Lecture Notes in Control and Information Sciences 184

T.E. Duncan, B. Pasik-Duncan(Eds.)

Stochastic Theory and Adaptive Control

Proceedings of a Workshop
held in Lawrence, Kansas, September 26 - 28, 1991

Springer-Verlag

CONSISTENT ESTIMATION OF THE ORDER OF HIDDEN MARKOV CHAINS [1]

John S. Baras
Electrical Engineering Department
and Systems Research Center
University of Maryland at College Park

and

Lorenzo Finesso
CNR Ladseb
Corso Stati Uniti, 4
35020 Padova, Italy

Abstract

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 of Hidden Markov Chain models. The specification of the penalty term requires precise information on the rate of growth of the maximized likelihood ratio. 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.

1. 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 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 [4], 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*

[1] This research was supported by National Science Foundation grant NSFD CDR 8803012, under the Engineering Research Centers Program.

that admit a HMC representation. This problem was solved by Heller [11] 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 [3] 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 [7]. 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 [12] which also includes an extensive bibliography.

Although much work has been dedicated to parameter 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 paper we describe our recent results on the problem of order estimation for hidden Markov chains. The detailed proofs can be found in [8]. 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], [10]), but we are not aware of any previous work that employs this approach for hidden Markov chains. 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 resort to a result from Information Theory to get the necessary asymptotics of the maximum likelihood.

2. Towards a Realization Theory for HMC's

There are many equivalent ways of defining HMC's. We particularly like the definition that originated in Realization Theory [16] and we will borrow it.

Definition 2.1 (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) $P(Y_{t+1} = y_{t+1}, X_{t+1} = x_{t+1} \mid Y_1^t = y_1^t, X_1^t = x_1^t) = P(Y_{t+1} = y_{t+1}, X_{t+1} = x_{t+1} \mid X_t = x_t)$

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 2.2 (HMC): A stochastic process Y_t with values 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 2.2 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 $(\Omega, \mathcal{F}, P)$ 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.

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 2.2. 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 Definition 2.2. Conversely, let Y_t be a HMC according to Definition 2.2 and X_t be the state process of a SFSS associated with Y_t. If we sum (ii) of Definition 2.1 over y_{t+1} we get $P(X_{t+1} = x_{t+1} \mid X_1^t = x_1^t, Y_1^t = y_1^t) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$ 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 Definition 2.1 (ii) 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.

In general HMCs 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. 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 $B_y := \text{diag}\,\{b_{1y}, b_{2y}, \cdots b_{cy}\}$ (where $c := \mid \mathcal{X} \mid$). The factorization hypothesis now gives: $M(y) = AB_y$.

In [11] Heller characterized the finite valued stationary processes Y_t that are HMC's. Let $\mathcal{Y}$ 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

$$q_i(\cdot \mid y) := \frac{q_i(y\cdot)}{q_i(y)}\epsilon\mathcal{C}$$

Then

Theorem 2.1 (Heller [11]): $P_Y(\cdot)$ is the *pdf* of a HMC iff the set $C_Y := conv\{P_Y(\cdot \mid u) \quad u\epsilon\mathcal{Y}*\}$ is contained in a polyhedral stable subset of $\mathcal{C}^*$.

Consider now a HMC Y_t with the set of parameters $\mathcal{M} := \{c, M(y), \pi\}$ where $c = \mid \mathcal{X} \mid$. It is 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. Two questions now arise naturally.

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?

Past work contains some partial answers. Unless otherwise noted the following summary is derived from the works of Gilbert [9], Carlyle [5] and Paz [14]. 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. Then 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 $g(v_i)$, the j-th column of H is $h(v_j')$ and $p(v_j v_j') = \pi M(v_i v_j')e = \pi M(v_i)M(v_j')e = g(v_i)h(v_j')$.

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

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

It is important to note that the concept of the rank of a *pdf* is only loosely related to the HMC property because there are examples of *pdf's* with finite rank that do not correspond to HMC's. 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. It is not difficult to establish that regular representations are minimal. As it was just noted 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. Or more precisely : *A finite number of operations is sufficient to determine the regularity of a given representation* $\mathcal{M} = \{c, M(\cdot), \pi\}$. 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. Then: *The non-regular elements of* Γ *are a closed subset of* $\mathcal{R}^k$*-Lebesgue measure zero.*

3. Families of HMC's

In this section we introduce the families of HMC's that will be used as model classes. 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 and simply write $P_\theta(y_1^n) = \pi M(y_1^n)e$).

Define $\Theta_q := \{\theta \epsilon \Theta; c_\theta = q\}$. Note that $\forall\, q \forall \theta \epsilon \Theta_q\ \exists \bar{\theta} \epsilon \Theta_{q+1}$ such that $P_{\bar{\theta}}(\cdot) = P_\theta(\cdot)$ or, abusing the notation, $\Theta_q \subset \Theta_{q+1}$. Statisticians refer to families having the last property 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 $\theta' \neq \theta (\theta' \epsilon \Theta_q) P_\theta(\cdot) \neq P_{\theta'}(\cdot)$ i.e. for at least one word $w, P_\theta(w) \neq P_{\theta'}(w)$. 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 $P_{\theta'}(\cdot) = P_\theta(\cdot)$. We will say that $\theta \epsilon \Theta_q$ is *identifiable modulo permutations* (i.m.p.) if the only points $\theta' \epsilon \Theta_q$ with $P_\theta(\cdot) = P_{\theta'}(\cdot)$ 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. We have adapted to our case the following theorem from Petrie [15] on identifiability.

Definition 3.1: $\theta = \{q, M(y), \pi\}$ is a Petrie point if: θ is regular, $M(y)$ is invertible $\forall y$, and $\exists y \epsilon \mathcal{Y}$ such that b_{iy}, $(i = 1, 2, \cdots q)$, are distinct.

Theorem 3.1(Petrie [15] adapted): The Petrie points of Θ_q are identifiable modulo permutation.

Theorem 3.2[8]:The set of Petrie points is open and of full Lebesgue measure in Θ_q.

It will often be convenient to somewhat restrict the family Θ in order to simplify statistical considerations. To this end we have the **Definition 3.2**:for $0 < \delta < 1/q$ define:
$\Theta_q^\delta := \{\theta \epsilon \Theta_q;\ a_{ij} \geq \delta,\ b_{jy} \geq \delta, \quad \forall i, j, y\}$.

With the abuse of notation introduced earlier we have:

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

This nested property will be essential later.

4. HMC's as Models of Stationary Processes

The consistency of the Maximum Likelihood Estimator (MLE) for HMC's was established in [3] under the assumption that the true distribution of the observations comes from a HMC. In our work we have shown 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 [3] is therefore a special case of ours. In the course of this work we have also obtained a slightly generalized version of the Shannon-McMillan-Breiman theorem.

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. 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.

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. 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.

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

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

Following the terminology from [13] we define the *quasi-maximum likelihood estimator* $\hat{\theta}(n)$ as:

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

Note that $\hat{\theta}(n)$ is defined as a set because no uniqueness is guaranteed for this class of models. It is easy to see that in the last equation 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:

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

It can also be shown that:

$$D(Q \parallel P_\theta) = H_Q - H_Q(\theta) \geq 0$$

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:

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

An equivalent description is

$$\mathcal{N} = \{\bar{\theta}\epsilon\Psi;\ H_Q(\bar{\theta}) = \max_{\theta\epsilon\Psi} H_Q(\theta)\}$$

For the proof of Theorem 4.1 below we need the following result, established in [].

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

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 4.1[8]:

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

This proof [8] is even simpler than the one given by Baum and Petrie [3] 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 $h_n(\theta,\ Y)$.

We now present a slightly generalized version of the Shannon-McMillan-Breiman (SMB). 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 4.2: Let Y_t be a finitely valued stationary ergodic process with probability distribution $Q(\cdot)$. Then:

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

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 4.2 have appeared for the case of real valued processes. Our result generalizes Theorem 4.2 to reference measures M of the HMC type but it applies only to finitely valued processes.

Theorem 4.3[8]: 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 $q(Y_1^k) = Q(Y_1^k)/P(Y_1^k)$ 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:

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

5. Estimation of the Order of a Hidden Markov Chain

The technique that was employed in [8, 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 [8] 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 $O_{a.s.}(\log\log n)$ 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 [8, 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 of [8] does *not* state that $\alpha_n = O_{a.s.}(\log\log n)$ *uniformly with respect to the realization* ω.

In Section 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 $\mid \mathcal{X} \mid = q$. 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. 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 5.1:

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 $\Theta_{q_0}^\delta$ for some $\delta > 0$.

The class of parametric models that will be used is

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

The results of Section 2 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.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 5.1 is observed. Find a compensator sequence $\delta_n(q)$ such that the estimator $\hat{q}(n)$ is strongly consistent i.e. $\hat{q} \to q_0$ $a.s.$ P_{θ_0}.

The analog of Theorem 3.4.2 of [8] is valid and we can easily give a sufficient condition on $\delta_n(q)$ that avoids underestimation.

Theorem 5.1 (Compensators avoiding underestimation)[8]: Let Y_t be a process satisfying Assumption 5.1. If $\lim_{n\to\infty}\ \delta_n(q) = 0 \quad (\forall\ q)$, then $\underline{\lim}_{n\to\infty} \hat{q}(n) \geq q_0 \quad P_{\theta_0} - a.s.$

To estimate the rate of convergence in $\Theta_{q_0}^{\delta}$, we study next 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, $\hat{\theta}_{q_0}(n)$ 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 5.2:

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

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 $0_{a.s}(\log\log n)$. Recall that:

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

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

The second preliminary result establishes the following bound needed for the application of the Law of Iterated Logarithm.

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

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

We are now ready to study the rate of convergence.

Theorem 5.2[8]:

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

We next 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 $\theta\epsilon\Theta^{\delta}_q$. We denote by $P_{ML_q}(Y_1^n)$ the corresponding maximized probability when $\theta\epsilon\Theta_q$. The next result is crucial. A complete proof is to be found in Csiszar [6].

Theorem 5.3: There exists a probability measure Q on $\mathcal{Y}^{\infty}$ such that

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

where c is a constant and $d(q) := q(q+r-2)$ As a sketch of the proof we observe first that:

$$P_{ML_q}(y_1^n) = \max_{\theta\epsilon\Theta_q} P_{\theta}(y_1^n) = \max_{\theta\epsilon\Theta_q}\sum_{x_1^n} P_{\theta}(y_1^n \mid x_1^n)P_{\theta}(x_1^n)$$

$$\leq \sum_{x_1^n}\max_{\theta} P_{\theta}(y_1^n \mid x_1^n)\cdot\max_{\theta} P_{\theta}(x_1^n)$$

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

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

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

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 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 the above.

The following Theorem, based on Theorem 5.3, will be essential to finding estimators of the order that avoid overestimation.

Theorem 5.4[8]:

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

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 5.5 is complementary to Theorem 5.1: together they allow us to construct compensators $\delta_n(q)$ that guarantee strong consistency of the order estimator $\hat{q}(n)$.

Theorem 5.5(Compensators avoiding overestimation)[8]: Let Y_t be a process satisfying Assumptions 5.1 and 5.2. 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:

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

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 5.1 and Theorem 5.5.

Theorem 5.6[8]: 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 5.1. The function $\varphi(n) := 2(\log n/n)$ is such that $\underline{\lim}\,(\frac{\log n}{n})^{-1}\varphi(n) = 2 > 1$ and therefore satisfies the condition imposed by Theorem 5.5. 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 left-hand side is increasing in $\hat{q}$ and that for $\hat{q} = q+1$ the inequality is verified.

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

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

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

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

[7] 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.

[8] Finesso, L. "Consistent Estimation of the Order for Markov and Hidden Markov Chains", Ph. D. Thesis Electrical Engin. Department, Technical Report Ph.D. 91-1, Systems Research Center, University of Maryland, College Park, 1991.

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

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

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

[12] 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.

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

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

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

[16] 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.