid_{\theta_0} \mid}{\sqrt{n \log\log n}} = \sqrt{2}\beta$$

As discussed in comment 3.1.2d this is equivalent to saying that $\frac{\partial}{\partial\theta_l} \log P_\theta(y_1^n) \mid_{\theta_0} = O_{a.s.}(\sqrt{n \log\log n})$.

Step 2:

Here we determine the order of $\hat{\theta}_n - \theta_0$. Clearly:

$$\begin{aligned}\frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta(y_1^n)\mid_{\hat{\theta}_n}=0 &= \frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta(y_1^n)\mid_{\theta_0} \\ &+ \frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\theta_0}(\hat{\theta}_n-\theta_0) \\ &+ o(\parallel\hat{\theta}_n-\theta_0\parallel^2)\end{aligned}$$

It follows that:

$$\begin{aligned}&\left[-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\theta_0}\right](\hat{\theta}_n-\theta_0) \\ =\ &\frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta(y_1^n)\mid_{\theta_0}+o(\parallel\hat{\theta}_n-\theta_0\parallel^2)\end{aligned}$$

Lemma 4.2.1 plays a crucial rôle here because it guarantees that $\hat{\theta}_n - \theta_0 \to 0 a.s.$ and therefore $o(\parallel\hat{\theta}_n-\theta_0\parallel^2)$ is negligible.

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\theta_0}\to -\frac{\partial^2}{\partial\theta^2}H_{\theta_0}(\theta)\mid_{\theta_0}>0$$

by hypothesis and therefore for n large enough the LHS is invertible and we have:

$$(\hat{\theta}_n-\theta_0)=\left[-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\theta_0}\right]^{-1}\left[\frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta(y_1^n)\mid_{\theta_0}+o(\parallel\hat{\theta}_n-\theta_0\parallel^2)\right]$$

From Step 1 we conclude that

$$(\hat{\theta}_n-\theta_0)=O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$$

Step 3

To complete the proof expand $\frac{1}{n}\log P_{\hat{\theta}_n}(y_1^n)$ in Taylor's series around θ_0:

$$\begin{aligned}\frac{1}{n}\log P_{\hat{\theta}_n}(y_1^n) &= \frac{1}{n}\log P_{\theta_0}(y_1^n) \\ &+ \frac{1}{n}\frac{\partial}{\partial\theta}\log P_\theta(y_1^n)\mid_{\theta_0}(\hat{\theta}_n-\theta_0) \\ &+ (\hat{\theta}_n-\theta_0)^T\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\tilde{\theta}_n}(\hat{\theta}_n-\theta_0)\end{aligned}$$

where $\tilde{\theta}_n$ is a point in $\Theta_{q_0}^\delta$ such that $\parallel\tilde{\theta}_n-\theta_0\parallel\ \leq\ \parallel\hat{\theta}_n-\theta_0\parallel$ and therefore $\tilde{\theta}_n\to\theta_0$. The matrix $\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log P_\theta(y_1^n)\mid_{\tilde{\theta}_n}$ is bounded (it actually converges) and from the results of Steps 1 and 2 we conclude that

$$\frac{1}{n}\log P_{\hat{\theta}_n}(y_1^n) = \frac{1}{n}\log P_{\theta_n}(y_1^n) + O_{a.s.}(\frac{\log\log n}{n}).$$

□

4.3 Finite memory approximation

We present here a result on the rate of growth of the MLR obtained using an approximation technique. The idea is to approximate the HMC process with a sequence of Markov chains of increasing order m. The result is too weak for our purpose of estimating the order of the HMC, but we believe the technique of proof to be interesting in itself. In this section we will not need assumption SP'. The only hypothesis on Y_t will be the following:

$\boxed{Y_t \text{ is an HMC of order } q \text{ admitting a representation } \theta_0 \text{ with } \theta_0 \epsilon \Theta_q^\delta}$

The standard log-likelihood ratio will be denoted $R_n(\theta)$ i.e.:

$$R_n(\theta) := \log \frac{P_\theta(y_1^n)}{P_{\theta_0}(y_1^n)}$$

and for $1 \leq m < n$ the m order approximation $R_n^m(\theta)$ is defined as:

$$R_n^m(\theta) := \sum_{k=1}^{m} \log \frac{P_\theta(y_k \mid y_1^{k-1})}{P_{\theta_0}(y_k \mid y_1^{k-1})} + \sum_{k=m+1}^{n} \log \frac{P_\theta(y_k \mid y_{k-m}^{k-1})}{P_{\theta_0}(y_k \mid y_{k-m}^{k-1})}$$

i.e. $R_n^m(\theta)$ is obtained keeping track of only the m most recent samples. Our first result gives a bound on how well $R_n^m(\theta)$ approximates $R_n(\theta)$.

Lemma 4.3.1

$$\mid R_n(\theta) - R_n^m(\theta) \mid \leq \frac{2}{\delta(1-\rho)}\, n\, \rho^m$$

for some $0 < \rho < 1$

Proof: From the definitions we get:

$$\begin{aligned} \mid R_n(\theta) - R_n^m(\theta) \mid \quad \leq \quad & \sum_{k=m+1}^{n} \mid \log P_\theta(y_k \mid y_1^{k-1}) - \log P_\theta(y_k \mid y_{k-m}^{k-1}) \mid \\ + \quad & \sum_{k=m+1}^{n} \mid \log P_{\theta_0}(y_k \mid y_1^{k-1}) - \log P_{\theta_0}(y_k \mid y_{k-m}^{k-1}) \mid \end{aligned}$$

For each term of the first sum we have:

$$\begin{aligned} & \mid \log P_\theta(y_k \mid y_1^{k-1}) - \log P_\theta(y_k \mid y_{k-m}^{k-1}) \mid \\ \leq & \sum_{h=1}^{k-m-1} \mid \log P_\theta(y_k \mid y_h^{k-1}) - \log P_\theta(y_k \mid y_{h+1}^{k-1}) \mid \end{aligned}$$

Using inequality $A.0.3$ we have:

$$\mid \log P_\theta(y_k \mid y_h^{k-1}) - \log P_\theta(y_k \mid y_{h+1}^{k-1}) \mid \leq \frac{1}{\delta}\, \rho^{k-h-1}$$

where $0 < \rho < 1$. Observe that the bound does not depend on θ and is therefore valid also for the terms of the second sum. Adding all the terms we get:

$$\mid R_n(\theta) - R_n^m(\theta) \mid \leq \frac{2}{\delta} \sum_{k=m+1}^{n} \sum_{h=1}^{k-m-1} \rho^{k-h-1}$$

Some minor algebra gives:

$$\begin{aligned} \sum_{k=m+1}^{n} \sum_{h=1}^{k-m-1} \rho^{k-h-1} &= \frac{\rho^m}{1-\rho} \sum_{k=m+1}^{n} (1 - \rho^{k-m-1}) \\ &= \rho^m \left[\frac{n-m-1}{1-\rho} + \frac{1-\rho^{n-m}}{1-\rho} \right] \leq \frac{n\rho^m}{1-\rho} \end{aligned}$$

□

It is possible to interpret $R_n^m(\theta)$ as the log-likelihood ratio of two m order Markov chains. Define:

$$P_\theta^m(y_1^n) := P_\theta(y_1^m) \prod_{k=m+1}^{n} P_\theta(y_k \mid y_{k-m}^{k-1})$$

and analogously for $P_{\theta_0}^m$

The process Y_t becomes an m order Markov chain under P_θ^m and it is easily seen that $R_n^m(\theta) = \log \frac{P_\theta^m(y_1^n)}{P_{\theta_0}^m(y_1^n)}$. Both P_θ^m and $P_{\theta_0}^m$ are elements of the set:

$\mathcal{P}_m :=$ { all transition probabilities $P(y_0 \mid y_{-m}^{-1})$ with elements $\geq \delta$}.

(This is the set that in Section 3.4 was called Θ_m, we introduce a new name for it to avoid confusion with the parameter set of the HMC). Let P be the generic element of $\mathcal{P}_m$ and define $R_n^m(P) := \log \frac{P(y_1^n)}{P_{\theta_0}^m(y_1^n)}$. Clearly $\{P_\theta(y_0 \mid y_{-m}^{-1})\ ;\ \theta \epsilon \Theta_q^\delta\} \subset \mathcal{P}_m$ and therefore

$$\sup_{\theta \epsilon \Theta_q^\delta} R_n^m(\theta) \leq \sup_{P \epsilon \mathcal{P}_m} R_n^m(P) := R_n^m(\hat{P})$$

where $\hat{P}$ denotes the maximum likelihood estimator of the transition probabilities $P^m_{\theta_0}(y_0 \mid y_{-m}^{-1})$. As usual r denotes the cardinality of the set of values of Y_t.

Lemma 4.3.2

$$R_n(\hat{\theta}_n) = r^m \; O_{a.s.}\; (\log\log n) + \frac{2}{\delta(1-p)}\; n\; \rho^m$$

Proof: Follows from Lemma 4.3.1 and Theorem 3.3.2.

□

Lemma 4.3.2 can be used to estimate the rate of growth of $R_n(\hat{\theta}_n)$ or the rate of convergence to zero of $\frac{1}{n}R_n(\hat{\theta}_n)$. Ideally we would like to prove that $\frac{1}{n}R_n(\hat{\theta}_n) = O_{a.s.}(\frac{\log\log n}{n})$ but the bound in 4.3.2 is too weak for this, nevertheless we have:

Lemma 4.3.3

$$\frac{1}{n}R_n(\hat{\theta}_n) = O_{a.s.}(\frac{\log\log n}{n^\alpha})$$

for all

$$\alpha \leq \alpha_M := \frac{\log_r \frac{(q-1)-\delta^2}{(q-1)+\delta^2}}{\log_r \frac{(q-1)-\delta^2}{(q-1)+\delta^2} - 1}$$

Proof: From Lemma 4.3.2 we have for all n large and some finite C:

$$\frac{n^{\alpha-1}}{\log\log n}R_n(\hat{\theta_n}) \leq C(r^m n^{\alpha-1} + \rho^m \frac{n^\alpha}{\log\log n}) + 1$$

The idea is to choose m as a function of n in such a way that the right hand side remains bounded for $n \to \infty$.

Clearly $r^m n^{\alpha-1} = 1(\forall\ n)$ if we choose $m = (1-\alpha)\log_r n$. (Since m must be positive for all n we get $\alpha < 1$). Adopting this value for m the second term becomes:

$$\begin{aligned} \rho^m \frac{n^\alpha}{\log\log n} &= \frac{n^{(1-\alpha)\log_r \rho} n^\alpha}{\log\log n} \\ &= \frac{n^{\alpha+(1-\alpha)\log_r \rho}}{\log\log n} \end{aligned}$$

This term remains bounded in n ($as\ n \to \infty$) if: $\alpha + (1-\alpha)\log_r \rho \leq 0$ i.e. for $0 < \alpha \leq \frac{\log_r \rho}{\log_r \rho - 1} < 1$. Using the expression for ρ obtained in $A.0.2$ we complete the proof.

□

4.4 Information theoretic approach

In this section we use a result from Information Theory to get a useful bound on the MLR valid for all values of q. Recall that by $P_{\hat{\theta}_q(n)}(y_1^n)$ we denoted the maximized probability $P_\theta(y_1^n)$ for P_θ a HMC with $\theta \epsilon \Theta_q^\delta$. We denote by $P_{ML_q}(Y_1^n)$ the corresponding maximized probability when $\theta \epsilon \Theta_q$. The next lemma is crucial. A complete proof is to be found in Csiszar [7]. In this section log denotes $\log_2$.

Lemma 4.4.1 *There exists a probability measure Q on $\mathcal{Y}^\infty$ such that*

$$\log \frac{P_{ML_q}(y_1^n)}{Q(y_1^n)} \leq \frac{d(q)}{2} \log n - c \quad \textit{for all } n \textit{ and } y_1^n$$

where c is a constant and $d(q) := q(q + r - 2)$

Sketch of the proof:

First we observe that:

$$\begin{aligned} P_{ML_q}(y_1^n) &= \max_{\theta \epsilon \Theta_q} P_\theta(y_1^n) = \max_{\theta \epsilon \Theta_q} \sum_{x_1^n} P_\theta(y_1^n \mid x_1^n) P_\theta(x_1^n) \\ &\leq \sum_{x_1^n} \max_\theta P_\theta(y_1^n \mid x_1^n) \cdot \max_\theta P_\theta(x_1^n) \end{aligned} \tag{4.1}$$

The proof proceeds by showing the existence of probability measures Q_1 and Q_2 such that:

$$\max_\theta P_\theta(y_1^n \mid x_1^n) \leq Q_1(y_1^n \mid x_1^n) n^{q(r-1)/2} \tag{4.2}$$

$$\max_\theta P_\theta(x_1^n) \leq Q_2(x_1^n) n^{q(q-1)/2} \tag{4.3}$$

Clearly $Q(y_1^n) := \sum_{x_1^n} Q_1(y_1^n \mid x_1^n) Q_2(x_1^n)$ is a probability measure on $\mathcal{Y}^\infty$ and substituting into (4.1) completes the proof. The existence of Q_1 and Q_2 is proved directly by actually constructing measures Q_1 and Q_2 that satisfy (4.2) and (4.3) respectively.

□

The following Theorem, based on Lemma 4.4.1, will be essential to finding estimators of the order that avoid overestimation.

Theorem 4.4.1

$$\overline{\lim}\ (\log n)^{-1} \log \frac{P_{\hat{\theta}_q(n)}(y_1^n)}{P_{\theta_0}(y_1^n)} \leq \frac{d(q)}{2} + 2 \quad a.s.\ P_{\theta_0}$$

Proof: Introducing the measure Q from Lemma 4.4.1 we have:

$$\log \frac{P_{\hat{\theta}_q(n)}(y_1^n)}{P_{\theta_0}(y_1^n)} = \log \frac{P_{\hat{\theta}_q(n)}(y_1^n)}{Q(y_1^n)} + \log \frac{Q(y_1^n)}{P_{\theta_0}(y_1^n)}$$

We multiply by $(\log n)^{-1}$ and apply the inequality $\overline{\lim}(a_n + b_n) \leq \overline{\lim} a_n + \overline{\lim} b_n$. The first term is evaluated using Lemma 4.4.1:

$$\overline{\lim}(\log n)^{-1} \log \frac{P_{\hat{\theta}_q(n)}(y_1^n)}{Q(y_1^n)} \leq \frac{d(q)}{2}$$

To evaluate the second term define:

$$A_n := \{y_1^n;\ (\log n)^{-1} \log \frac{Q(y_1^n)}{P_{\theta_0}(y_1^n)} > 2\}$$

Clearly:

$$A_n := \{y_1^n;\ Q(y_1^n) > n^2 P_{\theta_0}(y_1^n)\}$$

It follows that:

$$P_{\theta_0}(A_n) = \sum_{y_1^n \epsilon A_n} P_{\theta_0}(y_1^n) \leq \sum_{y_1^n \epsilon A_n} \frac{1}{n^2} Q(y_1^n) \leq \frac{1}{n^2}$$

Thus $\sum_n P_{\theta_0}(A_n) < \infty$ and from the easy direction of the Borel-Cantelli lemma we conclude that $P_{\theta_0}(A_n\ i.o.) = 0$. This is equivalent to $\overline{\lim}\ (\log n)^{-1} \log \frac{Q(y_1^n)}{P_{\theta_0}(y_1^n)} \leq 2$

□

4.5 Compensators avoiding overestimation

We are finally able to give a set of sufficient conditions on the compensators of the maximized likelihood (the sequences $\delta_n(q)$) to avoid overestimation of the order. Theorem 4.5.1 is complementary to Theorem 4.1.1, together they allow us to construct compensators $\delta_n(q)$ that guarantee strong consistency of the order estimator $\hat{q}(n)$. Theorem 4.5.1 is the analog of Theorem 3.4.3 and should be compared with it.

Theorem 4.5.1 *Compensators avoiding overestimation*
Let Y_t be a process satisfying assumptions SP' and PH. If the compensator is of the form:

$$\delta_n(q) := \varphi(n)h(q)$$

where the function φ satisfies:

$$\underline{\lim}\left(\frac{\log n}{n}\right)^{-1}\varphi(n) > 1$$

and the function h satisfies:

$$h(q') - h(q) \geq \frac{d(q')}{2} + 2 \qquad \forall q' > q \geq 1$$

Then:

$$\overline{\lim}\,\hat{q}(n) \leq q_0 \quad a.s.\ P_{\theta_0}$$

Proof: Let $q > q_0$. From the definitions we have:

$$C(q_0,n) - C(q,n) = \frac{1}{n}\log\frac{P_{\hat{\theta}_q}(y_1^n)}{P_{\hat{\theta}_{q_0}}(y_1^n)} + \varphi(n)(h(q_0) - h(q))$$

Therefore:

$$\begin{aligned}&\overline{\lim}\left(\frac{\log n}{n}\right)^{-1}[C(q_0,n) - C(q,n)]\\ \leq\ &\overline{\lim}\left(\frac{\log n}{n}\right)^{-1}\frac{1}{n}\log\frac{P_{\hat{\theta}_q}(y_1^n)}{P_{\hat{\theta}_{q_0}}(y_1^n)} + \overline{\lim}\left(\frac{\log n}{n}\right)^{-1}\varphi(n)(h(q_0) - h(q))\end{aligned}$$

The first term on the RHS can be bounded using Theorem 4.4.1 as:

$$\overline{\lim}\left(\frac{\log n}{n}\right)^{-1}\left(\frac{1}{n}\log\frac{P_{\hat{\theta}_q}(y_1^n)}{P_{\theta_0}(y_1^n)} - \frac{1}{n}\log\frac{P_{\hat{\theta}_q}(y_1^n)}{P_{\theta_0}(y_1^n)}\right) \leq \frac{d(q)}{2} + 2$$

This follows from the fact that $\frac{1}{n}\log\frac{P_{\hat{\theta}_q}(y_1^n)}{P_{\theta_{q_0}}(y_1^n)} = O_{a.s.}(\frac{\log\log n}{n})$ as a consequence of Theorem 4.2.1 and assumption PH. On the second term on the RHS we use the hypothesis on the h function ($q > q_0$) to get:

$$\overline{\lim}\left(\frac{\log n}{n}\right)^{-1}[C(q_0,n) - C(q,n)] \leq (\frac{d(q)}{2} + 2)[1 - \underline{\lim}(\frac{\log n}{n})^{-1}\varphi(n)]$$

The hypothesis on $\varphi(\cdot)$ now gives:

$$\overline{\lim}\left(\frac{\log n}{n}\right)^{-1}[C(q_0,n)-C(q,n)]<0$$

On the other hand $[C(q_0,n)-C(\hat{q}(n),n)]\geq 0$ by definition of $\hat{q}(n)$. We conclude that $\overline{\lim}\,\hat{q}(n)\leq q_0$

□

The existence of a strongly consistent estimator $\hat{q}(n)$ of the order q_0 will be established by giving examples of functions $h(\cdot)$ and $\varphi(\cdot)$ satisfying both the conditions imposed by Theorem 4.1.1 and Theorem 4.5.1.

Theorem 4.5.2 *The compensator:*

$$\delta_n(q):=2d^2(q)\frac{\log n}{n}$$

produces a strongly consistent estimator $\hat{q}(n)$ *of* q_0.

Proof: Clearly $\lim\delta_n(q)=0\ \forall q$ thus satisfying the conditions of Theorem 4.1.1. The function $\varphi(n):=2^{\frac{\log n}{n}}$ is such that $\underline{\lim}\,(\frac{\log n}{n})^{-1}\varphi(n)=2>1$ and therefore satisfies the condition imposed by Theorem 4.5.1. For the function $h(q):=d^2(q)$ we must check the condition:

$$h(\hat{q})-h(q)\geq\frac{d(\hat{q})}{2}+2\quad\forall\,\hat{q}>q\geq 1$$

Recall that $d(q):=q(q+r-2)$. The condition to be verified is equivalent to:

$$\hat{q}(\hat{q}+r-2)[\hat{q}(\hat{q}+r-2)-\frac{1}{2}]\geq q^2(q+r-2)^2+2$$

for all $\hat{q}>q\geq 1$. This is easily established observing that the LHS is increasing in $\hat{q}$ and that for $\hat{q}=q+1$ the inequality is verified.

□

Appendix A

We collect here some basic inequalities for HMC's found in Baum-Petrie [4] and often used in the text. The proofs of these results imitate closely analogous results for Markov chains given by Doob [9].

Lemma A.0.1
Let Y_t be a HMC with p.d.f. P_θ where $\theta \epsilon \Theta_q^\delta$ then

$$P_\theta(X_{t+1} = j \mid X_t = i,\ Y_{t_k}, t_k \epsilon T) \ \geq \ \mu_\delta$$

where $\mu_\delta = (1 + \frac{q-1}{\delta^2})^{-1}$ is independent of θ, T, Y_{t_k}, i, j.

Proof: Let j and j' be elements of $\mathcal{X}$ and suppose that $t_k \neq t+1$ for all k;

$$\begin{aligned}
&\frac{P_\theta(X_{t+1} = j \mid X_t = i, Y_{t_k})}{P_\theta(X_{t+1} = j' \mid X_t = i, Y_{t_k})} \\
= \ &\frac{P_\theta(X_{t+1} = j, X_t = i, Y_{t_k})}{P_\theta(X_{t+1} = j', X_t = i, Y_{t_k})} \\
= \ &\frac{P_\theta(X_{t+1} = j, Y_{t_k} t_k \geq t+1 \mid X_t = i,)}{P_\theta(X_{t+1} = j', Y_{t_k} t_k \geq t+1 \mid X_t = i,)} \\
= \ &\frac{\sum_{j_0} P_\theta(X_{t+2} = j_0, X_{t+1} = j, Y_{t_k} t_k \geq t+1 \mid X_t = i)}{\sum_{j_0} P_\theta(X_{t+2} = j_0, X_{t+1} = j', Y_{t_k} t_k \geq t+1 \mid X_t = i)} \\
= \ &\frac{\sum_{j_0} P(Y_{t_k} t_k \geq t+2 \mid X_{t+2} = j_0) a_{jj_0} a_{ij}}{\sum_{j_0} P(Y_{t_k} t_k \geq t+2 \mid X_{t+2} = j_0) a_{j'j_0} a_{ij'}}
\end{aligned}$$

Let $\alpha_{j_0} := P(Y_{t_k} t_k \geq t+2 \mid X_{t+2} = j_0)$. The last expression is:

$$\frac{a_{ij}}{a_{ij'}} \frac{\sum_{j_0} \alpha_{j_0} a_{jj_0}}{\sum_{j_0} \alpha_{j_0} a_{j'j_0}} = (*)$$

Since $a_{ij} \geq \delta \; \forall i,j$ we have

$$\frac{\sum_{j_0} \alpha_{j_0} a_{jj_0}}{\sum_{j_0} \alpha_{j_0} a_{j'j_0}} = \frac{\sum_{j_0} \alpha_{j_0} a_{j'j_0} \frac{a_{jj_0}}{a_{j'j_0}}}{\sum_{j_0} \alpha_{j_0} a_{j'j_0}} \leq \max_{j_0} \left(\frac{a_{jj_0}}{a_{j'j_0}} \right)$$

Therefore:

$$(*) \leq \frac{a_{ij}}{a_{ij'}} \max_{j_0} \left(\frac{a_{jj_0}}{a_{j'j_0}} \right) \leq \max_{i,j,j',j_0} \left(\frac{a_{ij} a_{jj_0}}{a_{ij'} a_{j'j_0}} \right) \leq \frac{1}{\delta^2}$$

since all elements are $\geq \delta$.

Let now $\rho_j := P_\theta(X_{t+1} = j \mid X_t = i, Y_{t_k} t_k \epsilon T)$ then $1 = \rho_j + \sum_{j' \neq j} \rho_{j'} \leq \rho_j + (q+1)\frac{\rho_j}{\delta^2}$ i.e. $\rho_j \geq (1 + \frac{q-1}{\delta^2})^{-1}$

If $t_k = t+1$ for some k the proof needs a minor modification.

□

To introduce the next Lemma we need to introduce some notation. C_t denotes a cylinder set in $\mathcal{X}_t^\infty$ i.e.

$$C_t := \{X_{t_1} = i_1, X_{t_2} = i_2, \cdots X_{t_n} = i_n \; where \; t_k \geq t \quad k = 1, 2, \cdots n\}$$

D denotes a cylinder set in $\mathcal{Y}_1^\infty$ i.e.

$$D := \{Y_{t_1} = y_1, Y_{t_2} = y_2, \cdots Y_{t_m} = y_m \; where \; t_h \; are \; arbitrary\}$$

$$\begin{aligned} M_\theta^+(d, C_t, D) &:= \max_i P_\theta(C_t \mid X_{t-d} = i, D) \\ M_\theta^-(d, C_t, D) &:= \min_i P_\theta(C_t \mid X_{t-d} = i, D) \end{aligned}$$

Lemma A.0.2

$$M_\theta^+(d, C_t, D) - M_\theta^-(d, C_t, D) \leq \rho^{d-1}$$

where $\rho = 1 - 2\mu_\delta$

Proof: (Follows closely the proof of an analogous result for Markov chains given in Lamperti [15]). We simplify notation to M_d^+, M_d^- moreover let $\gamma_k := P_\theta(C_t \mid X_{t-d} = k, D)$, $\beta_{ik} := P_\theta(X_{t-d} = k \mid X_{t-d-1} = i, D)$ and define i_0 and k_0 by: $M_{d+1}^+ = P_\theta(C_t \mid X_{t-d-1} = i_0, D)$ and $M_d^- = \gamma_{k_0}$.

With the new notation we have:

$$\begin{aligned} M_{d+1}^{+} &= P_\theta(X_{t-d} = k \mid X_{t-d-1} = i_0, D) = \sum_k \gamma_k \beta_{i_0 k} \\ &= \mu M_d^- + (\beta_{i_0 k_0} - \mu) M_d^- + \sum_{k \neq k_0} \gamma_k \beta_{i_0 k} \\ &\leq \mu M_d^- + (\beta_{i_0 k} - \mu + \sum_{k \neq k_0} \gamma_k) M_d^+ = \mu M_d^- + (1-\mu) M_d^+ \end{aligned}$$

Similary we get:

$$M_{d+1}^- \geq (1-\mu) M_{d+1}^- + \mu M_{d+1}^+$$

Together the last two inequalities give:

$$M_{d+1}^+ - M_{d+1}^- \leq (1-2\mu)(M_d^+ - M_d^-)$$

The result follows immediately

□

Lemma A.0.3

$$\mid P_\theta(C_t \mid Y_n^k) - P_\theta(C_t \mid Y_{n+1}^k) \mid \leq \rho^{t-n-1}$$

for all k and all $n \leq t-1$

Proof:

$$P_\theta(C_t \mid Y_n^k) = \sum_j P_\theta(C_t \mid Y_n^k, X_{n-1} = j) P_\theta(X_{n-1} = j \mid Y_n^k)$$

and therefore:

$$M_{t-n+1}^- \leq P_\theta(C_t \mid Y_n^k) \leq M_{t-n+1}^+$$

On the other hand

$$P_\theta(C_t \mid Y_{n+1}^k) = \sum_{y_n} P_\theta(C_t \mid Y_n^k) P_\theta(Y_n \mid Y_{n+1}^k)$$

is an average of the $P_\theta(Y_n \mid Y_{n+1}^k)$ probabilities and therefore:

$$M_{t-n+1}^- \leq P_\theta(C_t \mid Y_{n+1}^k) \leq M_{t-n+1}^+$$

The result now follows from Lemma A.0.2

□

References

[1] Azencott, R. and Dacunha-Castelle, D., *Series of Irregular Observations*, New York: Springer Verlag, 1986.

[2] Baker, J.K., "Stochastic Modeling for Automatic Speech Understanding", in *Speech Recognition*, Reddy, R. ed., New York: Academic Press, 1975.

[3] Barron, A.R., "The Strong Ergodic Theorem for Densities: Generalized, Shannon McMillan Breiman Theorem", *Ann. Probab.*, 16 (1985) 1292-1303.

[4] Baum, L.E. and Petrie, T., "Statistical Inference for Probabilistic Functions of Finite State Markov Chains", *Ann. Math. Stat.*, 37 (1966), 1554-63.

[5] Blackwell, D. and Koopmans, L., "On the Identifability Problem for Functions of Finite Markov Chains", *Ann. Math. Stat.*, 28 (1957), 1011-15.

[6] Carylye, J.W., "Stochastic Finite-State System Theory", in *System Theory*, Zadeh, L.A., Polak, E. eds., New York: McGraw-Hill, 1969.

[7] Csiszar, I., *Information Theoretic Methods in Statistics*, Notes for course ENEE 728F, University of Maryland, Spring 1990.

[8] Dempster, A.P., Laird, N.M. and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Roy. Statist. Soc., Ser. B*, 39 (1977), 1-38.

[9] Doob, J.L., *Stochastic Processes*, New York: Wiley 1953.

[10] Gilbert, E.J., "On the Identifiability Problem for Functions of Finite Markov Chains", *Ann. Math. Stat.*, 30 (1959), 688-697.

[11] Hannan, E.J. and Deistler, M., *The Statistical Theory of Linear Systems*, New York: Wiley, 1988.

[12] Heller, A., "On Stochastic Processes Derived from Markov Chains", *Ann. Math. Stat.*, 36, (1965), 1286-91.

[13] Ito, H., Amari, S. and Kobayashi, K., "Identifiability of Hidden Markov Information Sources and Their Minimum Degrees of Freedom", Proceedings of the 1990 IE^3 Intern. Workshop on Inf. Theory, June 10-15, 1990 (Eindhowen).

[14] Kieffer, J.C., "A Counter example to Perez's Generalization of the Shannon-McMillan Theorem", *Ann. Probab.*, 1 (1973), 362-64.

[15] Lamperti, J. *Stochastic Processes*, New York: Springer Verlag 1977.

[16] Levinson. S.E., Rabiner, L.R. and Sondhi, M.M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", *Bell Syst. Tech. J.*, 62, (1983), 1035-74.

[17] Merhav, N., Gutman, M. and Ziv, J., "On the Estimation of the Order of a Markov Chain and Universal Data Compression", *IEEE Trans. I.T.*, 35 (1989), 1014-1019.

[18] Neveu, J., *Martingales à temps discret*, Paris: Masson, 1972.

[19] Nishii, R., "Maximum Likelihood Principle and Model Selection when the True Model is Unspecified", *J. Multiv. Anal.*, 27 (1988), 392-403.

[20] Ornstein, D.S., "An Application of Ergodic Theory to Probability Theory", *Ann. Probab.*, 1 (1973), 43-58.

[21] Paz, A., *Introduction to Probabilistic Automata*, New York: Academic Press 1971.

[22] Petrie, T., "Probabilistic Functions of Finite State Markov Chains", *Ann. Math. Stat.*, 40 (1969), 97-115.

[23] Picci, G., "On the Internal Structure of Finite State Stochastic Processes", in *Recent Developments in Variable Structure Systems*, New York: Springer Verlag (Lecture Notes in Economics and Math. Systems, Vol. 162) 1978.

[24] Walters, P., *An Introduction to Ergodic Theory*, New York: Springer Verlag 1982.