Probability Surveys
Vol. 18 (2021) 77–131
ISSN: 1549-5787
https://doi.org/10.1214/20-PS345

On universal algorithms for classifying and predicting stationary processes*

Gusztáv Morvai [1,2] **and Benjamin Weiss** [3]

[1] *Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, 13-15 Reáltanoda utca, H-1053, Budapest, Hungary*

[2] *MTA-BME Stochastics Research Group, 1 Egry József utca, Building H, Budapest,1111, Hungary*
e-mail: morvai@math.bme.hu

[3] *Hebrew University of Jerusalem, Jerusalem 91904 Israel*
e-mail: weiss@math.huji.ac.il

Abstract: This is a survey of results on universal algorithms for classification and prediction of stationary processes. The classification problems include discovering the order of a k-step Markov chain, determining memory words in finitarily Markovian processes and estimating the entropy of an unknown process. The prediction problems cover both discrete and real valued processes in a variety of situations. Both the forward and the backward prediction problems are discussed with the emphasis being on pointwise results. This survey is just a teaser. The purpose is merely to call attention to results on classification and prediction. We will refer the interested reader to the sources. Throughout the paper we will give illuminating examples.

AMS 2000 subject classifications: Primary 60G25, 60G10.
Keywords and phrases: Stationary processes, prediction theory.

Received May 2020.

Contents

*The first author was supported partly by the Alfréd Rényi Institute of Mathematics, the Bolyai János Research Scholarship and OTKA grant No. K75143.

1. Introduction

Fourty five years ago David Bailey wrote a PhD thesis under the direction of Donald Ornstein [4] entitled "Sequential schemes for classifying and predicting ergodic processes". Even though the thesis was never published it was very influential and gave rise to a great deal of work and it is our purpose to survey some of the developments in this research program. To put things in a proper historical perspective we will begin by reviewing the main results from that thesis.

The general problem considered there was that of extracting as much information as possible from a sequence of observations of a finite alphabet stationary stochastic process $X_0, X_1, ...X_n$. He gave the first universal estimation scheme for the evaluation of the Shannon entropy, prior to the schemes which arose from the universal data compression algorithms of J. Ziv and A. Lempel [105]. He then showed that for each k there was a sequence of functions g_n which when applied to $X_0, X_1, ...X_n$ would with probability one eventually equal YES/NO according to the alternative "the process IS/IS NOT a k-step mixing Markov chain". On the other hand he showed the non existence of a similar sequence of functions for deciding membership in the union over all k of these classes.

In contrast to the pioneering universal scheme of D. Ornstein [82] for estimating the conditional probability of X_0 given the infinite past $\{X_i : i \leq 0\}$ in a sequential fashion he showed the nonexistence of such a universal scheme for the forward problem of estimating the conditional probability of X_{n+1} given the observations $X_0, X_1, ...X_n$.

Since then much work has been done on questions of this type, by researchers such as Scarpellini [97, 98, 99], Paul Algoet [1, 2, 3], Amir Dembo and Yuval Peres [17], Meir Feder and Neri Merhav [50], Györfi, Kohler, Krzyzak and Walk [24], Andrew Nobel [81], Florentina Bunea and Andrew Nobel [11] Boris Ryabko [90], Daniel Jones, Michael Kohler and Harro Walk [35], Daniil Ryabko [92, 95, 96] Tina Felber, Daniel Jones, Michael Kohler and Harro Walk [18], Gusztáv Morvai and Benjamin Weiss [64, 59, 72, 76], Patrizia Berti, Irene Crimaldi, Luca Pratelli and Pietro Rigo [6], Hayato Takahashi [102], Eva Löcherbach and Enza Orlandi [48], Ramon van Handel [31], Dariusz Kalocinski and Tomasz Steifer [38, 39] and others and we will clearly be unable to describe all of the work that has been done. We shall give some results from these papers and others but will devote much of the survey to our own work. We turn now to a more detailed description of the survey.

In the first part we concentrate on discrete (finite or countably infinite) valued processes and begin by taking up the questions that relate to learning about general features of a process in a sequential fashion. We start by addressing the

problem of estimating the order k of a k-step Markov chain, including countable state chains. In contrast to Bailey's negative result for two valued decision schemes, we show that there is a sequence of functions g_n which when applied to the outputs $X_0, X_1, ...X_n$ of any ergodic process will converge with probability one to the order k if the process is k-step Markov and to infinity otherwise. We will also describe some further negative results, generalizing Bailey's, for classification of the class of processes called finitarily Markovian, where the next output depends on a finite segment of the past but the length of this segment is not bounded.

Following this we will describe some more general classification problems giving a variety of conditions under which one can, with eventual certainty, decide between membership in two disjoint classes of processes. In the last part of this section we will describe the recent striking characterization of the Shannon entropy of a process as essentially the only finitely observable isomorphism invariant of a process.

Most of the next section deals with estimation problems for finitarily Markovian processes (also called finite context processes or variable length Markov processes). Before continuing the introduction we pause to give an intuitive definition of this class. The memory length for a sequence of past observations $\{X_i : i \leq 0\}$ of a process is the smallest possible $0 \leq K(\ldots, X_{-1}, X_0) \leq \infty$ such that the conditional distribution of X_1 given the entire past is equal to the conditional distribution of X_1 given only $X_{1-K}, ..., X_0$. The least such value of K is called the memory length. When it is finite it should have the property that the same value is obtained for any other continuation $\{X'_j : j \leq -K\}$. A process is finitarily Markovian if with probability one this K is always finite. If it is bounded by k then the process is a Markov chain with order at most k.

We describe universal backward schemes for the estimation of this memory length which almost surely converge to the correct value $K(\ldots, X_{-2}, X_{-1}, X_0)$. The forward estimation problem of the memory length is the problem of determining $K(X_0, X_1, ...X_n)$, based on the observations of $(X_0, X_1, ...X_n)$. Here there is no universal scheme. We will show that even within the class of two step countable Markov chains one cannot successfully guess along a sequence of stopping times of density one whether the minimal memory length is one or two. We will also show that within the class of binary finitarily Markovian processes one cannot guess for $K(X_0, X_1, ..., X_{\lambda_n})$ on a sequence of stopping times λ_n with $\lambda_n/n \to 1$. The last part of this section deals with the special class of binary renewal processes and the problem of estimating the residual waiting time until the next occurrence of the renewal state.

The second part of the survey is devoted to real valued processes. In his thesis, Bailey [4] showed that for finite valued processes even though no scheme can be universally successful for forward estimation any universal backward scheme when used for forward prediction will converge almost surely in Cesaro mean, cf. also Ornstein [82]. Several authors have extended this to bounded real valued processes using quantization to reduce to the finite valued case see for example Algoet [1, 3], Morvai [53], Morvai Yakowitz and Györfi [56]. Yet another approach to the sequential prediction used a weighted average of expert schemes,

and with these schemes the results were extended to the general unbounded case by Nobel [80] and Györfi and Ottucsak[28], (see also the survey of Feder and Merhav [50]). However none of these results were optimal in the sense that moment conditions higher than those strictly necessary were assumed. We will describe some optimal results that we recently obtained for this forward prediction for real valued processes.

We have already mentioned the use of stopping times in devising universal schemes and we will describe a few results of this kind in the next subsection where we focus our attention on those processes where the conditional distribution of X_0 given the past becomes a continuous function of the past outputs after a set of probability zero is omitted. Next we take up the case of Gaussian processes which have been considered by Schäfer [100]. He constructed an algorithm which can estimate the conditional expectation for every time instance n for an extremely restricted class of Gaussian processes. A more general result giving an estimate for the conditional mean along a stopping time sequence will be described for stationary Gaussian (not necessarily ergodic) processes that include a much wider class of processes than that in Schäfer [100]. The disadvantage of these estimators is the rapid growth of the stopping times. A more realistic scheme will be given with a more moderate growth.

Throughout the survey we will give specific examples to illustrate the ideas.

2. Part I. Discrete valued processes

2.1. Discovering features of a process by sequential sampling

A stochastic process $\mathbf{X} = \{X_n : 0 \leq n < \infty\}$ is determined by the joint distributions of the random variables $\{X_0, X_1, ..., X_k\}$ for all k. We will be interested in stationary stochastic processes. These are those processes for which the joint distribution of $\{X_t, X_{t+1}, ..., X_{t+k}\}$ is the same as that of $\{X_0, X_1, ..., X_k\}$ for all t and all k. The simplest examples are independent identically distributed random variables and stationary Markov chains. Stationary processes can be uniquely extended into the past. This means that on a possibly enlarged sample space we have random variables $\{X_n : -\infty < n < \infty\}$ whose distributions are stationary.

For notational convenience, we will use the following notation throughout this survey $X_m^n = (X_m, \ldots, X_n)$, where $m \leq n$. We shall deal primarily with ergodic processes. These are stationary processes that cannot be decomposed into an average of stationary processes in a non-trivial fashion. Irreducible Markov chains are always ergodic. It is an easy consequence of Birkhoff's ergodic theorem that if a process $\{X_n\}$ is both stationary and ergodic, then from almost every sample sequence of the process one can determine the joint distributions. Indeed, in that case, for a fixed k, with probability 1, the empirical distributions on k-tuples determined by the sample will converge to the true distribution and the knowledge of these finite distributions gives the original process $\mathbf{X}$. In brief, with probability 1, a single sampling of an ergodic stationary process suffices to determine the nature of the process exactly.

A more realistic situation is one in which as time goes on we are presented with more and more observations and we are asked to give some information about $\mathbf{X}$ based on a finite sampling $x_0, x_1, ..., x_n$, which will get better and better as n increases. In this first section we will survey several kinds of specific problems that correspond to this general situation. We will begin with a simple problem in which we want to determine the order of K-step Markov chain, and then go on to discuss the more basic question of determining whether or not the process that we are observing is a Markov chain of some finite order. After these more specific classes of processes we will discuss more general classification problems and then conclude this section with a remarkable characterization of the entropy of a process the unique finitely observable isomorphism invariant. These notions will be defined below.

2.1.1. Estimating the order of a Markov chain

For a stationary stochastic process $\{X_n\}$ with values in some set $\mathcal{X}$, finite or countably infinite, a word $w \in \mathcal{X}^k$ of length k is called a memory word if the conditional probability of X_0 given the past is constant on the cylinder set defined by $X_{-k}^{-1} = w$. For a formal definition we introduce some notation for the distributions and conditional distributions: let $p(x_{-k}^0)$ denote the probability of the event $X_{-k}^0 = x_{-k}^0$ and let $p(y|x_{-k}^0)$ denote the conditional probability of the event $X_1 = y$ given that the event $X_{-k}^0 = x_{-k}^0$ occurred.

Note that the random variables are denoted by capital letters and particular realizations by lower case letters. For example, $p(y|X_{-k}^0)$ denotes the random variable which is a function of the random variables X_{-k}^0 taking the value $P(X_1 = y|X_{-k}^0 = x_{-k}^0)$ when $X_{-k}^0 = x_{-k}^0$.

Definition 2.1. *We say that the empty word $\emptyset$ with length zero is a memory word if for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-i+1}^0 \in \mathcal{X}^i$ such that $p(z_{-i+1}^0, y) > 0$:*

$$p(y) = p(y|z_{-i+1}^0).$$

If the empty word is a memory word then it is also called a minimal memory word.
For $k \geq 1$ we say that w_{-k+1}^0 is a memory word if $p(w_{-k+1}^0) > 0$ and for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-k-i+1}^{-k} \in \mathcal{X}^i$ such that $p(z_{-k-i+1}^{-k}, w_{-k+1}^0, y) > 0$:

$$p(y|w_{-k+1}^0) = p(y|z_{-k-i+1}^{-k}, w_{-k+1}^0).$$

If no proper suffix of w is a memory word then w is called a minimal memory word.

Note that the empty word is a memory word if and only if the stationary stochatic process is independent and identically distributed. Define the set $\mathcal{W}_k$ of those memory words w_{-k+1}^0 with length k and let $\mathcal{W}^*$ denote the set of all memory words. Note that $\mathcal{W}_0$ is either the empty set or it conains exactly the

empty word. Note also that if the empty word is a memory word then it is the only minimal memory word.

For example in a k-step Markov processes all words of length k are memory words. However, in general, a k-step Markov processes may also have shorter memory words, cf. Bühlmann and Wyner [10]. Naturally any left extension of a memory word is also a memory word.

Example 2.1. *Consider an independent and identically distributed process $\{X_n\}$ on a countable alphabet. Then the empty word is a memory word and it is the only minimal memory word. Now the length of the shortest minimal memory word is zero and the length of the longest minimal memory word is also zero.*

Example 2.2. *Consider the binary periodic Markov chain X_n with transition probabilities*

$$P(X_{n+1}=1|X_n=0)=P(X_{n+1}=0|X_n=1)=1.$$

This Markov chain yields a stationary process by choosing the initial distribution

$$P(X_0=0)=P(X_0=1)=0.5.$$

This stationary process is an ergodic process. Indeed, the process has only two possible realizations $\omega_{-\infty}^{\infty}$, either

$$\omega_0=1 \text{ and for } 1\le i<\infty\text{: } \omega_{-i}=\omega_i=0 \text{ if } i \text{ is odd and } \omega_{-i}=\omega_i=1 \text{ if } i \text{ is even}$$

or

$$\omega_0=0 \text{ and for } 1\le i<\infty\text{: } \omega_{-i}=\omega_i=1 \text{ if } i \text{ is odd and } \omega_{-i}=\omega_i=0 \text{ if } i \text{ is even}$$

each of the two realizations occures with probability 0.5 and an invariant set is either the empty set (wich has probability zero) or it must contain both of these realizations (in which case it has probability one). The minimal memory words are the '0' and the '1'. The other memory words w_{-k+1}^{0} with length k, $k\ge 2$, are those for which either

$$w_0=1 \text{ and for } 1\le i\le k-1\text{: } w_{-i}=0 \text{ if } i \text{ is odd and } w_{-i}=1 \text{ if } i \text{ is even}$$

or

$$w_0=0 \text{ and for } 1\le i\le k-1\text{: } w_{-i}=1 \text{ if } i \text{ is odd and } w_{-i}=0 \text{ if } i \text{ is even.}$$

The rest of the words have probability zero.

Example 2.3. *Consider the Markov chain $\{M_n\}$ with state space $S=\{0,1,2\}$ and transition probabilities*

$$P(M_2=1|M_1=0)=P(M_2=2|M_1=1)=1,$$

$$P(M_2=0|M_1=2)=P(M_2=1|M_1=2)=0.5.$$

This yields a stationary and ergodic process $\{M_n\}$. *Define*

$$Z_n = I_{\{M_n=1\}}.$$

Then $\{Z_n\}$ *is a stationary and ergodic binary Markov chain with order 2. The minimal memory words of the process* $\{Z_n\}$ *are the '1', the '10' and the '00'. Note that the length of the shortest minimal memory word is one and the length of the longest minimal memory word is two.*

The next example shows that the right extension of a memory word is not necessarily a memory word.

Example 2.4. *Consider the Markov chain* $\{M_n\}$ *with state space* $S = \{0, 1, 2, 3, 4\}$ *and transition probabilities*

$$P(M_2 = 0|M_1 = 0) = P(M_2 = 1|M_1 = 0) = 0.5,$$

$$P(M_2 = 0|M_1 = 3) = P(M_2 = 2|M_1 = 3) = 0.5$$

and

$$P(M_2 = 3|M_1 = 1) = P(M_2 = 4|M_1 = 2) = P(M_2 = 3|M_1 = 4) = 1.$$

This yields a stationary and ergodic process $\{M_n\}$. *Let function* $f : S \to \{a, b, c\}$ *be defined as*

$$f(x) = \begin{cases} c & \text{if } x = 0 \\ a & \text{if } x = 1 \text{ or } x = 2 \\ b & \text{if } x = 3 \text{ or } x = 4. \end{cases}$$

Define

$$Z_n = f(M_n).$$

Then $\{Z_n\}$ *is a stationary and ergodic Markov chain with order 3. The minimal memory words of the process* $\{Z_n\}$ *are 'a', 'c', 'cab' and 'bab'. Notice that though 'a' is a memory word the right extension 'ab', even though it has positive probability, is not a memory word. Note that the length of the shortest minimal memory word is one and the length of the longest minimal memory word is three.*

Example 2.5. *Consider the Markov chain* $\{M_n\}$ *with countably infinite state space* $S = \{0, 1, 2, \dots\}$ *and transition probabilities*

$$P(M_1 = n + 1|M_0 = n) = \left(\frac{1}{2}\right)^{n+1},$$

$$P(M_1 = 0|M_0 = n) = 1 - \left(\frac{1}{2}\right)^{n+1}$$

where $n \in S$. *This yields a stationary and ergodic first order Markov chain* $\{M_n\}$. *Define* $Z_n = I_{\{M_n \neq 0\}}$. *Then* $\{Z_n\}$ *is a stationary and ergodic binary renewal process with renewal state '0'. The minimal memory words of the process* $\{Z_n\}$ *are '0', '01', '011'. '0111', '01111', ...*

Example 2.6. *Consider a stationary and ergodic binary renewal process with renewal state '0'. Then any word w with positive probability which contains at least one '0' is a memory word, though not necessarily minimal. Any word w which contains more than one '0' can not be a minimal memory word.*

Consider the problem of determining the order of a Markoc chain, based on sequentially observing the outputs of a single sample $\{X_1, X_2, ..., X_n\}$. That is to say we would like to have sequences of functions L_n so that $L_n(X_1, X_2, ..., X_n)$ will converge almost surely to M, in case the process is a M-step Markov process but not a $(M-1)$-step Markov chain, and to infinity otherwise.

Early work on this problem like that of Merhav, Gutman and Ziv [51], Finesso [19, 20] Csiszár and Shields [13], Csiszár [14] and Peres and Sields [87] was restricted to finite state processes. This enabled them to use a priori rates for the convergence of empirical distributions and entropy estimators. Morvai and Weiss [63] gave the first universal order estimator for countable state Markov processes. However, in that scheme, the data segment was unnecessarily divided into two parts. Later, in [67], a simpler, better scheme was given which does not divide the data segment into two. To review this scheme we begin with a formal definition of the memory length.

Definition 2.2. *For a stationary time series $\{X_n\}$ the (random) length $K(X_{-\infty}^0)$ of the memory of the sample path $X_{-\infty}^0$ is the smallest possible $0 \leq K < \infty$ such that for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-K-i+1}^{-K} \in \mathcal{X}^i$*

$$p(y|X_{-K+1}^0) = p(y|z_{-K-i+1}^{-K}, X_{-K+1}^0)$$

provided $p(z_{-K-i+1}^{-K}, X_{-K+1}^0, y) > 0$, and $K(X_{-\infty}^0) = \infty$ if there is no such K.

In terms of the memory words this is simply the minimal K such that X_{-K+1}^0 is a memory word, if such a K exists, and is infinity otherwise.

Example 2.7. *Consider an independent and identically distributed process $\{X_n\}$ on a countable alphabet. Then*

$$K(X_{-\infty}^0) = 0$$

almost surely.

Example 2.8. *Consider a stationary and ergodic first order finite or countably infinite Markov chain $\{X_n\}$. Then*

$$K(X_{-\infty}^0) = 1$$

almost surely.

Example 2.9. *Consider the stationary and ergodic binary second order Markov chain $\{Z_n\}$ in Example 2.3. Then*

$$K(Z_{-\infty}^0) = \begin{cases} 1 & \text{if } Z_0 = 1 \\ 2 & \text{if } Z_0 = 0 \end{cases}$$

almost surely.

Example 2.10. *Consider a stationary and ergodic second order finite or countably infinite Markov chain* $\{X_n\}$*. Then*

$$K(X_{-\infty}^0) \leq 2$$

almost surely.

Example 2.11. *Consider a stationary and ergodic binary renewal process* $\{X_n\}$ *with renewal state '0'. Let* $\tau(X_{-\infty}^0)$ *be the smallest* $t \geq 0$ *such that* $X_{-t} = 0$ *and* $X_i = 1$ *for all* $-t < i \leq 0$*. Then*

$$K(X_{-\infty}^0) \leq \tau(X_{-\infty}^0) + 1$$

almost surely. Consider the stationary and ergodic binary renewal process $\{Z_n\}$ *in Example 2.5. Then*

$$K(Z_{-\infty}^0) = \tau(Z_{-\infty}^0) + 1$$

almost surely.

The goal is now to estimate the essential supremum of the function $K(X_{-\infty}^0)$. The essential supremum of $K(X_{-\infty}^0)$ is equal to the order of the Markov chain if the process is Markov of some order and infinity otherwise. In other words, the essential supremum of $K(X_{-\infty}^0)$ is the smallest $k \geq 0$ such that $P(X_1^k \in \mathcal{W}_k) = 1$ if there is such k and infinite otherwise.

In order to describe the estimate for this function we first give a formal definition of how to find the essential supremum of the function $K(X_{-\infty}^0)$. For $k \geq 0$ let $\mathcal{S}_k$ denote the support of the distribution of X_{-k}^0. Define

$$\Delta_0 = \sup_{1 \leq i} \sup_{(z_{-i+1}^0, x) \in \mathcal{S}_i} \left| p(x) - p(x|z_{-i+1}^0) \right|.$$

Define

$$\Delta_1 = \sup_{1 \leq i} \sup_{(z_{-i}^0, x) \in \mathcal{S}_{1+i}} \left| p(x|z_0) - p(x|z_{-i}^0) \right|.$$

In general, define

$$\Delta_k = \sup_{1 \leq i} \sup_{(z_{-k-i+1}^0, x) \in \mathcal{S}_{k+i}} \left| p(x|z_{-k+1}^0) - p(x|z_{-k-i+1}^0) \right|.$$

If for some k, $\Delta_k = 0$ then the process is a k-step Markov chain and the least such k is the order of the chain.

Example 2.12. *Consider the stationary and ergodic binary process* $\{X_n\}$ *in Example 2.2. Then*

$$\begin{aligned} \Delta_0 &= \max(|p(0) - p(0|0)|, |p(0) - p(0|1)|, |p(1) - p(1|0)|, |p(1) - p(1|1)|) \\ &= \max(|0.5 - 0|, |0.5 - 1|, |0.5 - 1|, |0.5 - 0|) \\ &= 0.5 > 0 \end{aligned}$$

and $\Delta_i = 0$ *for* $i \geq 1$.

Example 2.13. *Consider the stationary and ergodic binary process* $\{Z_n\}$ *in Example 2.3. Then*

$$\Delta_0 > 0,$$

$$\Delta_1 > 0,$$

and $\Delta_i = 0$ *for* $i \geq 2$.

Example 2.14. *Consider the stationary and ergodic binary renewal process* $\{Z_n\}$ *in Example 2.5. Then* $\Delta_i > 0$ *for all* $i \geq 0$.

We would like to define a statistic to estimate Δ_k. The key fact which we will use is the pointwise ergodic theorem. It follows from that theorem that with probability one, for all fixed k, the empirical distributions on k-tuples determined by the sample taken from 0 up to time n will converge as n tends to infinity to the true distribution. However at any finite stage we only have a finite sample at our disposal. It follows that we have to make sure that we have seen a specific k-block enough times to be sure that we are close to the truth. Here is the procedure in detail. (Cf. Morvai and Weiss [67].)

We denote the usual empirical distribution estimates for the conditional distributions $p(x|z_{-k+1}^0)$ from the samples X_0^n as $\hat{p}_n(x|z_{-k+1}^0)$. (In other words, $\hat{p}_n(x|z_{-k+1}^0)$ is the ratio of the number of occurrences of the string (z_{-k+1}^0, x) in the observed X_0^n to the number of occurrences of the string z_{-k+1}^0 in X_0^n.) These $\hat{p}$'s are functions of X_0^n, but we suppress this dependence.

As we have said we only want to consider this statistic if the sample afforded us is sufficiently large. One kind of such restriction is the following one.

For a fixed $0 < \gamma < 1$ let $\mathcal{S}_k^n$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in X_0^n. These are the strings which occur sufficiently often so that we can rely on their empirical distribution. Now define the empirical version of Δ_0 as follows:

$$\hat{\Delta}_0^n = \max_{1 \leq i \leq n} \max_{(z_{-i+1}^0, x) \in \mathcal{S}_i^n} \left| \hat{p}_n(x) - \hat{p}_n(x|z_{-i+1}^0) \right|.$$

Define the empirical version of Δ_1 as follows:

$$\hat{\Delta}_1^n = \max_{1 \leq i \leq n} \max_{(z_{-i}^0, x) \in \mathcal{S}_{1+i}^n} \left| \hat{p}_n(x|z_0) - \hat{p}_n(x|z_{-i}^0) \right|.$$

In general, define the empirical version of Δ_k as follows:

$$\hat{\Delta}_k^n = \max_{1 \leq i \leq n} \max_{(z_{-k-i+1}^0, x) \in \mathcal{S}_{k+i}^n} \left| \hat{p}_n(x|z_{-k+1}^0) - \hat{p}_n(x|z_{-k-i+1}^0) \right|.$$

By ergodicity, the empirical conditional probabilities tend to the true conditional probabilities. Now it is immediate that for any fixed k, by ergodicity,

$$\hat{\Delta}_k^n \geq \frac{\Delta_k}{2}$$

eventually almost surely. Now the key idea is that if the process is not Markov of any order then for any fixed $k \geq 0$,

$$\hat{\Delta}_k^n \geq \frac{\Delta_k}{2} > 0$$

eventually almost surely and if the process is Markov with order M then for each $0 \leq k < M$,

$$\hat{\Delta}_k^n \geq \frac{\Delta_k}{2} > 0$$

eventually almost surely, and for each $k \geq M$ not just

$$\lim_{n\to\infty} \hat{\Delta}_k^n = \Delta_k = 0$$

almost surely, but $\hat{\Delta}_k^n$ tends to zero with a rate. Thus define an estimate χ_n for the order from samples X_0^n as follows. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Set $\chi_0 = 0$, and for $n \geq 1$ let χ_n be the smallest $0 \leq k < n$ such that

$$\hat{\Delta}_k^n \leq n^{-\beta}$$

if there is such a k and n otherwise. The algorithm works because if the process is not Markov of any order or Markov but k is smaller than the order then $\hat{\Delta}_k^n$ will be bounded away from zero eventually almost surely and so $\hat{\Delta}_k^n$ will be greater than $n^{-\beta}$ eventually almost surely while if k is greater than or equal to the order of the Markov chain then $\hat{\Delta}_k^n$ tends to zero with a rate, that is, $\hat{\Delta}_k^n$ will not be greater than $n^{-\beta}$ eventually almost surely.

The next theorem asserts that this estimator is pointwise universally consistent.

Theorem 2.1 (Morvai and Weiss [67])**.** *For any ergodic, stationary process $\{X_n\}$ taking values from a finite or countably infinite alphabet if the observed process is Markov then the sequence of estimators χ_n converges to the order of the Markov chain almost surely and if the observed process is not Markov of any order then the sequence of estimators χ_n tends to infinity almost surely. In other words, for any ergodic, stationary process $\{X_n\}$ taking values from a finite or countably infinite alphabet the sequence of estimators χ_n converges almost surely to the essential supremum of the memory function $K(\cdot)$.*

Now if $M > 0$ is arbitrary but fixed then for the class of all stationary and ergodic processes $\chi_n < M$ eventually if the process is Markov with order less than M and $\chi_n \geq M$ eventually almost surely otherwise, cf. Morvai and Weiss [67]. A result in Morvai and Weiss [67] assserts that even when we restrict attention to countable second order Markov chains there is no universal estimator for the length of the shortest memory word that converges even in probability.

For further reading on related topics see also [16] and [88].

2.1.2. Classification for special processes

In this subsection we take up classification problems which seem simpler since all that we want to do is to determine if our observations are coming from a certain class or not. Here is how to formalize the situation.

Let $\mathcal{X}$ be discrete (finite or countably infinite) alphabet. Let $\{X_n\}$ be a stationary and ergodic time series.

If $\mathcal{G}$ is a subclass of all stationary and ergodic binary processes then a sequence of functions $g_n : \{0,1\}^n \to \{YES, NO\}$ is a classification for $\mathcal{G}$ in probability if $\lim_{n\to\infty} P(g_n(X_1,\ldots,X_n) = YES) = 1$ for all processes in $\mathcal{G}$, and $\lim_{n\to\infty} P(g_n(X_1,\ldots,X_n) = NO) = 1$ for all processes not in $\mathcal{G}$.

Similarly, $g_n : \{0,1\}^n \to \{YES, NO\}$ is a classification for $\mathcal{G}$ in a pointwise sense if $g_n(X_1,\ldots,X_n) = YES$ eventually almost surely for all processes in $\mathcal{G}$, and $g_n(X_1,\ldots,X_n) = NO$ eventually almost surely for all processes not in $\mathcal{G}$. Of course, if g_n is a classification in a pointwise sense then it is a classification in probability but a classification in probability is not necessarily a classification in a pointwise sense.

For the class $\mathcal{M}_k$ of k-step mixing Markov chains of fixed order k, there are pointwise estimators of the type we have just described. Bailey [4] gave such a scheme for independent processes ($k = 0$) and indicated how to generalize the result for the class of $\mathcal{M}_k$.) For the class $\mathcal{M}_{mix} = \bigcup_{k=0}^{\infty} \mathcal{M}_k$ of mixing Markov chains of any order, Bailey showed that no such classification exists.

Theorem 2.2. *(Bailey [4]) There is no sequence of functions*

$$g_n : \{0,1\}^n \to \{YES, NO\}$$

such that for all stationary and ergodic binary processes $\{X_n\}$

$$g_n(X_1,\ldots,X_n) = YES \text{ eventually almost surely}$$

if process $\{X_n\}$ *is in* $\mathcal{M}_{mix}$*, and*

$$g_n(X_1,\ldots,X_n) = NO \text{ eventually almost surely}$$

if the processes $\{X_n\}$ *is not in* $\mathcal{M}_{mix}$*.*

See Ornstein and Weiss [84] for some further results on this kind of question. For a generalization of this non-existence result of Bailey see Morvai and Weiss [61]. Now consider the class of finitarily Markovian processes. These are processes such that with probability one we will encounter a memory word but their lengths are not bounded. Simple examples of such processes are renewal processes where as we look back as soon we see a recurrent event we will have a memory word in our hand.

Definition 2.3. *The stationary time series* $\{X_n\}$ *is said to be finitarily Markovian if* $K(X_{-\infty}^0)$ *is finite (though not necessarily bounded) almost surely.*

In other words the stationary and egodic discrete process $\{X_n\}$ is finitarily Markovian if and only if $P(\bigcup_{k=0}^{\infty}\{X_{-k+1}^0 \in \mathcal{W}^*\}) = 1$ where $\mathcal{W}^*$ denotes the set of all memory words of the process. This class includes all finite order Markov chains (mixing or not) and many other processes such as the finitarily deterministic processes of Kalikow, Katznelson and Weiss [37].

Here is another example which includes all binary renewal processes with finite expected inter-arrival time. Let $\{M_n\}$ be any stationary and ergodic first order Markov chain with finite or countably infinite state space S. Let $s \in S$ be an arbitrary state with $P(M_1 = s) > 0$. Now let $X_n = I_{\{M_n=s\}}$. The resulting binary time series $\{X_n\}$ is stationary and ergodic. It is also finitarily Markovian. (Indeed, the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ does not depend on values beyond the first (going backwards) occurrence of one in $X_{-\infty}^0$ which identifies the first (going backwards) occurrence of state s in the Markov chain $\{M_n\}$.) The resulting time series $\{X_n\}$ is not a Markov chain of any order in general. (Indeed, consider the Markov chain $\{M_n\}$ with state space $S = \{0, 1, 2\}$ and transition probabilities $P(M_2 = 1|M_1 = 0) = P(M_2 = 2|M_1 = 1) = 1$, $P(M_2 = 0|M_1 = 2) = P(M_2 = 1|M_1 = 2) = 0.5$. This is the same Markov chain as in Example 2.3 and it yields a stationary and ergodic Markov chain $\{M_n\}$. The resulting time series $X_n = I_{\{M_n=0\}}$ will not be Markov of any order. The conditional probability $P(X_1 = 0|X_{-\infty}^0)$ depends on whether until the first (going backwards) occurrence of one you see an even or odd number of zeros.)

A result in Morvai and Weiss [61] asserts that there is no classification for membership in the class of binary finitarily Markovian processes. The result applies to both pointwise classifications and classifications in probability. For details see Morvai and Weiss [61].

In contrast to the negative result on classification for the class of finitarily Markovian processes, one can construct a classification rule for the class of renewal processes since in the case of the class of binary renewal processes (with renewal state zero) it is enough to check if each of the words from the countable set $\{0, 01, 011, \dots\}$ is a memory word, cf. Morvai and Weiss [73]. For more results see D. Ryabko [93, 94] or Morvai and Weiss [71, 76].

2.1.3. On classifying general processes

The general problem of when can one discriminate between two classes of processes has been studied by several authors. In order to obtain positive results the testing schemes considered are not restricted to being simply two valued as were the schemes considered in the previous section. Some sufficient conditions for this to be possible were given by A. Dembo and Y.Peres in [17] and more general ones by A. Nobel in [81]. Here is a brief description of one of Nobel's result. First a formal definition of what is meant by a testing scheme.

Definition 2.4. *A sequence of measurable functions $\phi_n : R^n \to [0, 1]$, $n > 1$, will be called a testing scheme. A testing scheme is continuous if each of its constituent functions is continuous. Families of ergodic processes, H_0 and H_1,*

are discernible with probability one if there exists a testing scheme such that:

$$\phi_n(X_1^n) \to i$$

with probability one exactly when $\mathbf{X} \in H_i$.

With this definition Nobel proves:

Theorem 2.3. *(Nobel [81]) Two families,* H_0 *and* H_1*, of stationary ergodic processes are continuously discernible if the following two conditions are satisfied. (i)* $H_0 \bigcup H_1$ *is contained in a countable union of uniformly tight subsets of the space of real valued stationary processes,* $\mathbf{M}_s$*. (ii) There exist two families* $U_1, U_2, ..., V_1, V_2, .. \subset \mathbf{M}_s$ *such that: (a) each* U_i, V_j *is contained in* $\mathbf{E}$*, the ergodic processes, and closed in Ms; (b)* $H_0 \subset U = \bigcup_i U_i$ *and* $H_1 \subset V = \bigcup_i V_i$*:. (c)* $U \bigcap V = \emptyset$.

Here is another result of this type drawn from [103]. One of the motivations was the desire to recognize in an effective way when a process is a function of a Markov chain. These are very popular today in the mathematical biology literature under the name "Hidden Markov Models" (HMM). In [21] one can find a very nice characterization of these processes as those which can be defined by a finite number of finite dimensional stochastic matrices. Essentially the same characterization was rediscovered several years later by A. Heller in [32]. There has been much work in finding methods for determining the best HMM to fit some given data. In light of this it is natural to ask – can one determine membership in this class or not by successive observations of $\{X_1, X_2, ...X_n\}$. D. Bailey showed in his thesis [4] that this is not even possible for the class of all k-step Markov chains (k arbitrary, fixed number of states). In [61] we give a similar negative result for another extension of the class of all Markov chains – the finitary Markov processes.

On the other hand, if one restricts the order and the size of the state space then there are guessing schemes g_n which will converge almost surely and test for membership, see for example [44], [13]. (In these papers there are integer valued schemes which are shown to converge to the least k such that the process is a k-step Markov chain, and with an a priori bound on the value of k this can be used to produce a two valued scheme which tests for membership in the class).

One can find such schemes for any family of ergodic processes with uniform rates in the ergodic theorem and a variant of this can be used for the class of all ergodic HMM where there is an a priori bound on the number of states in the Markov chain.

Let $\mathcal{F}$ denote some family of ergodic stochastic processes on a fixed state space S with a finite number of symbols. Identify these processes with the shift invariant measures on the compact space, S^Z, of bi-infinite sequences of elements from S. On this space of measures put the weak* topology to obtain a compact space. Convergence in this topology coincides exactly with convergence of all finite dimensional distributions. We will be concerned mainly with ergodic measures, since by the ergodic decomposition almost every sequence produced

by any stationary process is a typical sequence for some ergodic process. On the ergodic processes we take the induced topology. Thus when we speak of a closed family of ergodic processes we mean closed in this relative topology.

The estimation scheme will be based on the properties of the empirical distribution of k-blocks in n-strings based on the alphabet S. Let us introduce the following notation for this empirical distribution. Let $b \in S^k$ be a fixed k-block and $u \in S^n$ an n-string, then define

$$\mathbf{D}(b|u) = |\{1 \leq i \leq n-k+1 : u[i,...,i+k-1] = b\}|/(n-k+1).$$

Definition 2.5. *A closed family of ergodic stochastic processes $\mathcal{F}$ has uniform rates, if for every $k \in N$, and every $\epsilon > 0$ there is some $n = n(k,\epsilon)$ such that for all $P \in \mathcal{F}$ we have that $P\{u \in S^n : |P(b) - \boldsymbol{D}(b|u)| < \epsilon$, for all $b \in S^k\} > 1-\epsilon$.*

With this definition, for any closed family with uniform rates, a guessing scheme with two values, {YES,NO}, can be constructed which will almost surely stabilize on YES if the process belongs to $\mathcal{F}$ and to NO in the contrary case. To this end let $\mathcal{F}$ be a family with uniform rates, and fix a sequence ϵ_k such that it is summable.

Let $n_k = n(k,\epsilon_k)$ be the sequence which the definition supplies for us, and define g_n as follows:

For n in the range $[n_k, n_{k+1}-1]$ if for some $P \in \mathcal{F}$ we have that

$$|P(b) - \mathbf{D}(b|x_1, x_2, ...x_{n_k})| < \epsilon_k \text{ for all } b \in S^k$$

then set $g_n(x_1, x_2, ...x_n) = YES$ and if not set $g_n(x_1, x_2, ...x_n) = NO$.

With this definition we have that if the closed family of ergodic processes, $\mathcal{F}$, has uniform rates and the g_n are defined by (3.4)-(3.6) then for almost every realization of a process P from the family $\mathcal{F}$ we have that eventually $g_n(x_1, x_2, ...x_n) =$ YES, while for almost every realization of an ergodic process that is not in $\mathcal{F}$ eventually $g_n(x_1, x_2, ...x_n) =$ NO.

It is not hard to show that if $\mathcal{K}$ is a compact set of ergodic distributions then $\mathcal{K}$ has uniform rates.

For example, all Markov processes defined by transition matrices of a fixed size and a uniform positive lower bound on their entries, have uniform rates, since the set is clearly compact and consists of ergodic processes only. We can now formulate a theorem which is sufficiently general and whose assumptions are purely toplogical.

Theorem 2.4. *(Weiss [103]) If the family of ergodic processes, $\mathcal{E}$, is closed (in the set of all ergodic processes) and is also σ-compact, then there are g_n such that for almost every realization of a process P from the family $\mathcal{E}$ we have that eventually $g_n(x_1, x_2, ...x_n) = YES$, while for almost every realization of an ergodic process that is not in $\mathcal{E}$ eventually $g_n(x_1, x_2, ...x_n) = NO$.*

Note that in constract to Nobel's result the hypotheses refer only to the class $\mathcal{E}$, and not to its complement which would be needed to apply his theorem.

As examples of this theorem one can take all ergodic Markov processes with a fixed number of states. The σ-compactness can be seen by taking for the $\mathcal{K}_k$ all

those ergodic Markov processes defined by transition matrices where if an entry is non zero it is at least $1/k$. In a similar fashion one sees that all ergodic hidden Markov models with a fixed number of states and a bound on the window size of the function satisfy the hypotheses of the theorem. For further reading on related topics see [5], [34], [17], [84], [47], [103],[91], [71], [22] and [42].

2.1.4. Finite observability and entropy

We can put the questions that we have been considering in a yet more general framework. For simplicity we will consider only finite valued processes in this subsection. If J is a function of ergodic processes taking values in a metric space (Ω, d), then we say that J is *finitely observable* (*FO*) if there is some sequence of functions $S_n(x_1, x_2, ..., x_n)$ that converges to $J(\mathbf{X})$ for almost every realization of the process $\mathbf{X}$, for all ergodic processes. A weaker notion would involve convergence in probability of the functions S_n to J rather than convergence almost everywhere. The particular labels that a process carries play no role in the following and so we may assume that all our processes take values in finite subsets of Z.

Here are some examples of *FO* functions. If $J(\mathbf{X}) = E\{X_0\}$ is the expected value of X_0 then the basic pointwise ergodic theorem of G. D. Birkhoff implies that J is *FO* via the estimators $S_n(x_1, x_2, ..., x_n) = (x_1 + x_2 + ... + x_n)/n$.

This may easily be generalized as follows. Denote by $\mathbf{P}$ the shift-invariant probability measures on Z^Z with support on a finite number of symbols and the topology of convergence in finite dimensional distributions. This means that a sequence of probability measures μ_n converges to a limiting measure μ if and only if for each finite block b the measures $\mu_n([b])$ of the finite cylinder sets defined by the block b converge to $\mu([b])$. Then to each finite-valued stationary process there will correspond a unique element of $\mathbf{P}$, namely its distribution function DIST($\mathbf{X}$). This function is also *FO* by the same argument, replacing the arithmetic averages of the x_i by the empirical distributions of finite blocks. Next consider the **memory order** $L(\mathbf{X})$ of a process. This equals the minimal m such that the process is an m-Markov process, and $+\infty$ if no such m exists. (Note that $L(\mathbf{X})$ is a number associated with the distribution of process $\mathbf{X}$.) In §2.1 it is shown that this function is *FO*.

A better-known example is the Shannon entropy of a process. Here, several different estimators S_n are known to converge to the entropy; cf.[4, 106, 84, 85, 46]. The expected value of X_0 will clearly change if we change the labeling of our states but the Shannon entropy is not sensitive to such changes. In fact it is invariant under a very broad notion of equivalence of processes which we proceed to describe.

Processes $\mathbf{X}$ and $\mathbf{X}'$ are *isomorphic* if there is a stationary invertible coding going from one to the other. More formally, let us denote the bi-infinite sequence $...x_{-2}, x_{-1}, x_0, x_1, x_2, ...$ by $x_{-\infty}^{\infty}$, and the shift by T where $(Tx)_n = x_{n+1}$ for all n. A *coding* from $\mathbf{X}$ to $\mathbf{X}'$ is a mapping ϕ defined on the sequences $x_{-\infty}^{\infty}$ with values in $\mathbf{X}'$, which maps the probability distribution of the $\mathbf{X}$ random

variables to that of the $\mathbf{X}'$ random variables. It is *stationary* if almost surely $\phi T = T'\phi$, where T' is the shift on $\mathbf{X}'$. Finally, it is invertible if it is almost surely one-to-one. In this case it is not hard to see that the inverse mapping, where defined, will yield a stationary coding from $\mathbf{X}'$ to $\mathbf{X}$.

While the definition of the entropy of a process was given by C. Shannon [101] it was the great insight of A. Kolmogorov [45] that it is in fact an isomorphism invariant. This enabled him to solve an outstanding problem in ergodic theory; namely, he proved that independent processes with differing entropies are not isomorphic. Since that time entropy has turned out to be fundamental in many areas of ergodic theory. It is perhaps somewhat surprising that no new invariants of that kind were discovered and the next theorem of Ornstein and Weiss [86] explains this to some extent:

Theorem 2.5. *(Ornstein and Weiss[85]) If J is a finitely observable function, defined on all ergodic finite-valued processes, that is an isomorphism invariant, then J is a continuous function of the entropy.*

Note that there is no a priori assumption about the nature of the function J, such as measurability. An even stronger version of the theorem replaces isomorphism by the more restricted notion of finitary isomorphism. These are isomorphisms where the codings, in both directions, depend only on a finite (but variable) number of the variables. These are codings that are continuous after the removal of a null set. About ten years after Kolmogorov's result D. Ornstein [83] showed the converse; namely, independent processes with the same entropy are isomorphic. This was strengthened to finitary isomorphism by M. Keane and M. Smorodinsky [41], and is a strictly stronger notion than isomorphism, since there are many examples of processes that are isomorphic but not finitarily isomorphic.

It is natural to ask what happens when we restrict attention to smaller families of processes. That is, we now suppose that the finitely observable isomorphism invariant is only defined on a particular class and ask can one find any new invariants. Y. Gutman and M. Hochman ([23]) have proved a rather general theorem which shows that for many natural examples of classes of processes the answer remains negative. These classes include the main classes of the various mixing types. We will content ourselves with formulating just two of their results here.

Theorem 2.6 (Gutman and Hochman [23])**.** *If J is a finitely observable invariant on one of the following classes:*

1. *the Kronecker systems (the class of systems with pure point spectrum),*
2. *the zero entropy weakly mixing processes,*
3. *the zero entropy mildly mixing processes,*
4. *the zero entropy strongly mixing processes,*

Then J is constant.

For the class of irrational rotations the general problem is still open but they did obtain a partial result.

Theorem 2.7 (Gutman and Hochman [23]). *For every finitely observable invariant J on the class of irrational rotations, there is a Borel set $\Theta \subseteq [0,1)$ of full Lebesgue measure such that J assigns the same value to processes arising from rotations by angles in Θ. In particular, there is no complete finitely observable invariant for irrational rotations.*

2.2. Estimation for finitarily Markovian processes

In this section we will concentrate on the class of finitarily Markov processes and discuss several specific estimation problems for them. For our first problem we take up the basic question of detection of memory words (cf. Morvai and Weiss [65]). This problem has been discussed often in the context of modelling processes but mostly only for finite alphabet processes. We will show here how it relates to prediction questions.

To begin with, recall that K was the minimal length of the context that determines the conditional probability. Consider the problem of estimating the value of K, both in the backward sense, where we observe more and more of the past and in the forward sense, where one observes successive values of $\{X_n\}$ for $n \geq 0$ and asks for the least value K such that the conditional distribution of X_{n+1} given $\{X_i\}_{i=n-K+1}^{n}$ is the same as the conditional distribution of X_{n+1} given $\{X_i\}_{i=-\infty}^{n}$. We will not restrict to the finite alphabet case and include the possibility that the process takes countably infinite values.

Similar questions have been studied by Bühlman and Wyner in [10] but only for the case of finite alphabet finite order Markov chains. The possibility of countable alphabets complicates matters significantly. The reason is that while for finite alphabet Markov chains empirical distributions converge exponentially fast and one can establish universal rates of convergence for countable alphabet Markov chains no universal rates are available at all.

As for the classification problem, namely determining whether the observed process is finitarily Markovian or not, in Morvai and Weiss [61] it was shown that there is no classification rule for discriminating the class of finitarily Markovian processes from the other ergodic processes that are not.

In the first subsection we will review how to determine the value of $K(X_{-\infty}^{0})$ from observations of increasing length of the data segments X_{-n}^{0}. We will describe a universal consistent estimator which will converge almost surely to the memory length $K(X_{-\infty}^{0})$ for any ergodic finitarily Markovian process on a countable state space. Then we turn our attention to the forward estimation problem. This is the attempt to determine $K(X_{-\infty}^{n})$ from successive observations of X_0^n. The stationarity means that results in probability can be carried over automatically. However, almost sure results present serious problems as we have already mentioned previously. For more results in related to these questions of what can be learned about processes by forward observations see Ornstein and Weiss [84], Dembo and Peres [17], Nobel [79], and Csiszár and Talata [15].

In this last paper the authors define a finite context to be a memory word w of minimal length, that is, no proper suffix of w is a memory word. An infinite context for a process is an infinite string with all finite suffixes having

positive probability but none of them being a memory word. They treat there the problem of estimating the entire context tree in case the size of the alphabet is finite. For a bounded depth context tree, the process is Markovian, while for an unbounded depth context tree the universal pointwise consistency result there is obtained only for the truncated trees which are again finite in size. This is in contrast to the results discussed here which deal with infinite alphabet size and consistency in estimating memory words of arbitrary length. It is this generality that forces us to restrict to estimating at specially chosen times.

Finally, in the last subsection we will discuss estimating the residual waiting time in binary renewal processes. Recall that the classical binary renewal process is a stochastic process $\{X_n\}$ taking values in $\{0,1\}$ where the lengths of the runs of 1's between successive zeros are independent. These arise for example, in the study of Markov chains since the return times to a fixed state form such a renewal process. In many applications, the occurrences of a zero, which represent the failure times of some system which is renewed after each failure, are of importance and so the problem arises of estimating when the next failure will occur. Since this is usually unbounded this problem is rather difficult. We will give a rather detailed discussion of this problem and defer a more detailed description of the results to the subsection itself.

2.2.1. Estimation of the memory length for finitarily Markovian processes

Let $\{X_n\}$ be stationary and ergodic finitarily Markovian with finite or countably infinite alphabet $\mathcal{X}$. In this subsection we will first show how to determine the value of $K(X_{-\infty}^0)$ from observations of increasing length of the data segments X_{-n}^0. We will describe a universal consistent estimator which will converge almost surely to the memory length $K(X_{-\infty}^0)$ for any ergodic finitarily Markovian process on a countable state space.

In order to estimate $K(X_{-\infty}^0)$ (for the definition cf Definition 2.2) some explicit statistics are needed to be defined. These will be the same as those that we used when estimating its essential supremum in finding the order of a Markov chain. For the convenience of the reader we brifly repeat their definition. (Cf. e.g. Morvai and Weiss [65] or [73].) The first is a measurement of the failure of w_{-k+1}^0 to be a memory word. For the empty word $\emptyset$ with length zero $\Delta_0(\emptyset)$ is defined as

$$\Delta_0(\emptyset) = \sup_{1\le i} \sup_{\{z_{-i+1}^0 \in \mathcal{X}^i, x\in\mathcal{X}: p(z_{-i+1}^0, x)>0\}} \left| p(x) - p(x|z_{-i+1}^0) \right|.$$

If $\Delta_0(\emptyset) = 0$ then the process is independent and identically distributed. In general, for any $k \ge 1$ and for any word $w_{-k+1}^0 \in \mathcal{X}^k$, $\Delta_k(w_{-k+1}^0)$ is defined as

$$\Delta_k(w_{-k+1}^0) = \sup_{1\le i} \sup_{\{z_{-k-i+1}^{-k} \in \mathcal{X}^i, x\in\mathcal{X}: p(z_{-k-i+1}^{-k}, w_{-k+1}^0, x)>0\}} \left| p(x|w_{-k+1}^0) - p(x|z_{-k-i+1}^{-k}, w_{-k+1}^0) \right|.$$

This vanishes precisely when w_{-k+1}^0 is a memory word.

Example 2.15. *Consider the stationary and ergodic binary process* $\{X_n\}$ *in Example 2.2. Then*

$$\begin{aligned}\Delta_0(\emptyset) &= \max(|p(0)-p(0|0)|, |p(0)-p(0|1)|, |p(1)-p(1|0)|, |p(1)-p(1|1)|)\\ &= \max(|0.5-0|, |0.5-1|, |0.5-1|, |0.5-0|)\\ &= 0.5 > 0,\end{aligned}$$

$\Delta_1(1) = 0$ *and* $\Delta_1(0) = 0$.

Example 2.16. *Consider the stationary and ergodic binary process* $\{Z_n\}$ *in Example 2.3. Then*

$$\Delta_0(\emptyset) > 0,$$
$$\Delta_1(0) > 0,$$
$$\Delta_1(1) = 0,$$
$$\Delta_2(10) = 0$$

and

$$\Delta_2(00) = 0.$$

Example 2.17. *Consider the stationary and ergodic binary renewal process* $\{Z_n\}$ *in Example 2.5. Then*

$$\Delta_0(\emptyset) > 0,$$
$$\Delta_1(0) = 0,$$
$$\Delta_1(1) > 0,$$
$$\Delta_2(01) = 0$$
$$\Delta_2(11) > 0$$
$$\Delta_3(011) = 0$$
$$\Delta_3(111) > 0$$
$$\textit{etc.}$$

An empirical version of this based on the observation of a finite data segment X_{-n}^0 is needed. Let $\hat{p}_{-n}(x|w_{-k+1}^0)$ denote tne usual empirical version of the conditional probability $p(x|w_{-k+1}^0)$ from samples X_{-n}^0. These $\hat{p}$'s are functions of X_{-n}^0, but the dependence is suppressed to keep the notation manageable.

For a fixed $0 < \gamma < 1$ let $\mathcal{L}_k^n$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in X_{-n}^0. Now the empirical version of $\Delta_0(\emptyset)$ is as follows:

$$\hat{\Delta}_0^n(\emptyset) = \max_{1\le i\le n} \max_{(z_{-i+1}^0, x)\in\mathcal{L}_i^n} \left|\hat{p}_{-n}(x) - \hat{p}_{-n}(x|z_{-i+1}^0)\right|.$$

For any $k \ge 1$ and for any word $w_{-k+1}^0 \in \mathcal{X}^k$ the empirical version of Δ_k is as follows:

$$\hat{\Delta}_k^n(w_{-k+1}^0) =$$
$$\max_{1\le i\le n} \max_{(z_{-k-i+1}^{-k}, w_{-k+1}^0, x)\in\mathcal{L}_{k+i}^n} \left|\hat{p}_{-n}(x|w_{-k+1}^0) - \hat{p}_{-n}(x|z_{-k-i+1}^{-k}, w_{-k+1}^0)\right|.$$

By ergodicity, the ergodic theorem implies that almost surely the empirical

distributions $\hat{p}$ converge to the true distributions p and so for any $w_{-k+1}^0 \in \mathcal{X}^k$,

$$\liminf_{n\to\infty} \hat{\Delta}_k^n(w_{-k+1}^0) \geq \Delta_k(w_{-k+1}^0) \quad \text{almost surely.}$$

The key idea is that if w_{-k+1}^0 is not a memory word then

$$\liminf_{n\to\infty} \hat{\Delta}_k^n(w_{-k+1}^0) \geq \Delta_k(w_{-k+1}^0) > 0$$

almost surely and if w_{-k+1}^0 is a memory word then not just

$$\lim_{n\to\infty} \hat{\Delta}_k^n(w_{-k+1}^0) = \Delta_k(w_{-k+1}^0) = 0$$

almost surely, but $\hat{\Delta}_k^n(w_{-k+1}^0)$ tends to zero with a rate.

Now we review a test for w_{-k+1}^0 to be a memory word. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Let $NTEST_n(w_{-k+1}^0) = YES$ if $\hat{\Delta}_k^n(w_{-k+1}^0) \leq n^{-\beta}$ and NO otherwise. Note that $NTEST_n$ depends on X_{-n}^0. ('N' in NTEST stands for 'negative' since the data segment grows in negative (backward) direction.) By Morvai and Weiss [65], eventually almost surely, $NTEST_n(w_{-k+1}^0) = YES$ if and only if w_{-k+1}^0 is a memory word. Now we define an estimate χ_n for $K(X_{-\infty}^0)$ from samples X_{-n}^0 as follows. Set $\chi_0 = 0$, and for $n \geq 1$ let χ_n be the smallest $0 \leq k < n$ such that $NTEST_n(X_{-k+1}^0) = YES$ if there is such and n otherwise.

Theorem 2.8 (Morvai and Weiss [65])**.** *Let $\{X_n\}$ be a stationary and ergodic finitarily Markovian process taking values from a finite or countably infinite alphabet. Then*

$$\chi_n(X_{-n}^0) = K(X_{-\infty}^0)$$

eventually almost surely.

Now we turn our attention to the forward estimation problem where we are allowed to use growing segments of successive observations of X_0^n. Since when the word is a memory word one can use conditional independence and hence specific rates, either going backward or forward, and if the word is not a memory word one can use the forward ergodic theorem instead of the backward, it makes sense to define the forward version of the previous test as $PTEST_n(w_{-k+1}^0)(X_0^n) = NTEST_n(w_{-k+1}^0)(T^n X_0^n)$ where T is the left shift operator. ('P' in PTEST stands for 'positive' since the data segment grows in positive (forward) direction.) Now by Morvai and Weiss [65], eventually almost surely, $PTEST_n(w_{-k+1}^0) = YES$ if and only if w_{-k+1}^0 is a memory word.

PTEST tests a single word if it is a memory word or not. It is also possible to test a countable list of words (instead of a single word) if all of the words on the list are memory words or not, cf [73].

Now we shall examine how well can one estimate the local memory length for finite order Markov chains. In the case of finite alphabets this can be done with stopping times that eventually cover all time epochs (cf. Morvai and Weiss [65]). However, as soon as one goes to a countable alphabet, even if the order is known to be two and we are just trying to decide whether the X_n alone is a

memory word or not, there is no sequence of stopping times which is guaranteed to succeed eventually and whose density is one, cf. Morvai and Weiss [65].

Theorem 2.9. *(Morvai and Weiss [67]) There are no strictly increasing sequence of stopping times $\{\lambda_n\}$ and estimators $\{h_n(X_0,\ldots,X_{\lambda_n})\}$ taking the values one and two, such that for all countable alphabet Markov chains of order two*

$$\lim_{n\to\infty}\frac{\lambda_n}{n}=1 \quad \textit{almost surely}$$

and

$$h_n(X_0,\ldots,X_{\lambda_n})=K(X_0^{\lambda_n}) \quad \textit{eventually almost surely.}$$

We discussed that we cannot achieve density one in the forward memory length estimation problem even in the class of Markov chains on a countable alphabet. Now we shall show something similar in the class of binary (i.e. $0,1$) valued finitarily Markov processes. We will assume that there is given a sequence of estimators and stopping times, (h_n,λ_n) that do succeed to estimate successfully the memory length for binary Markov chains of finite order and construct a finitarily Markovian binary process on which the scheme fails infinitely often. Here is a precise statement:

Theorem 2.10. *(Morvai and Weiss [65]) For any strictly increasing sequence of stopping times $\{\lambda_n\}$ and sequence of estimators $\{h_n(X_0,\ldots,X_{\lambda_n})\}$, such that for all stationary and ergodic binary Markov chains with arbitrary finite order,*

$$\lim_{n\to\infty}\frac{\lambda_n}{n}=1 \quad \textit{almost surely}$$

and

$$h_n(X_0,\ldots,X_{\lambda_n})=K(X_0^{\lambda_n}) \quad \textit{eventually almost surely}$$

there is a stationary, ergodic finitarily Markovian binary time series such that

$$K(X_{-\infty}^n)\le n \quad \textit{eventually almost surely}$$

and with positive probability

$$h_n(X_0,\ldots,X_{\lambda_n})\neq K(X_{-\infty}^{\lambda_n}) \quad \textit{infinitely often.}$$

We emphasize that in the final counterexample process X_n that was constructed in Morvai and Weiss [65], eventually almost surely $K(X_{-\infty}^n)\le n$ and $K(X_{-\infty}^n)=K(X_0^n)$. For further reading cf. [73], [60] and [65].

2.2.2. On estimating the residual waiting time

In this subsection we investigate the possibility of giving a universal estimator at time n for the residual waiting time to the next zero in the binary renewal process $\{X_n\}$.

As for motivation consider a big system, e.g. a telephone exchange or a computer system. The sytem can be either in a good state or in a bad state. When the system breaks down (the system gets into a bad state) it is restarted (renewal). We observe the sequence of the states (good or bad states) of the system and observing these states to a certain time we would like to give an estimate to the residual waiting time to the next bad state / renewal. More precisely, we would like to estimate the conditional expectation of the residual waiting time until the next such renewal state without prior knowledge of the distribution.

Consider the renewal process $\{X_n\}$ with renewal state '0'. (For a formal definition see Morvai and Weiss [68].) We will assume that the process is stationary and ergodic. Even though our primary interest is in one sided processes, stationarity implies that there exists a two sided process with the same statistics and we will use the two sided version whenever it is convenient to do so. Note that these renewal processes are finitarily Markovian processes. Indeed, any word with positive probability from $\{0, 01, 011, 0111, \dots\}$ is a memory word, though not necessarily a minimal one.

Our interest is in the waiting time to renewal (the state 0) given some previous observations, in particular given X_0^n. We introduce the notation $\tau(X_{-\infty}^n)$ as the look back time for the last zero occurred in $X_{-\infty}^n$. Formally put

$$\tau(X_{-\infty}^n) = \text{the } t \geq 0 \text{ such that } X_{n-t} = 0, \text{ and } X_i = 1 \text{ for } n - t < i \leq n.$$

If a zero occurs in X_0^n then $\tau(X_{-\infty}^n)$ depends only on X_0^n and so we will also write for $\tau(X_{-\infty}^n)$, $\tau(X_0^n)$ with the understanding that this is defined only if a zero occurs in X_0^n.

Now for the classical binary renewal process $\{X_n\}$ define σ_i as the length of runs of 1's starting at position i. Formally put

$$\sigma_0 = \max\{0 \leq l : X_j = 1 \text{ for } 0 < j \leq i + l\}.$$

$$\sigma_1 = \max\{0 \leq l : X_j = 1 \text{ for } 1 < j \leq 1 + l\}$$

and in general put

$$\sigma_i = \max\{0 \leq l : X_j = 1 \text{ for } i < j \leq i + l\}.$$

For a simple example consider

$$(X_0, X_1, \dots, X_{11}) = (1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1).$$

Then

$$\begin{aligned} \sigma_0 &= 1 \\ \sigma_1 &= 0 \\ \sigma_2 &= 1 \\ \sigma_3 &= 0 \end{aligned}$$

$$\begin{aligned}\sigma_4 &= 4\\ \sigma_5 &= 3\\ \sigma_6 &= 2\\ \sigma_7 &= 1\\ \sigma_8 &= 0\end{aligned}$$

and σ_9, σ_{10} and σ_{11} are not yet defined.

For $k = 0, 1, \ldots$ let p_k denote the conditional probability that given $X_0 = 0$ it will be followed exactly by k ones until the next zero. Formally put

$$p_k = P(\sigma_0 = k | X_0 = 0).$$

Our goal is to estimate $E(\sigma_n | X_0^n)$ without prior knowledge of the distribution function of the process. In earlier works such as [43] attention is restricted to those renewal processes which arise from Markov chains with a finite number of states. In that case the problem is much easier since the probabilities p_k decay exponentially and one can use this information in trying to find not only the distribution but even the hidden Markov chain itself. We are considering the general case where the number of hidden states might be infinite and this exponential decay no longer holds in general.

For the estimator itself it is most natural to use the empirical distribution observed in the data segment $X_0, X_1, \ldots, X_n$. However if there were an insufficient number of occurrences of 1-blocks of length at least $\tau(X_0, X_1, \ldots, X_n)$ then we do not expect to give a good estimate. In particular if no block of that length has occurred yet, clearly no intelligent estimate can be given. For this reason we will estimate only along stopping times.

Unfortunately, there is no strictly increasing sequence of stopping times $\{\xi_n\}$ with density one, and sequence of estimators $\{h_n(X_0, \ldots, X_{\xi_n})\}$, such that for all binary classical renewal processes the error

$$|h_n(X_0, \ldots, X_{\xi_n}) - E(\sigma_{\xi_n} | (X_0, \ldots, X_{\xi_n})|$$

tends to zero almost surely as n tends to infinity, without higher moment assumptions on the p_k's. To obtain a positive result some higher moment assumptions on the p_k's are needed, cf. Morvai and Weiss [68]. Note also that the process is stationary means that the first moment of the p_k's must be finite. Furthermore, in order that the expected value of σ_0, that is, $E(\sigma_0)$ (not conditioned on the event that $X_0 = 0$) be finite the second moment of the p_k's has to be finite.

Now we describe the stopping times and the estimators. Define ψ as the position of the first zero, that is, $\psi = \min\{t \geq 0 : X_t = 0\}$. Let $0 < \delta < 1$ be arbitrary. Define the stopping times ξ_n as

$$\xi_0 = \psi,$$

$$\xi_1 = \min\left\{k > \xi_0 : \left|\left\{\psi \leq i < k : \tau(X_0^i) = \tau(X_0^k)\right\}\right| \geq k^{1-\delta}\right\},$$

$$\xi_2 = \min\left\{k > \xi_1 : \left|\left\{\psi \leq i < k : \tau(X_0^i) = \tau(X_0^k)\right\}\right| \geq k^{1-\delta}\right\}$$

and in general let

$$\xi_n = \min\left\{k > \xi_{n-1} : \ \left|\left\{\psi \le i < k : \ \tau(X_0^i) = \tau(X_0^k)\right\}\right| \ge k^{1-\delta}\right\}.$$

These are the successive times i when the value $t = \tau(X_0^i)$ has occurred previously enough times so that we can safely estimate the residual renewal time by empirical distributions derived from observations already made. We also need to fix κ_n as the index where reading backwards from X_{ξ_n} we will have seen for the first time $\ge \xi_n^{1-\delta}$ occurrences of an i with $\tau(X_0^i) = \tau(X_0^{\xi_n})$. Formally put

$$\kappa_1 = \max\{K : \left|\left\{K \le k < \xi_1 : \tau(X_0^k) = \tau(X_0^{\xi_1})\right\}\right| = \lceil \xi_1^{1-\delta} \rceil\},$$

$$\kappa_2 = \max\{K : \left|\left\{K \le k < \xi_2 : \tau(X_0^k) = \tau(X_0^{\xi_2})\right\}\right| = \lceil \xi_2^{1-\delta} \rceil\}$$

and in general, let

$$\kappa_n = \max\{K : \left|\left\{K \le k < \xi_n : \tau(X_0^k) = \tau(X_0^{\xi_n})\right\}\right| = \lceil \xi_n^{1-\delta} \rceil\}.$$

For $n > 0$ define our estimator $h_n(X_0, \ldots, X_{\xi_n})$ at time ξ_n as

$$h_1(X_0, \ldots, X_{\xi_1}) = \frac{1}{\lceil (\xi_1)^{1-\delta} \rceil} \sum_{i=\kappa_1}^{\xi_1 - 1} I_{\{\tau(X_0^i) = \tau(X_0^{\xi_1})\}} \sigma_i,$$

$$h_2(X_0, \ldots, X_{\xi_2}) = \frac{1}{\lceil (\xi_2)^{1-\delta} \rceil} \sum_{i=\kappa_2}^{\xi_2 - 1} I_{\{\tau(X_0^i) = \tau(X_0^{\xi_2})\}} \sigma_i$$

and in general, define

$$h_n(X_0, \ldots, X_{\xi_n}) = \frac{1}{\lceil (\xi_n)^{1-\delta} \rceil} \sum_{i=\kappa_n}^{\xi_n - 1} I_{\{\tau(X_0^i) = \tau(X_0^{\xi_n})\}} \sigma_i.$$

Note that κ_n ensures that we take into consideration exactly $\lceil (\xi_n)^{1-\delta} \rceil$ pieces of occurrences. The n-th estimate is simply the average of the residual waiting times that we have already observed in the data segment $X_{\kappa_n}^{\xi_n}$ when we were at the same value of τ as we see at time ξ_n.

Example 2.18. *Fix $\delta = 0.25$. Let $X_0^{12} = 1000101001100$. Note that $\psi = 1$. Calculate the τ's to get*

$$\begin{aligned} \tau(X_0^1) &= 0 \\ \tau(X_0^2) &= 0 \\ \tau(X_0^3) &= 0 \\ \tau(X_0^4) &= 1 \\ \tau(X_0^5) &= 0 \end{aligned}$$

$$\begin{aligned}
\tau(X_0^6) &= 1\\
\tau(X_0^7) &= 0\\
\tau(X_0^8) &= 0\\
\tau(X_0^9) &= 1\\
\tau(X_0^{10}) &= 2\\
\tau(X_0^{11}) &= 0\\
\tau(X_0^{12}) &= 0.
\end{aligned}$$

Calculate the σ's to get

$$\begin{aligned}
\sigma_0 &= 0\\
\sigma_1 &= 0\\
\sigma_2 &= 0\\
\sigma_3 &= 1\\
\sigma_4 &= 0\\
\sigma_5 &= 1\\
\sigma_6 &= 0\\
\sigma_7 &= 0\\
\sigma_8 &= 2\\
\sigma_9 &= 1\\
\sigma_{10} &= 0\\
\sigma_{11} &= 0.
\end{aligned}$$

Now calculating the ξ's one gets

$$\begin{aligned}
\xi_0 &= \psi = 1\\
\xi_1 &= 8\\
\xi_2 &= 11\\
\xi_3 &= 12.
\end{aligned}$$

Calculating $\frac{\xi_n}{n}$'s one gets

$$\begin{aligned}
\frac{\xi_1}{1} &= \frac{8}{1} = 8\\
\frac{\xi_2}{2} &= \frac{11}{2} = 5.5\\
\frac{\xi_3}{3} &= \frac{12}{3} = 4.
\end{aligned}$$

Calculate $\lceil \xi^{1-\delta} \rceil$'s to get

$$\begin{aligned}
\lceil \xi_1^{1-\delta} \rceil &= \lceil 8^{1-0.25} \rceil = 5\\
\lceil \xi_2^{1-\delta} \rceil &= \lceil 11^{1-0.25} \rceil = 7
\end{aligned}$$

$$\lceil \xi_3^{1-\delta} \rceil = \lceil 12^{1-0.25} \rceil = 7$$

Calculating the κ's one gets

$$\begin{aligned} \kappa_1 &= 1 \\ \kappa_2 &= 1 \\ \kappa_3 &= 2. \end{aligned}$$

Finally calculate the h's to get

$$\begin{aligned} h_1(X_0^8) &= \frac{\sigma_1+\sigma_2+\sigma_3+\sigma_5+\sigma_7}{5} = \frac{0+0+1+1+0}{5} = \frac{2}{5} \\ h_2(X_0^{11}) &= \frac{\sigma_1+\sigma_2+\sigma_3+\sigma_5+\sigma_7+\sigma_8}{6} = \frac{0+0+1+1+0+2}{6} = \frac{4}{6} \\ h_3(X_0^{12}) &= \frac{\sigma_2+\sigma_3+\sigma_5+\sigma_7+\sigma_8+\sigma_{11}}{6} = \frac{0+1+1+0+2+0}{6} = \frac{4}{6}. \end{aligned}$$

Theorem 2.11. *(Morvai and Weiss [68]) Assume $\sum_{k=0}^{\infty} k^{\alpha+1} p_k < \infty$ for some $\alpha > 2$. Let $0 < \delta < \min(1 - 2/\alpha, 1/3)$. Then*

$$\lim_{n\to\infty} \frac{\xi_n}{n} = 1$$

and

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\xi_n}) - E(\sigma_{\xi_n} | X_0, \ldots, X_{\xi_n})| = 0$$

almost surely.

Note that the fact that ξ_n/n tends to one means that we are estimating on a sequence that has density one, in other words, we rarely fail to give an estimate.

Note that both h_n and ξ_n depend on δ and so on α. We also constructed a more involved sequence of stopping times ξ_n^* and estimator $h_n^*(X_0, \ldots, X_{\xi_n^*})$ the constructions of which do not depend on a-priori knowledge of the α and we also managed to reduce our assumption from $\alpha > 2$ to $\alpha > 1$, cf. Morvai and Weiss [68]. We also constructed intermittent schemes for estimating the residual waiting time to the next zero for all binary stationary and ergodic processes. The scheme consists of a sequence of stopping times λ_n and estimators $f_n(X_0^{\lambda_n})$. For all binary stationary and ergodic processes,

$$\lim_{n\to\infty} \left| f_n(X_0^{\lambda_n}) - E(\sigma_{\lambda_n} | X_0^{\lambda_n}) \right| = 0$$

almost surely. If the process turnes out to be a binary renewal process then

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

almost surely. Cf. Morvai and Weiss [74]. For further reading see [75] and [78].

3. Part II. Estimation for real valued processes

In the first part of this survey we dealt exclusively with discrete valued processes. In this part we will deal with real valued processes. If the one dimensional marginal distribution is continuous then with probability one in a finite number of observations there will be no repetitions. This means that in order to be able to use any of the methods that we were considering before we will have to introduce quantizers which will group the data so that there will be repetitions. We will discuss in this section several positive results for the forward prediction problem for real valued processes. The first of these is based on an observation of Bailey that despite the fact that a backward scheme when used in the forward direction needn't converge pointwise it may be that it converges in Cesaro mean. The subsequent last section is based on the idea of **intermittent estimation**. This means that we do not predict at every time instant, but when we do predict we want to be certain that eventually our predictions are optimal.

3.1. Pointwise sequential estimation of the conditional expectation in Cesaro mean

In this section we consider the problem of estimating the conditional expectation $E(X_n|X_0^{n-1})$ from a single sample of length n. (For the origin of this problem cf. Cover [12].) We observe a longer and longer finite segment of the single sample path X_0^∞ and from the data segment X_0^{n-1} we want to estimate the conditional expectation $E(X_n|X_0^{n-1})$. Unfortunately this can not be done even for binary processes as the next theorem shows.

Theorem 3.1. *(Bailey [4], Ryabko [89]) For any estimator $\{\hat{E}_n(X_0^{n-1})\}$ there is a stationary ergodic binary-valued process $\{X_i\}$ such that*

$$\limsup_{n\to\infty} |\hat{E}_n(X_0^{n-1}) - E(X_n|X_0^{n-1})| > 0$$

with positive probability.

(Cf. Györfi, Morvai, and Yakowitz [27] also.)

In his thesis, Bailey [4] constructed a backward estimator $\hat{E}_{-n}(X_{-n}^{-1})$ which tries to approximate $E(X_0|X_{-n}^{-1})$. It turned out that to estimate the conditional expectation of a fixed random variable X_0 is possible as the next theorem shows.

Theorem 3.2. *(Bailey [4], Ornstein [82]) For the backward estimator $\hat{E}_{-n}(X_{-n}^{-1})$ constructed in Bailey [4] (cf. Ornstein [82] also) and for all stationary and ergodic binary processes $\{X_i\}$*

$$\lim_{n\to\infty} |\hat{E}_{-n}(X_{-n}^{-1}) - E(X_0|X_{-n}^{-1})| = 0$$

almost surely.

(Algoet [1], Morvai [53], Morvai, Yakowitz and Györfi [56] have extended this from binary processes to bounded real-valued stationay and ergodic processes.

Györfi et. al. [24] and Algoet [3] extended the above result further to unbounded real-valued stationary processes.)

In his thesis, Bailey [4] (cf. Ornstein [82] also) indicated how Maker's (also known as Breiman's) generalized ergodic theorem can be used to turn the backward estimator into a forward estimator for which the error will tend to zero in Cesaro average.

Theorem 3.3. *(Maker [49], Breiman [8, 9], Algoet [2]) Consider a stationary and ergodic dynamical system with the usual left shift oparator T. Let f_n be a sequence of real valued functions such that*

$$\lim_{n\to\infty} f_n = f$$

almost surely. If in addition

$$E(\sup_{n\geq 1} |f_n|) < \infty$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(T^i\omega) = E(f)$$

almost surely.

Note that if the f_n's are bounded then the condition $E(\sup_{n\geq 1} |f_n|) < \infty$ is trivially true. Now combine the above theorems with

$$f_n = |\hat{E}_{-n}(X_{-n}^{-1}) - E(X_0|X_{-n}^{-1})|$$

and

$$f = 0$$

to get

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \left|\hat{E}_i(X_0^{i-1}) - E(X_i|X_0^{i-1})\right| = 0 \quad \text{almost surely}$$

where $\hat{E}_i(X_0^{i-1}) = \hat{E}_{-i}(X_{-i}^{-1})(T^i\omega)$ cf. also Ornstein [82]. Several authors have extended this from binary processes to bounded real valued processes using quantization to reduce to the finite valued case see for example Algoet [1, 3], Morvai [53], Morvai, Yakowitz and Györfi [56]. The extension to the unbounded case turned out to be difficult because of the requirement of the integrability of the supremum in Maker's theorem.

A different approach to the sequential prediction uses a weighted average of simple estimators called 'experts', cf. e.g. Györfi and Lugosi [25]. The simple estimators can be partition-based, kernel-based etc. (cf. e.g. Györfi and Ottucsák and Walk [29]) The weight of an expert in the weighted average depends on its past performance as an estimator of the next outcome. These schemes are constructed directly as forward schemes and with these, results were extended to the general unbounded case by Nobel [80] and Györfi and Ottucsák[28].

Theorem 3.4. *(Györfi and Ottucsák[28]) Let $\{X_n\}$ be stationary and ergodic real-valued process with $E\left(|X_0|^4\right) < \infty$. Then for the estimator $\hat{E}_n(X_0^{n-1})$ defined in [28] (which is based on the idea of combining simple estimators called 'experts'):*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \left|\hat{E}_i(X_0^{i-1}) - E(X_i|X_0^{i-1})\right|^2 = 0 \quad \textit{almost surely}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \left|\hat{E}_i(X_0^{i-1}) - X_i\right|^2 = E\left(\left|E(X_0|X_{-\infty}^{-1}) - X_0\right|^2\right) \quad \textit{almost surely.}$$

(In fact, Györfi and Ottucsák considered a little bit more general framework when side information is also available, cf. [28], but for the case of simplicity we stated their result in a little bit simpler setting.) However none of these results were optimal in the sense that moment conditions higher than those strictly necessary were assumed. In our work [70] we have obtained optimal results by managing to prove the integrability of the supremum for the backward estimator and it is these results that we shall now review briefly. (For the the algorihm cf. Morvai, Yakowitz and Györfi [56], Algoet [3] and Morvai and Weiss[70].) Let $\{X_n\}$ be a real-valued doubly infinite stationary ergodic time series.

Since the process is real-valued and the scheme is based on pattern matching, quantization is needed. Let $[\cdot]^k$ denote the quantizer

$$[x]^k = \begin{cases} 0 & \text{if } -2^{-k} < x < 2^{-k} \\ -i2^{-k} & \text{if } -(i+1)2^{-k} < x \le -i2^{-k} \text{ for some } i = 1, 2, \dots \\ i2^{-k} & \text{if } i2^{-k} \le x < (i+1)2^{-k}. \end{cases}$$

(Cf. Algoet [3].) Let $[X_m^n]^k$ denote $([X_m]^k, \dots, [X_n]^k)$.

Example 3.1. *Assume that $X_0 = \frac{1}{2^3}$. Then*

$$\begin{aligned} [X_0]^1 &= \left[\frac{1}{2^3}\right]^1 = 0 \\ [X_0]^2 &= \left[\frac{1}{2^3}\right]^2 = 0 \\ [X_0]^3 &= \left[\frac{1}{2^3}\right]^3 = \frac{1}{2^3} \\ [X_0]^4 &= \left[\frac{1}{2^3}\right]^4 = \frac{1}{2^3}. \end{aligned}$$

Example 3.2. *Assume that $X_0 = \pi$ and $X_1 = 0$. Then*

$$[X_0^1]^1 = ([X_0]^1, [X_1]^1) = ([\pi]^1, [0]^1) = (3, 0)$$

$$\begin{aligned}
[X_0^1]^2 = ([X_0]^2, [X_1]^2) = ([\pi]^2, [0]^2) &= (3, 0) \\
[X_0^1]^3 = ([X_0]^3, [X_1]^3) = ([\pi]^3, [0]^3) &= (\frac{25}{8}, 0) \\
[X_0^1]^4 = ([X_0]^4, [X_1]^4) = ([\pi]^4, [0]^4) &= (\frac{25}{8}, 0).
\end{aligned}$$

The sequences λ_{k-1}, R_{k-1} and τ_k are defined recursively ($k = 1, 2, \dots$). Put $\lambda_0 = 1$ and $R_0 = 0$. Let τ_1 be the time between the occurrence of the pattern

$$[X_{-1}]^1$$

at time -1 and the last occurrence of the same pattern prior to time -1. More precisely, let

$$\tau_1 = \min\{t > 0 : [X_{-1-t}]^1 = [X_{-1}]^1\}.$$

Put

$$\lambda_1 = \tau_1 + \lambda_0.$$

Define

$$R_1 = X_{-\tau_1}.$$

Let τ_2 be the time between the occurrence of the pattern

$$([X_{-\lambda_1}]^2, \dots, [X_{-1}]^2) = [X_{-\lambda_1}^{-1}]^2.$$

at time -1 and the last occurrence of the same pattern prior to time -1. More precisely, let

$$\tau_2 = \min\{t > 0 : [X_{-\lambda_1-t}^{-1-t}]^2 = [X_{-\lambda_1}^{-1}]^2\}.$$

Put

$$\lambda_2 = \tau_2 + \lambda_1.$$

Define

$$R_2 = \frac{X_{-\tau_1} + X_{-\tau_2}}{2}.$$

In general, let τ_k be the time between the occurrence of the pattern

$$([X_{-\lambda_{k-1}}]^k, \dots, [X_{-1}]^k) = [X_{-\lambda_{k-1}}^{-1}]^k.$$

at time -1 and the last occurrence of the same pattern prior to time -1. More precisely, let

$$\tau_k = \min\{t > 0 : [X_{-\lambda_{k-1}-t}^{-1-t}]^k = [X_{-\lambda_{k-1}}^{-1}]^k\}.$$

Put

$$\lambda_k = \tau_k + \lambda_{k-1}.$$

Define

$$R_k = \frac{1}{k} \sum_{1 \le j \le k} X_{-\tau_j}.$$

(Cf. Morvai and Weiss [70], Algoet [3] and Morvai et. al. [56].)

Example 3.3. *Let* $X_{-9}^{-1} = (X_{-9}, X_{-8}, \ldots, X_{-2}, X_{-1}) = 010010010$. *Note that* $\lambda_0 = 1$, $R_0 = 0$. *The* τ*'s are:*

$$\begin{aligned} \tau_1 &= 2 \\ \tau_2 &= 3 \\ \tau_3 &= 3. \end{aligned}$$

The λ*'s are:*

$$\begin{aligned} \lambda_0 &= 1 \\ \lambda_1 &= \tau_1 + \lambda_0 = 3 \\ \lambda_2 &= \tau_2 + \lambda_1 = 6 \\ \lambda_3 &= \tau_3 + \lambda_2 = 9. \end{aligned}$$

The $X_{-\tau}$*'s are:*

$$\begin{aligned} X_{-\tau_1} &= X_{-2} = 1 \\ X_{-\tau_2} &= X_{-3} = 0 \\ X_{-\tau_3} &= X_{-3} = 0. \end{aligned}$$

The R*'s are:*

$$\begin{aligned} R_0 &= 0 \\ R_1 &= \frac{1}{1} \sum_{1 \le j \le 1} X_{-\tau_j} = \frac{1}{1} = 1 \\ R_2 &= \frac{1}{2} \sum_{1 \le j \le 2} X_{-\tau_j} = \frac{1+0}{2} = \frac{1}{2} \\ R_3 &= \frac{1}{3} \sum_{1 \le j \le 3} X_{-\tau_j} = \frac{1+0+0}{3} = \frac{1}{3}. \end{aligned}$$

To obtain a fixed sample size $t > 0$ version, let κ_t be the maximum of non-negative integers k for which $\lambda_k \le t$. For $t > 0$ put

$$\hat{R}_{-t} = R_{\kappa_t}.$$

Note that

$$\hat{R}_{-t} = R_k \quad \text{as long as } \lambda_k \le t < \lambda_{k+1}$$

and $\hat{R}_{-t}$ depends solely on X_{-t}^{-1}.

Example 3.4. *Let* $X_{-9}^{-1} = (X_{-9}, X_{-8}, \ldots, X_{-2}, X_{-1}) = 110111011$. *Note that* $\lambda_0 = 1$, $R_0 = 0$. *The* τ*'s are:*

$$\begin{aligned} \tau_1 &= 1 \\ \tau_2 &= 3 \\ \tau_3 &= 4. \end{aligned}$$

The λ*'s are:*

$$\begin{aligned}\lambda_0 &= 1\\ \lambda_1 &= \tau_1 + \lambda_0 = 2\\ \lambda_2 &= \tau_2 + \lambda_1 = 5\\ \lambda_3 &= \tau_3 + \lambda_2 = 9.\end{aligned}$$

The $X_{-\tau}$'s are:

$$\begin{aligned}X_{-\tau_1} &= X_{-1} = 1\\ X_{-\tau_2} &= X_{-3} = 0\\ X_{-\tau_3} &= X_{-4} = 1.\end{aligned}$$

The R's are:

$$\begin{aligned}R_0 &= 0\\ R_1 &= \frac{1}{1}\sum_{1\le j\le 1} X_{-\tau_j} = \frac{1}{1} = 1\\ R_2 &= \frac{1}{2}\sum_{1\le j\le 2} X_{-\tau_j} = \frac{1+0}{2} = \frac{1}{2}\\ R_3 &= \frac{1}{3}\sum_{1\le j\le 3} X_{-\tau_j} = \frac{1+0+1}{3} = \frac{2}{3}.\end{aligned}$$

The kappa's are:

$$\begin{aligned}\kappa_1 &= 0\\ \kappa_2 &= 1\\ \kappa_3 &= 1\\ \kappa_4 &= 1\\ \kappa_5 &= 2\\ \kappa_6 &= 2\\ \kappa_7 &= 2\\ \kappa_8 &= 2\\ \kappa_9 &= 3.\end{aligned}$$

The $\hat{R}$'s are:

$$\begin{aligned}\hat{R}_{-1} &= 0\\ \hat{R}_{-2} &= 1\\ \hat{R}_{-3} &= 1\\ \hat{R}_{-4} &= 1\\ \hat{R}_{-5} &= \frac{1}{2}\\ \hat{R}_{-6} &= \frac{1}{2}\end{aligned}$$

$$\begin{aligned}\hat{R}_{-7} &= \frac{1}{2}\\ \hat{R}_{-8} &= \frac{1}{2}\\ \hat{R}_{-9} &= \frac{2}{3.}\end{aligned}$$

Algoet [3] managed to prove that $\hat{R}_{-t}$ converges to $E(X_0|X_{-\infty}^{-1})$ almost surely provided that $E|X_0|$ is finite. For a somewhat weaker result see Györfi et. al. [24]. However none of them was able to prove the integrability of the supremum of the estimates $\hat{R}_{-t}$ in case of unbounded random variables. This missing link was proved by Morvai and Weiss [70] under the condition that

$$E\left(|X_0|\log^+(|X_0|)\right) < \infty.$$

(What is more, we proved that merely having $E|X_0| < \infty$ is not enough, cf. [70].)

For $t > 0$ consider the estimator $\hat{R}_t$ as

$$\hat{R}_1(\omega) = \hat{R}_{-1}(T^1\omega),$$

$$\hat{R}_2(\omega) = \hat{R}_{-2}(T^2\omega)$$

and in general

$$\hat{R}_t(\omega) = \hat{R}_{-t}(T^t\omega)$$

which is defined in terms of

$$(X_0, \ldots, X_{t-1})$$

in the same way as

$$\hat{R}_{-t}(\omega)$$

was defined in terms of

$$(X_{-t}, \ldots, X_{-1}).$$

(T denotes the left shift operator.)

The next example shows how the left shift operator T works. We will use these numerical calculations later.

Example 3.5. *Let*

$$X_0^8(\omega) = (X_0(\omega), X_1(\omega), \ldots, X_7(\omega), X_8(\omega)) = 110111011.$$

Then

$$\begin{aligned}X_{-1}(T^1\omega) &= X_0(\omega) = 1\\ (X_{-2}(T^2\omega), X_{-1}(T^2\omega)) &= (X_0(\omega), X_1(\omega)) = (1,1)\\ (X_{-3}(T^3\omega), X_{-2}(T^3\omega), X_{-1}(T^3\omega)) &= (X_0(\omega), X_1(\omega), X_2(\omega)) = (1,1,0)\\ (X_{-4}(T^4\omega), X_{-3}(T^4\omega), \ldots, X_{-1}(T^4\omega)) &= (X_0(\omega), X_1(\omega), \ldots, X_3(\omega)) = (1,1,0,1)\end{aligned}$$

$$X_{-5}^{-1}(T^5\omega) = X_0^4(\omega) = (1,1,0,1,1)$$
$$X_{-6}^{-1}(T^6\omega) = X_0^5(\omega) = (1,1,0,1,1,1)$$
$$X_{-7}^{-1}(T^7\omega) = X_0^6(\omega) = (1,1,0,1,1,1,0)$$
$$X_{-8}^{-1}(T^8\omega) = X_0^7(\omega) = (1,1,0,1,1,1,0,1)$$
$$X_{-9}^{-1}(T^9\omega) = X_0^8(\omega) = (1,1,0,1,1,1,0,1,1)$$

The next example shows how to calculate the estimator $\hat{R}_t$ for $t = 1, 2, \ldots$. We will use the same data as in the previous example. The numerical calculations in the previous example are useful as auxiliary calculations for the next one.

Example 3.6. *Let*

$$X_0^8(\omega) = (X_0(\omega), X_1(\omega), \ldots, X_7(\omega), X_8(\omega)) = 110111011.$$

Then

$$\hat{R}_1(\omega) = \hat{R}_{-1}(T^1\omega) = R_{\kappa_1}(T^1\omega) = R_0(T^1\omega) = 0$$
$$\hat{R}_2(\omega) = \hat{R}_{-2}(T^2\omega) = R_{\kappa_2}(T^2\omega) = R_1(T^2\omega) = \frac{1}{1} = 1$$
$$\hat{R}_3(\omega) = \hat{R}_{-3}(T^3\omega) = R_{\kappa_3}(T^3\omega) = R_0(T^3\omega) = 0$$
$$\hat{R}_4(\omega) = \hat{R}_{-4}(T^4\omega) = R_{\kappa_4}(T^4\omega) = R_1(T^4\omega) = \frac{0}{1} = 0$$
$$\hat{R}_5(\omega) = \hat{R}_{-5}(T^5\omega) = R_{\kappa_5}(T^5\omega) = R_2(T^5\omega) = \frac{1+0}{2} = \frac{1}{2}$$
$$\hat{R}_6(\omega) = \hat{R}_{-6}(T^6\omega) = R_{\kappa_6}(T^6\omega) = R_2(T^6\omega) = \frac{1+1}{2} = 1$$
$$\hat{R}_7(\omega) = \hat{R}_{-7}(T^7\omega) = R_{\kappa_7}(T^7\omega) = R_1(T^7\omega) = \frac{1}{1} = 1$$
$$\hat{R}_8(\omega) = \hat{R}_{-8}(T^8\omega) = R_{\kappa_8}(T^8\omega) = R_2(T^8\omega) = \frac{0+1}{2} = \frac{1}{2}$$
$$\hat{R}_9(\omega) = \hat{R}_{-9}(T^9\omega) = R_{\kappa_9}(T^9\omega) = R_3(T^9\omega) = \frac{1+0+1}{3} = \frac{2}{3}.$$

The estimator $\hat{R}_t$ may be viewed as an on-line predictor of X_t. This predictor has special significance not only because of potential applications, but additionally because Bailey [4] (cf. B. Ryabko [89] also) proved that it is impossible to construct estimators $\hat{R}_t$ such that always $|\hat{R}_t - E(X_t|X_0^{t-1})| \to 0$ almost surely.

Theorem 3.5. *(Morvai and Weiss [70]) Let $\{X_n\}$ be stationary and ergodic. Assume that $E\left(|X_0|\log^+(|X_0|)\right) < \infty$. Then*

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left|\hat{R}_i - E(X_i|X_0^{i-1})\right| = 0 \quad \textit{almost surely}$$

and

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left|\hat{R}_i - X_i\right| = E\left(\left|E(X_0|X_{-\infty}^{-1}) - X_0\right|\right) \quad \textit{almost surely.}$$

In the above theorem we assumed that X_0 was not merely in L^1 but in $L\log^+ L$. Indeed mere integrability is not ennough, cf. Morvai and Weiss [70]. For results on $\frac{1}{t}\sum_{i=1}^{t}\left|\hat{R}_i - E(X_i|X_0^{i-1})\right|^p$ and $\frac{1}{t}\sum_{i=1}^{t}\left|\hat{R}_i - X_i\right|^p$ under the condition of finite p-th moment of X_0, where $1 < p < \infty$, see Morvai and Weiss [70]. For further reading see Scarpellini [97, 98, 99], Morvai and Weiss [77], Merhav and Feder [50], Jones, Kohler and Walk [35], Felber, Jones, Kohler and Walk [18], Györfi and Lugosi [25], Györfi and Ottucsák [28], Nobel [80] Morvai and Weiss [70], Algoet [1, 2, 3] and Györfi, Ottucsák and Walk [29].

3.2. Pointwise consistent intermittent estimation schemes

Consider the forward estimation problem for countable alphabet first order Markov chains. Ryabko [89] showed that that problem can not be solved.

Theorem 3.6. *(Ryabko [89]) For any estimator $\{\hat{E}_n(X_0^{n-1})\}$ there is a stationary ergodic process $\{X_i\}$ with values from a countable subset of a bounded interval of real numbers such that $\{X_i\}$ is a first order Markov chain and*

$$\limsup_{n\to\infty} |\hat{E}_n(X_0^{n-1}) - E(X_n|X_{n-1})| > 0$$

with positive probability.

(Cf. Györfi, Morvai, and Yakowitz [27] also.)

If one insists on the error criteria then the two ways of getting around the negative results for forward estimation are intermittent schemes – where the estimates are given only at carefully chosen stopping times and restricting to processes with special properties. In this section first we will review results like this for the class of processes where the conditional distribution as a function of the past is continuous on a set of full measure. This class is more general than the processes with continuous conditional probabilities, as we shall see in an example which follows the definition.

Put R^{*-} the set of all one-sided sequences of real numbers, that is,

$$R^{*-} = \{(\ldots, x_{-1}, x_0) : \ x_i \text{ is real } \text{ for all } -\infty < i \le 0\}.$$

Define a metric on sequences $(\ldots, x_{-1}, x_0,)$ and $(\ldots, y_{-1}, y_0)$ as follows. Let

$$d^*((\ldots, x_{-1}, x_0), (\ldots, y_{-1}, y_0)) = \sum_{i=0}^{\infty} 2^{-i-1} \frac{|x_{-i} - y_{-i}|}{1 + |x_{-i} - y_{-i}|}. \tag{3.1}$$

We will consider two-sided stationary real-valued processes $\{X_n\}_{n=-\infty}^{\infty}$. Note that a one-sided stationary time series $\{X_n\}_{n=0}^{\infty}$ can be extended to be a two-sided stationary time series $\{X_n\}_{n=-\infty}^{\infty}$.

Definition 3.1. *The conditional expectation $E(X_1|X_{-\infty}^0)$ is almost surely continuous if for some set $C \subseteq R^{*-}$ which has probability one the conditional expectation $E(X_1|X_{-\infty}^0)$ restricted to this set C is continuous with respect to the metric $d^*(\cdot,\cdot)$ in (3.1).*

Consider any stationary and ergodic finitarily Markovian process $\{X_n\}$ such that the distribution of X_0 concentrates on $\{0, 1, 2, \dots\}$ and $E|X_0| < \infty$. Then obviously $E(X_1|X_{-\infty}^0)$ is almost surely continuous.

Example 3.7. *Consider the Markov chain $\{M_n\}$ with state space $S = \{0, 1, 2\}$ and transition probabilities*

$$P(M_2 = 1|M_1 = 0) = P(M_2 = 2|M_1 = 1) = 1,$$

$$P(M_2 = 0|M_1 = 2) = \frac{9}{10},$$

$$P(M_2 = 1|M_1 = 2) = \frac{1}{10}.$$

This yields a stationary and ergodic process $\{M_n\}$. Let

$$X_n = I_{\{M_n=0\}}.$$

The resulting time series $\{X_n\}$ will not be Markov of any order but it will be finitarily Markovian. The conditional expectation

$$E(X_1|X_{-\infty}^0) = P(X_1 = 1|X_{-\infty}^0)$$

takes values from the set $\{0, \frac{9}{10}\}$. If $X_0 = 1$ then it is zero. Otherwise its value depends solely on whether until the first (going backwards) occurrence of one you see an even or odd number of zeros. The conditional expectation $E(X_1|X_{-\infty}^0)$ is almost surely continuous, but it is not continuous on the whole space since it can not be made continuous at $X_{-\infty}^0 = (\dots, 0, 0, 0)$.

In the previous example X_0 was a binary random variable. In the next example X_0 will be uniformly distributed on the unit interval.

Example 3.8. *A transformation S will be defined on the unit interval. Consider the binary expansion r_1^∞ of each real-number $r \in [0, 1)$, that is, $r = \sum_{i=1}^\infty r_i 2^{-i}$. When there are two expansions, use the representation which contains finitely many $1's$. Now let*

$$\tau(r) = \min\{i > 0 : r_i = 1\}.$$

Notice that, aside from the exceptional set $\{0\}$, which has Lebesgue measure zero τ is finite and well-defined on the closed unit interval. The transformation is defined by

$$Sr = \begin{cases} r - 2^{-\tau(r)} & \text{if } \tau(r) = 1 \\ r - 2^{-\tau(r)} + \sum_{l=1}^{\tau(r)-1} 2^{-l} & \text{if } \tau(r) > 1. \end{cases}$$

All iterations S^k of S for $-\infty < k < \infty$ are well defined and invertible with the exception of the set of dyadic rationals which has Lebesgue measure zero. Now choose r uniformly on the unit interval. Set $X_0(r) = r$ and put $X_n(r) = S^n r$. The process $\{X_n\}$ is a stationary and ergodic first order Markov chain with conditional expectation $E(X_1|X_0 = x) = Sx$, (one observation determines the whole orbit of the process) cf. [27]. Since S is a continuous mapping disregarding

the set of dyadic rationals, the resulting conditional expectation is almost surely continuous. However, the conditional expectation is not continuous on the whole unit interval, since it can not be made continuous at e.g. 0.5.

Example 3.9. *Consider the binary periodic Markov chain* $\{M_n\}$ *which alternates between the states, that is, let*

$$P\left(M_1 = 1 | M_0 = 0\right) = P\left(M_1 = 0 | M_0 = 1\right) = 1.$$

This yields a stationary and ergodic process with marginal probabilities

$$P\left(M_0 = 1\right) = P\left(M_0 = 0\right) = \frac{1}{2}.$$

Let Z_n *be independent identically distributed with uniform distribution on* $(0, 1)$. *We assume that the* $\{Z_n\}$ *process is independent from the* $\{M_n\}$ *process. Now let*

$$X_n = M_n + Z_n.$$

Clearly, the $\{X_n\}$ *process is also stationary and ergodic. The conditional expectation*

$$E(X_1 | X_{-\infty}^0) = \begin{cases} \frac{3}{2} & \text{if } X_0 < 1 \\ \frac{1}{2}. & \text{if } X_0 > 1. \end{cases}$$

is almost surely continuous with respect to the metric $d^*(\cdot, \cdot)$ *in (3.1) even though*

$$\lim_{r \to 1^-} E\left(X_1 | X_0 = r\right) = \frac{3}{2} \neq \frac{1}{2} = \lim_{r \to 1^+} E\left(X_1 | X_0 = r\right).$$

(The event $\{X_0 = 1\}$ *occurs with probability zero and this event can be excluded.)*

The conditional expectation in the next example is not almost surely continuous with respect to the metric $d^*(\cdot, \cdot)$ in (3.1).

Example 3.10. *Consider the binary aperiodic Markov chain* $\{M_n\}$ *whith transition probabilities*

$$P\left(M_1 = 1 | M_0 = 0\right) = P\left(M_1 = 0 | M_0 = 1\right) = \frac{1}{10}$$

and

$$P\left(M_1 = 0 | M_0 = 0\right) = P\left(M_1 = 1 | M_0 = 1\right) = \frac{9}{10}.$$

This yields a stationary and ergodic process with marginal probabilities

$$P\left(M_0 = 1\right) = P\left(M_0 = 0\right) = \frac{1}{2}.$$

Let $\{Z_n\}$ *be independent and identically distributed with*

$$P(Z_n = 2^{-k}) = 2^{-k}$$

for $k = 1, 2, \ldots$. *Let*

$$m = E(Z_1).$$

We assume that the $\{Z_n\}$ *process is independent from the* $\{M_n\}$ *process. Now let*

$$X_n = M_n \cdot Z_n.$$

Obviously, the $\{X_n\}$ *process is also stationary and ergodic. The conditional expectation is*

$$E(X_1|X_{-\infty}^0) = \begin{cases} \frac{m}{10} & \text{if } X_0 = 0 \\ \frac{9m}{10}. & \text{if } X_0 = 2^{-k} \text{ for some } k = 1, 2, \ldots. \end{cases}$$

Now we argue by contradiction. Assume there exists

$$C \subseteq \{(\ldots, x_{-1}, x_0): \quad x_i \in \{0, 2^{-1}, 2^{-2}, \ldots\} \quad \text{for all } -\infty < i \leq 0\}.$$

such that $P(X_{-\infty}^0 \in C) = 1$ *and on* C *the conditional expectation* $E(X_1|X_{-\infty}^0)$ *is given as above and the conditional expectation* $E(X_1|X_{-\infty}^0)$ *is continuous on* C *with respect to the metric* $d^*(\cdot,\cdot)$ *in (3.1). Since* $P(X_0 = 0) > 0$ *there must be a sequence*

$$(\ldots, x_{-2}, x_{-1}, 0)$$

in C. *Since for any* $k = 1, 2, \ldots$, $P(X_0 = 2^{-k}) > 0$ *and since any word formed by the letters* $\{0, 2^{-1}, 2^{-2}, \ldots\}$ *has positive probability, there is a sequence*

$$(\ldots, y_{-k-2}^{(k)}, y_{-k-1}^{(k)}, x_{-k}, \ldots, x_{-2}, x_{-1}, 2^{-k})$$

in C. *(The* y*'s depend on* k.*) Obviously,*

$$d^*((\ldots, x_{-2}, x_{-1}, 0), (\ldots, y_{-k-2}^{(k)}, y_{-k-1}^{(k)}, x_{-k}, \ldots, x_{-2}, x_{-1}, 2^{-k})) \to 0.$$

But

$$\lim_{k\to\infty} E\left(X_1|X_0 = 2^{-k}\right) = \frac{9m}{10} \neq \frac{m}{10} = E\left(X_1|X_0 = 0\right).$$

This is a contradiction. Thus the conditional expectation $E(X_1|X_{-\infty}^0)$ *is not almost surely continuous with respect to the metric* $d^*(\cdot,\cdot)$ *in (3.1).*

The conditional expectation in the next example will not be almost surely continuous with respect to the metric $d^*(\cdot,\cdot)$ in (3.1).

Example 3.11. *Consider the binary aperiodic Markov chain* $\{M_n\}$ *whith transition probabilities*

$$P\left(M_1 = 1|M_0 = 0\right) = P\left(M_1 = 0|M_0 = 1\right) = \frac{2}{10}$$

and

$$P\left(M_1 = 0|M_0 = 0\right) = P\left(M_1 = 1|M_0 = 1\right) = \frac{8}{10}.$$

This yields a stationary and ergodic process with marginal probabilities

$$P\left(M_0=1\right)=P\left(M_0=0\right)=\frac{1}{2}.$$

Let $\{Z_n\}$ *be independent and identically distributed with uniform distribution on the interval* $(1,2)$*. We assume that the* $\{Z_n\}$ *process is independent from the* $\{M_n\}$ *process. Now let*

$$X_n=\left(Z_n\right)^{M_n}.$$

Obviously, the $\{X_n\}$ *process is also stationary and ergodic. The conditional expectation is*

$$E(X_1|X_{-\infty}^0)=\begin{cases}\frac{22}{20} & \text{if } X_0=1\\ \frac{28}{20}. & \text{if } 1<X_0<2.\end{cases}$$

Now we argue by contradiction. Assume there exists

$$C\subseteq\{(\ldots,x_{-1},x_0):\quad 1\le x_i<2\ \text{ for all } -\infty<i\le 0\}$$

such that $P(X_{-\infty}^0\in C)=1$ *and on* C *the conditional expectation* $E(X_1|X_{-\infty}^0)$ *is given as above and the conditional expectation* $E(X_1|X_{-\infty}^0)$ *is continuous on* C *with respect to the metric* $d^*(\cdot,\cdot)$ *in (3.1). Since* $P(X_0=1)>0$ *there must be a sequence*

$$(\ldots,x_{-2},x_{-1},1)$$

in C*. Since for any* $0<\epsilon_k\to 0$*,*

$$P(1<X_0<1+\epsilon_k,|X_{-i}-x_{-i}|<\epsilon_k\ \text{ for all } 1\le i\le k)>0,$$

for each k *there exists a sequence*

$$(\ldots,y_{-2}^{(k)},y_{-1}^{(k)},y_0^{(k)})$$

in C *such that* $1<y_0^{(k)}<1+\epsilon_k$ *and for all* $1\le i\le k$*,* $|y_{-i}^{(k)}-x_{-i}|<\epsilon_k$*. Obviously,*

$$d^*((\ldots,x_{-2},x_{-1},1),(\ldots,y_{-2}^{(k)},y_{-1}^{(k)},y_0^{(k)}))\to 0.$$

But

$$E\left(X_1|X_0=1\right)=\frac{22}{20}\neq\frac{28}{20}=\lim_{k\to\infty}E\left(X_1|X_0=y_0^{(k)}\right).$$

This is a contradiction. Thus the conditional expectation $E(X_1|X_{-\infty}^0)$ *is not almost surely continuous with respect to the metric* $d^*(\cdot,\cdot)$ *in (3.1).*

The conditional expectation in the next example is not almost surely continuous with respect to the metric $d^*(\cdot,\cdot)$ in (3.1). This is not immediately evident but a detailed proof can be found in our paper [62].

Example 3.12. *Define a Markov chain* $\{M_n\}$ *on the nonnegative integers. Let the transition probabilities be as follows.*

$$P(M_1=0|M_0=0)=P(M_1=1|M_0=0)=P(M_1=0|M_0=1)=2^{-1}$$

and for $i = 2, 3, \ldots,$ *let*

$$P(M_1 = i | M_0 = 1) = 2^{-i} \text{ and } P(M_1 = 0 | M_0 = i) = 1.$$

All other transitions happen with probability zero. This Markov chain yields a stationary and ergodic time series. Define the function h *as*

$$h(0) = 0,$$

$$h(1) = 1$$

and for $i \geq 2$ *put*

$$h(i) = \frac{2^{-2^i}}{2}.$$

Let $X_n = h(M_n)$. *Since* $h(\cdot)$ *is one to one,* $\{X_n\}$ *is also a stationary and ergodic Markov chain. The conditional expectation* $E(X_1 | X_{-\infty}^0)$ *is not almost surely continuous with respect to the metric* $d^*(\cdot, \cdot)$ *in (3.1).)*

However the conditional expectation in the next example is almost surely continuous with respect to the metric $d^*(\cdot, \cdot)$ in (3.1).

Example 3.13. *Consider the Markov chain* $\{M_n\}$ *with countably infinite state space* $S = \{0, 1, \frac{1}{2}, \frac{1}{3}, \ldots\}$ *and transition probabilities*

$$P\left(M_1 = 1 | M_0 = 0\right) = P\left(M_1 = 0 | M_0 = 0\right) = \left(\frac{1}{2}\right)$$

and for $n \in \{1, 2, 3, \ldots\}$

$$P\left(M_1 = \frac{1}{n+1} | M_0 = \frac{1}{n}\right) = \left(\frac{1}{2}\right)^{n+1},$$

$$P\left(M_1 = 0 | M_0 = \frac{1}{n}\right) = 1 - \left(\frac{1}{2}\right)^{n+1}.$$

This yields a stationary and ergodic real-vaued process $\{M_n\}$ *(the distribution of which concentrates on* S *and it is a first order Markov chain). The conditional expectation*

$$E(M_1 | M_{-\infty}^0) = \begin{cases} \frac{1}{2} & \text{if } M_0 = 0 \\ \left(\frac{1}{2}\right)^{n+1} \frac{1}{(n+1)} & \text{if } M_0 = \frac{1}{n} \text{ for } n = 1, 2, 3, \ldots. \end{cases}$$

is almost surely continuous with respect to the metric $d^*(\cdot, \cdot)$ *in (3.1) even though*

$$\lim_{k \to \infty} E\left(M_1 | M_0 = \frac{1}{k}\right) = 0 \neq \frac{1}{2} = E(M_1 | M_0 = 0).$$

To see that the conditional expectation $E(M_1 | M_{-\infty}^0)$ *is almost surely continuous observe that any fixed element of the past* $m_{-\infty}^0$ *which does not consist of all*

zeros must contain some value of n for which $m_{-n} \neq 0$. (The all zero case can be excluded since it has probability zero.) Let n_0 be the least n for which $m_{-n} \neq 0$. Say that $m_{-n_0} = 1/k$. If $n_0 = 0$ we are ready since $1/k$ can be approximated only by $1/k$. If $n_0 > 0$ this implies that $m_0 = 0$. Since all points of the state space different from zero are discrete this means that any sequence of points $s^0_{-\infty}$ which are sufficiently close to $m^0_{-\infty}$ must also satisfy $s_{-n_0} = 1/k$. But if for a sequence $s^0_{-\infty}$, $s_{-n_0} = 1/k$ then s_0 is either 0 or $s_0 \geq \frac{1}{k+n_0} > 0$. In the first case the conditional expectations agree, in the second case $s^0_{-\infty}$ can not be chosen to be arbitrarily close to $m^0_{-\infty}$. In this case the conditional expectations are not required to be close together.

Observe that both stationary and ergodic processes, $\{X_n\}$ in Example 3.12 and $\{M_n\}$ in Example 3.13, take values from a countable subset of the unit interval. Both of them are first order Markov chains. However the conditional expectation $E(M_1|M^0_{-\infty})$ in Example 3.13 is almost surely continuous whereas $E(X_1|X^0_{-\infty})$ in Example 3.12 is not (with respect to the metric $d^*(\cdot,\cdot)$ in (3.1)).

Now we will review an algorithm which will successfully estimate the conditional expectation of the next output (at time $n+1$) given the observations up to time n at carefully selected time instances n in case the process has almost surely continuous conditional expectations.

Define the nested sequence of partitions $\{\mathcal{P}_k\}_{k=0}^{\infty}$ of the real line as follows. Let

$$\mathcal{P}_k = \{[i2^{-k}, (i+1)2^{-k}) \ : \ \text{for } i = 0, 1, -1, 2, -2, \dots\}.$$

Let $x \to [x]^k$ denote the quantizer that assigns to any point x the unique interval in $\mathcal{P}_k$ that contains x. Let $[X_m^n]^k = ([X_m]^k, \dots, [X_n]^k)$.

We define the stopping times $\{\lambda_n\}$ along which we will estimate. Set $\lambda_0 = 0$. For $n = 1, 2, \dots$, define λ_n recursively. Let

$$\lambda_1 = \min\{t > 0 : [X_t]^1 = [X_0]^1\}.$$

Note that $\lambda_1 \geq 1$ and it is a stopping time on $[X_0^\infty]^1$. The first estimate m_1 is defined as

$$m_1 = X_1.$$

Let

$$\lambda_2 = \lambda_1 + \min\{t > 0 : [X_t^{\lambda_1+t}]^2 = [X_0^{\lambda_1}]^2\}.$$

Note that $\lambda_2 \geq 2$ and it is a stopping time on $[X_0^\infty]^2$. The second estimate m_2 is defined as

$$m_2 = \frac{X_1 + X_{\lambda_1+1}}{2}.$$

In general, let

$$\lambda_n = \lambda_{n-1} + \min\{t > 0 : [X_t^{\lambda_{n-1}+t}]^n = [X_0^{\lambda_{n-1}}]^n.$$

Note that $\lambda_n \geq n$ and it is a stopping time on $[X_0^\infty]^n$. The nth estimate m_n is defined as

$$m_n = \frac{1}{n}\sum_{j=0}^{n-1} X_{\lambda_j+1}.$$

This estimator can be viewed as a sampled version of the predictor in Morvai et al. [56], Weiss [104], Algoet [3]. (For the discrete case cf. Morvai [54] and Morvai and Weiss [57].)

Theorem 3.7 (Morvai and Weiss [62]). *Let $\{X_n\}$ be a real-valued stationary time series with $E(|X_0|^2) < \infty$. Then*

$$\lim_{n\to\infty} \left| m_n - E(X_{\lambda_n+1}|[X_0^{\lambda_n}]^n) \right| = 0$$

almost surely. If in addition the conditional expectation $E(X_1|X_{-\infty}^0)$ is almost surely continuous then

$$\lim_{n\to\infty} \left| m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n}) \right| = 0$$

almost surely.

Notice that the difference between the first and second statement in the theorem above is the quantization in the condition part of the conditional expectation. While the error $\left| m_n - E(X_{\lambda_n+1}|[X_0^{\lambda_n}]^n) \right|$ tends to zero almost surely for all real-valued stationary time series with $E(|X_0|^2) < \infty$, the error $\left| m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n}) \right|$ does not. E.g. for the stationary and ergodic Markov chain $\{X_n\}$ in Example 3.12 the error $\left| m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n}) \right|$ does not tend to zero with positive probability, cf. Morvai and Weiss [62]. (Of course, the conditional expectation $E(X_1|X_{-\infty}^0)$ for this counterexample process is not almost surely continuous with respect to the metric $d^*(\cdot,\cdot)$ in (3.1).) It turns out that the problem is caused by the quantization. If one knows in advance that the distribution of X_0 concentrates on finite or countably infinite subset of the real line then one may omit the partition $\mathcal{P}_k$ and the quantizer $[\cdot]^k$ entirely and so eliminate this problem. (Cf. Morvai and Weiss [62].)

Example 3.14. *Let $X_0^6 = (X_0, X_1, \ldots, X_5, X_6) = 0100101$. The λ's are:*

$$\begin{aligned} \lambda_0 &= 0 \\ \lambda_1 &= 2 \\ \lambda_2 &= 5. \end{aligned}$$

The $X_{\lambda+1}$'s are:

$$\begin{aligned} X_{\lambda_0+1} &= X_1 = 1 \\ X_{\lambda_1+1} &= X_3 = 0 \\ X_{\lambda_2+1} &= X_6 = 1. \end{aligned}$$

The m's are:

$$m_1 = X_{\lambda_0+1} = X_1 = 1$$

$$m_2 = \frac{X_{\lambda_0+1}+X_{\lambda_1+1}}{2} = \frac{X_1+X_3}{2} = \frac{1+0}{2} = \frac{1}{2}$$
$$m_3 = \frac{X_{\lambda_0+1}+X_{\lambda_1+1}+X_{\lambda_2+1}}{3} = \frac{X_1+X_3+X_6}{3} = \frac{1+0+1}{3} = \frac{2}{3}.$$

One of the drawbacks of this scheme is that the growth of the stopping times $\{\lambda_k\}$ is rather rapid.

Theorem 3.8. *(Morvai [54]) Let $\{X_n\}$ be a stationary and ergodic binary time series. Suppose that $H > 0$ where H denotes the entropy rate associated with the time series $\{X_n\}$. Let $0 < \epsilon < H$ be arbitrary. Then for k large enough,*

$$\lambda_k(X_0^\infty) \geq c^{c^{\cdot^{\cdot^{c}}}} \quad \textit{almost surely,}$$

where the height of the tower is $k-l$, $l(X_0^\infty)$ is a finite number which depends on X_0^∞, and $c = 2^{H-\epsilon}$.

Remark 3.1. *It is an OPEN PROBLEM if there is a better sequence of stopping times $\hat{\lambda}_n$ the growth of which is less rapid with estimator $\hat{e}_n(X_0, X_1, \ldots, X_{\hat{\lambda}_n})$ such that for all stationary and ergodic binary processes*

$$\lim_{n\to\infty} \left| \hat{e}_n - E(X_{\hat{\lambda}_n+1} | X_0^{\hat{\lambda}_n}) \right| = 0$$

almst surely.

At the end of the present section we will review an intermittent scheme where the stopping times grow less rapidly, but that scheme is not designed to succeed for all discrete valued processes.

From the proof of Bailey [4], Ryabko [89], Györfi, Morvai, Yakowitz [27] it is clear that even for the class of all stationary and ergodic binary time series with almost surely continuous conditional expectation $E(X_1|X_{-\infty}^0)$ one can not estimate $E(X_{n+1}|X_0^n)$ for all n in a pointwise consistent way. However, if one considers only a very narrow class of processes then one can succeed for all time instances.

Schäfer [100] considered stationary and ergodic Gaussian processes. He constructed an algorithm which can estimate the conditional expectation for every time instance n for an extremely restricted and narrow class of Gaussian processes. Note that if you want to estimate in time average (or Cesaro average) the problem becames much easier, cf. Györfi and Lugosi [25], Biau et. al. [7].

We consider stationary Gaussian (not necessarily ergodic) processes and estimate the conditional mean along a stopping time sequence for a much wider class of processes than in Schäfer [100].

Consider a stationary Gaussian process $\{X_n\}$ with autocovariance function $\gamma(k) = E(X_{n+k}X_n)$ and $EX_n = m$. Define the following subclasses of stationary Gaussian processes: In Φ_1 we have Gaussian processes satisfying the condition

$$\sum_{j=0}^{\infty} |\gamma(j)| < \infty \tag{3.2}$$

and are not Markovian of any order. In Φ_2 we have all Gaussian processes (not necessarily satifying (3.2)) which are Markov of some order. We are going to deal with processes in

$$\Phi = \Phi_1 \cup \Phi_2.$$

Although estimating the conditional mean in the class Φ_2 is much easier, our algorithm will be valid universally for every process in Φ.

Example 3.15. *Consider the class of Gaussian processes given by*

$$X_n = \sum_{j=0}^{\infty} \psi_j \epsilon_{n-j} + m,$$

where $\psi_0 = 1$, $\sum_{j=0}^{\infty} |\psi_j| < \infty$ *and* ϵ_i*-s are independent and identically distributed Gauss innovations distributed as* $N(0, \sigma)$. *Then condition (3.2) is satisfied and* $\{X_n\}$ *is a real-valued stationary and ergodic Gaussian process in* Φ, *see Hida and Hitsuda [33].*

Schäfer [100] investigated the restricted model class considered in the following example.

Example 3.16. *Consider the model class described in Example 3.15 with the very strong additional condition that the Taylor coefficients of*

$$\frac{1}{\psi(z)} = \sum_{k=0}^{\infty} \varphi_k z^k \qquad (|z| > 1)$$

satisfy

$$\sum_{k=d_n+1}^{\infty} |\varphi_k|^2 \le \left(\frac{C}{\log n}\right)^r \tag{3.3}$$

for sufficiently large n *with some* $C > 0$ *and* $r > 1$, *where* $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ *is the transfer function for* $|z| < 1$.

Theorem 3.9. *(Schäfer [100]) For any stationary Gaussian process from the model class defined in the above example, and the estimator* $\hat{E}_n(X_0^{n-1})$ *defined in Schäfer [100]*

$$\lim_{n\to\infty} \left| \hat{E}_n(X_0^{n-1}) - E(X_n | X_0^{n-1}) \right| = 0$$

almost surely.

For general Gaussian processes it is hard to check condition (3.3). Two special extremely narrow classes of Gaussian processes have been given in Schäfer [100] where this condition is satisfied.

At the beginning of this section we suggested an algorithm and sequence of stopping times along which the error tends to zero almost surely under the condition that the conditional expectation $E(X_1 | \dots, X_{-1}, X_0)$ is almost surely

continuous. Unfortunately the conditional expectation $E(X_1|\dots,X_{-1},X_0)$ is not almost surely continuous in the Gaussian case in general and so this result is not applicable for Gaussian processes in general, cf. Molnár-Sáska and Morvai [52]. We note that for Gauss-Markov processes the conditional expectation $E(X_1|X_{-\infty}^0)$ is continuous. Now we consider an extension of the algorithm discussed in at the beginning of this section.

Now consider the special nested sequence of partitions $\mathcal{P}_k$ of the real line as follows. Let

$$\mathcal{P}_k = \{[i2^{-(k+1)^3}, (i+1)2^{-(k+1)^3}) : \text{ for } i = 0, 1, -1, \dots\}.$$

The choice of $\mathcal{P}_k$ in such form has technical reasons, see [52]. Consider the same sequence of stopping times λ's and estimators m's using this sequence of $\mathcal{P}$'s.

Theorem 3.10. *(Molnár-Sáska and Morvai [52]) For any stationary Gaussian process from the model class* Φ*,*

$$\lim_{n\to\infty} \left| m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n}) \right| = 0$$

almost surely.

This estimator is also consistent for (not Gaussian) stationary processes with almost surely contionuous conditional expectations. For more on estimation for Gaussian processes see Györfi and Lugosi[25] and Biau et. al. [7]. Note that it is still unknown if one can estimate the conditional expectation for all n for all stationary and ergodic Gaussian processes.

Remark 3.2. *It is an OPEN PROBLEM if there is an estimator* $\{\hat{E}_n(X_0^{n-1})\}$ *such that for all stationary and ergodic Gaussian processes*

$$\lim_{n\to\infty} |\hat{E}_n(X_0^{n-1}) - E(X_n|X_0^{n-1})| = 0$$

almst surely.

(Cf. Györfi, Morvai, and Yakowitz [27] and Györfi and Sancetta [30].)

Now we will consider stationary real-valued (not necessarilily Gaussian) processes $\{X_n\}$. We will review a sequence of stopping times which grows slower than the previous ones.

Let $\{\mathcal{P}_k\}_{k=0}^{\infty}$ denote a nested sequence of finite or countably infinite partitions of the real line by intervals. Let $x \to [x]^k$ denote a quantizer that assigns to any point x the unique interval in $\mathcal{P}_k$ that contains x. For a set C of real numbers let $\text{diam}(C) = \sup_{y,z\in C} |z-y|$. We assume that

$$\lim_{k\to\infty} \text{diam}([x]^k) = 0 \quad \text{for all real number } x.$$

Let $[X_m^n]^k = ([X_m]^k, \dots, [X_n]^k)$. Let $1 \le l_k \le k$ be a nondecreasing sequence of positive integers such that $\lim_{k\to\infty} l_k = \infty$.

Define the stopping times as follows. Set $\zeta_0 = 0$. For $k = 1, 2, \ldots$, define the sequences η_k and ζ_k recursively. Each step we refine the quantization, and slowly increase the block length of the next repetition, as follows: let

$$\eta_1 = \min\{t > 0 : [X_t]^1 = [X_0]^1\}$$

and

$$\zeta_1 = \zeta_0 + \eta_1.$$

One denotes the estimate of $E(X_{\zeta_1+1}|X_0^{\zeta_1})$ by g_1, and defines it to be

$$g_1 = X_1.$$

Let

$$\eta_2 = \min\{t > 0 : [X_{\zeta_1-(l_2-1)+t}^{\zeta_1+t}]^2 = [X_{\zeta_1-(l_2-1)}^{\zeta_1}]^2\}$$

and

$$\zeta_2 = \zeta_1 + \eta_2.$$

One denotes the estimate of $E(X_{\zeta_2+1}|X_0^{\zeta_2})$ by g_2, and defines it to be

$$g_2 = \frac{X_1 + X_{\zeta_1+1}}{2}.$$

In general, let

$$\eta_k = \min\{t > 0 : [X_{\zeta_{k-1}-(l_k-1)+t}^{\zeta_{k-1}+t}]^k = [X_{\zeta_{k-1}-(l_k-1)}^{\zeta_{k-1}}]^k\}$$

and

$$\zeta_k = \zeta_{k-1} + \eta_k.$$

One denotes the kth estimate of $E(X_{\zeta_k+1}|X_0^{\zeta_k})$ by g_k, and defines it to be

$$g_k = \frac{1}{k}\sum_{j=0}^{k-1} X_{\zeta_j+1}.$$

Example 3.17. *Let $[\cdot]^k$ be the quantizer*

$$[x]^k = \begin{cases} 0 & \text{if } -2^{-k} < x < 2^{-k} \\ -i2^{-k} & \text{if } -(i+1)2^{-k} < x \le -i2^{-k} \text{ for some } i = 1, 2, \ldots \\ i2^{-k} & \text{if } i2^{-k} \le x < (i+1)2^{-k} \end{cases}$$

and let $l_k = k$. Let

$$(X_0, X_1, \ldots, X_5, X_6) = 0100101.$$

The ζ's and η's are:

$$\begin{aligned} \zeta_0 &= 0 \\ \eta_1 &= 2 \\ \zeta_1 &= 2 \end{aligned}$$

$$\begin{aligned}\eta_2 &= 3\\ \zeta_2 &= 5.\end{aligned}$$

The $X_{\zeta+1}$'s are:

$$\begin{aligned}X_{\zeta_0+1} &= X_1 = 1\\ X_{\zeta_1+1} &= X_3 = 0\\ X_{\zeta_2+1} &= X_6 = 1.\end{aligned}$$

The g's are:

$$\begin{aligned}g_1 &= \frac{1}{1}\sum_{j=0}^{0} X_{\zeta_j+1} = \frac{1}{1} = 1\\ g_2 &= \frac{1}{2}\sum_{j=0}^{1} X_{\zeta_j+1} = \frac{1+0}{2} = \frac{1}{2}\\ g_3 &= \frac{1}{3}\sum_{j=0}^{2} X_{\zeta_j+1} = \frac{1+0+1}{3} = \frac{2}{3}.\end{aligned}$$

The next theorem states the strong (pointwise) consistency of the estimator.

Theorem 3.11. *(Morvai and Weiss [58]) Let $\{X_n\}$ be a real-valued stationary time series with $E(|X_0|^2) < \infty$. Then*

$$\lim_{k\to\infty} \left| g_k - E(X_{\zeta_k+1}|X_0^{\zeta_k}) \right| = 0 \quad \textit{almost surely}$$

provided that the conditional expectation $E(X_1|X_{-\infty}^0)$ is almost surely continuous.

The consistency holds independently of how the sequence l_k and the partitions are chosen as long as l_k goes to infinity and the partitions become finer. However, the choice of these sequences has a great influence on the growth of the stopping times.

From the proof of [4], [89] and [27] it is clear that even for the class of all stationary and ergodic binary time series with almost surely continuous conditional expectation $E(X_1|\ldots,X_{-1},X_0)$ one can not estimate $E(X_{n+1}|X_0^n)$ for all n strongly (pointwise) consistently.

The stationary processes with almost surely continuous conditional expectation generalize the processes for which the conditional expectation is actually continuous. (Cf. [36] or [40].)

If one uses finite partitions then it is possible to give an upper bound on the growth of the stopping times $\{\zeta_k\}$. Let $\mathcal{P}_k$ be a nested sequence of finite partitions of the real line by intervals. If for some $\epsilon > 0$,

$$\sum_{k=1}^{\infty}(k+1)2^{-l_k\epsilon} < \infty$$

then for the stopping time ζ_k

$$\zeta_k < |\mathcal{P}_k|^{l_k} 2^{l_k \epsilon}$$

eventually almost surely, (cf. Morvai and Weiss [58], Algoet [3] and Morvai et. al. [55]).

Example 3.18. Consider $\epsilon = 1$, $l_k = \lfloor 4 \log_2(k+1) \rfloor$, and $|\mathcal{P}_k| = k+1$. Then

$$\zeta_k < (k+1)^{4(1+\log_2(k+1))}$$

which has a little bit faster growth than polynomial.

In case of finite alphabet processes you can achieve a slightly better upper bound. Indeed, let H denote the entropy rate associated with the stationary and ergodic finite alphabet time series $\{X_n\}$. Note that in this case no quantization is needed. Then

$$\zeta_k < 2^{l_k(H+\epsilon)}$$

eventualy almost surely provided that $(k+1)2^{-l_k \epsilon}$ is summable. (Cf. [57], [85], [55].)

References

[1] Algoet, P. (1992). Universal schemes for prediction, gambling and portfolio selection. *Annals of Probability.* **20** 901–941. Correction: *ibid.* vol. 23, pp. 474–478, 1995. MR1159579

[2] Algoet,P. (1994) The strong low of large numbers for sequential decisions under uncertainity. *IEEE Transactions on Information Theory* **40** 609–634. MR1295308

[3] Algoet, P (1999) Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Transactions on Information Theory*, **45** 1165–1185. MR1686250

[4] Bailey, D.H. (1976) *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph. D. thesis, Stanford University.

[5] Berger, A. (1951) On uniformly consistent tests. *Ann. Math. Statistics* **22** 289–293. MR0042653

[6] Berti, P. Crimaldi,I. Pratelli, L. and Rigo,P. (2009) Rate of convergence of predictive distributions for dependent data *Bernoulli* **15** 1351–1367. MR2597596

[7] Biau, G., Bleakley, K. Györfi, L., and Ottucsák, Gy. (2010) Nonparametric sequential prediction of time series, *Journal of Nonparametric Statistics*, **22** 297–317. MR2662595

[8] Breiman, L. (1957) The individual ergodic theorem of information theory, *Annals of Mathematical Statistics*, **28**, 809–811, MR0092710

[9] Breiman, L. (1960) The individual ergodic theorem of information theory: correction, *Annals of Mathematical Statistics*, **31**, 809–810.

[10] Bühlmann, P. and Wyner, A.J. (1999) Variable- length Markov chains. *Annals of Statistics* **27** 480–513. MR1714720

[11] Bunea, F. and Nobel, A. (2008) Sequential procedures for aggregating arbitrary estimators of a conditional mean *IEEE Transactions on Information Theory* **54** 1725–1735. MR2450298

[12] Cover T. (1975). Open Problems in Information Theory. *1975 IEEE-USSR Joint Workshop on Information Theory* 35–36.

[13] Csiszár, I. and Shields, P. (2000) The consistency of the BIC Markov order estimator. *Annals of Statistics.* **28** 1601–1619. MR1835033

[14] Csiszár, I. (2002) Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Transactions on Information Theory* **48** 1616–1628. MR1909476

[15] Csiszár, I. and Talata, Zs. (2006) Context tree estimation for not necessarily finite memory processes via BIC and MDL. *IEEE Transactions on Information Theory* **52** 1007–1016. MR2238067

[16] Cerqueti, R., Falbo, P., and Pelizzari, C. (2017) Relevant states and memory in Markov chain bootstrapping and simulation. *European Journal of Operational Research* **256** 163–177. MR3543093

[17] Dembo, A. and Peres, Y. (1994) A topological criterion for hypothesis testing. *Annals of Stat.* **22** 106–117. MR1272078

[18] Felber, T. Jones, D., Kohler, M., and Walk, H. (2013) Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates. *Journal of Statistical Planning and Inference* **143** 1689–1707. MR3082227

[19] Finesso, L. (1990) *Consistent Estimation of the Order for Markov and Hidden Markov Chains* PhD thesis, University of Maryland.

[20] Finesso, L. (1992) Estimation of the order of a finite Markov chain. *Recent advances in mathematical theory of systems, control, networks and signal processing, I (Kobe, 1991),* 643–645, Mita, Tokyo. MR1197985

[21] Furstenberg, H. (1960) *Stationary Processes and Prediction Theory* Princeton University Press. MR0140151

[22] Gallo, S. and Leonardi F. (2015) Nonparametric statistical inference for the context tree of a stationary ergodic process *Electronic Journal of Statistics.* **9** 2076–2098. MR3397402

[23] Gutman, Y. and Hochman, M. (2008) On processes which cannot be distinguished by finite observation. *Israel J. Math.* **164** 265–284. MR2391149

[24] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002) *A Distribution Free Theory of Nonparametric Regression* Springer-Verlag, New York MR1920390

[25] Györfi, L. Lugosi, G. (2002) Strategies for sequential prediction of stationary time series. *In M. Drop, P. L'Ecuyer, and F. Szidarovszky, editors, Examination of Stochastic Theory, Methods and Applications*, 225–248. Kluwer Academic Publishers. MR1893282

[26] Györfi, L. Lugosi, G. and Morvai, G. (1999) A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transac-*

tions on Information Theory **45** 2642–2650. MR1725166
[27] Györfi, L., Morvai, G. and Yakowitz, S. (1998) Limits to consistent on-line forecasting for ergodic time series. *IEEE Transactions on Information Theory*, **44** 886–892. MR1607704
[28] Györfi, L. and Ottucsák, Gy. (2007) Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory* **53** 1866–1872. MR2317147
[29] Györfi, L., Ottucsák, Gy. and Walk H., (2012) *Machine Learning for Financial Engineering* Imperial College Press, London.
[30] Györfi, L. and Sancetta, S. (2014) An open problem on strongly consistent learning of the best prediction for Gaussian processes. In *Topics in Nonparametric Statistics* 115–136. Springer, New York, NY. MR3333341
[31] Handel, R. (2011) On the minimal penalty for Markov order estimation. *Probability Theory and Related Fields* **150** 709–738. MR2824872
[32] Heller, A. (1965) On Stochastic Processes Derived from Markov Chains. *Annals of Math. Stat.* **36** 1286–1291. MR0176520
[33] Hida, T. and Hitsuda, M. (1993) *Gaussian Processes.* Providence, RI:AMS Translation of Mathematical Monographs, **10** MR1216518
[34] Hoeffding, W. and Wolfowitz, J. (1958) Distinguishability of sets of distributions. *Ann. Math. Statist.* **29** 700–718. MR0095555
[35] Jones, D. Kohler,M. and Walk, H. (2012) Weakly Universally Consistent Forecasting of Stationary and Ergodic Time Series. *IEEE Transactions on Information Theory* **58** 1191–1202. MR2918019
[36] Kalikow, S. (1990) Random Markov processes and uniform martingales. *Israel Journal of Mathematics*, **71** 33–54. MR1074503
[37] Kalikow, S., Katznelson, Y. and Weiss, B. (1992) Finitarily deterministic generators for zero entropy systems. *Israel Journal of Mathematics* **79** 33–45. MR1195252
[38] Kalocinski, D. and Steifer, T. (2019) An Almost Perfectly Predictable Process with No Optimal Predictor *In: 2019 IEEE International Symposium on Information Theory (ISIT)* NEW YORK: IEEE, 2504–2508.
[39] Kalocinski, D. and Steifer, T. (2019) On unstable and unoptimal prediction *Mathematical Logic Quarterly* **65** 218–227.
[40] Keane, M. (1972) Strongly mixing g-measures. *Invent. Math.* **16** 309–324. MR0310193
[41] Keane, M. and Smorodinsky, M. (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. *Ann. of Math.* **109** 397–406. MR0528969
[42] Khaleghi A., Ryabko, D., Mary, J. and Preux P. (2016) Consistent Algorithms for Clustering Time Series *Journal of Machine Learning Research* **3** 1–32. MR3482923
[43] Khudanpur, S. and Narayan, P. (2002) Order Estimation for a Special Class of Hidden Markov Sources and Binary Renewal Processses. *IEEE Transactions on Information Theory* **48** 1704–1713. MR1909484
[44] Kieffer, J. (1993) Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Transactions on Information Theory* **39** 893–902. MR1237719

[45] Kolmogorov, A.N. (1959) Entropy per unit time as a metric invariant of automorphisms. (Russian), *Dokl. Akad. Nauk* SSSR **124** 754–755. MR0103255

[46] Kontoyiannis, I., Algoet, P., Suhov, Yu. M. and Wyner, A.J. (1998) Nonparametric entropy estimation for stationary processes and random fields, with application to English text. *IEEE Transactions on Information Theory* **44** 1319–1327. MR1616653

[47] Kraft, Ch. (1955) Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.* **2**, pp. 125–141. MR0073896

[48] Löcherbach, E. and Orlandi, E.(2011) Neighborhood radius estimation for variable-neighborhood random fields. *Stochastic Process. Appl.* **121** 2151–2185. MR2819245

[49] Maker, Ph.T. (1940). The ergodic theorem for a sequence of functions, *Duke Math. J.*, **6**, 27–30. MR0002028

[50] Merhav, N. and Feder, M. (1998) Universal Prediction *IEEE Transactions on Information Theory* **44** 2124–2147. MR1658815

[51] Merhav, N., Gutman, M. and Ziv, J. (1989) On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Information Theory* **35** 1014–1019. MR1023240

[52] Molnár-Sáska, G. and Morvai, G. (2010) Intermittent Estimation for Gaussian Processes. *IEEE Transactions on Information Theory* **56** 2778–2782. MR2683434

[53] Morvai, G. (1994) Estimation of Conditional Distribution for Stationary Time Series. PhD Thesis, Technical University of Budapest.

[54] Morvai, G. (2003) Guessing the output of a stationary binary time series. In: *Foundations of Statistical Inference, (Eds. Y. Haitovsky, H.R.Lerche, Y. Ritov)* Physika-Verlag 207–215. MR2017826

[55] Morvai, G., Yakowitz, S., and Algoet, M. (1997) Weakly convergent nonparametric forecasting of stationary time series. *IEEE Transactions on Information Theory* **43** 483–498. MR1447529

[56] Morvai, G., Yakowitz, S. and Györfi, L. (1996) Nonparametric inferences for ergodic, stationary time series. *Annals of Statistics.*, **24** 370–379. MR1389896

[57] Morvai, G. and Weiss, B. (2003) Forecasting for stationary binary time series. *Acta Applicandae Mathematicae*, **79** 25–34. MR2021874

[58] Morvai, G. and Weiss, B. (2004) Intermittent estimation of stationary time series. *Test* **13** 525–542. MR2154012

[59] Morvai, G. and Weiss, B. (2005) Prediction for discrete time series. *Probability Theory and Related Fields* **132** 1–12. MR2136864

[60] Morvai, G. and Weiss, B. (2005) Limitations on intermittent forecasting. *Statistics and Probability Letters* **72** 285–290. MR2153125

[61] Morvai, G. and Weiss, B. (2005) On classifying processes. *Bernoulli* **11** 523–532. MR2146893

[62] Morvai, G. and Weiss, B. (2005) Inferring the conditional mean. *Theory of Stochastic Processes* **11** No. 1-2, 112–120. MR2327452

[63] Morvai, G. and Weiss, B. (2005) Order estimation of Markov chains. *IEEE Transactions on Information Theory* **51** 1496–1497. MR2241507

[64] Morvai, G. and Weiss, B. (2005) Forward estimation for ergodic time series. *Ann. I.H.Poincaré Probabilités et Statistiqoes* **41** 859–870. MR2165254

[65] Morvai, G. and Weiss, B. (2007) On estimating the memory for finitarily Markovian processes. *Ann. I.H.Poincaré-PR* **43** 15–30. MR2288267

[66] Morvai, G. and Weiss, B. (2007) On sequential estimation and prediction for discrete time series. *Stochastics and Dynamics* **7** 417–437. MR2378577

[67] Morvai, G. and Weiss, B. (2008) Estimating the Lengths of Memory Words. *IEEE Transactions on Information Theory* **54** 3804–3807. MR2451043

[68] Morvai, G. and Weiss, B. (2008) On universal estimates for binary renewal processes. *Ann. Appl. Probab.* **18** 1970–1992. MR2462556

[69] Morvai, G. and Weiss, B. (2009) Estimating the residual waiting time for binary stationary time series. *ITW 2009. IEEE Information Theory Workshop on Networking and Information Theory 10–12 June 2009* 67–70.

[70] Morvai, G. and Weiss, B. (2011) Nonparametric Sequential Prediction for Stationary Processes. *Annals of Probability* **39**, 1137–1160. MR2789586

[71] Morvai, G. and Weiss, B. (2011) Testing stationary processes for independence. *Ann. I.H.Poincaré-PR* **47** 1219–1225. MR2884232

[72] Morvai, G. and Weiss, B. (2012) A note on prediction for discrete time series. *Kybernetika* **48** 809–823,. MR3013400

[73] Morvai, G. and Weiss, B. (2013) Universal Tests for Memory Words. *IEEE Transactions on Information Theory* **59** 6873–6879. MR3106870

[74] Morvai, G. and Weiss, B. (2014) Inferring the Residual Waiting Time for Binary Stationary Time Series. *Kybernetika* **50** 869–882. MR3301776

[75] Morvai, G. and Weiss, B. (2016) A versatile scheme for predicting renewal times. *Kybernetika* **52** 348–358. MR3532511

[76] Morvai, G. and Weiss, B. (2019) A note on discriminating Poisson processes from other point processes with stationary inter arrival times. *Kybernetika* **55**, 802–808. MR4055577

[77] Morvai, G. and Weiss, B. (2020) Estimating the conditional expectations for continuous time stationary processes. *Kybernetika* **56**, 410–431. MR4131737

[78] Morvai, G. and Weiss, B. (2020) Universal rates for estimating the residual waiting time in an intermittent way. *Kybernetika* **56**, 601–616. MR4168527

[79] A. Nobel, "Limits to classification and regression estimation from ergodic processes," *Annals of Statistics*, vol. 27 pp. 262–273. MR1701110

[80] Nobel, A.B. (2003) On optimal sequential prediction for general processes. *IEEE Transactions on Information Theory* **49** 83–98. MR1965889

[81] Nobel, A. (2006) Hypothesis testing for families of ergodic processes. *Bernoulli* **12** 251–269. MR2218555

[82] Ornstein, D.S. (1978) Guessing the next output of a stationary process. *Israel Journal of Mathematics* **30** 292–296. MR0508271

[83] Ornstein, D.S. (1974) *Ergodic Theory, Randomness, and Dynamical Systems.* Yale University Press. MR0447525

[84] Ornstein, D.S. and Weiss, B. (1990) How sampling reveals a process. *The Annals of Probability* **18** 905–930. MR1062052

[85] Ornstein, D.S. and Weiss, B. (1993) Entropy and data compression schemes. *IEEE Transactions on Information Theory* **39** 78–83. MR1211492

[86] Ornstein, D.S. and Weiss, B. (2007) Entropy is the only finitely observable invariant. *J. Mod. Dyn.* **1** 93–105. MR2261073

[87] Peres, Y. and Shields, P. (2005) Two new Markov order estimators. *arXiv:math/0506080v1 [math.ST]*

[88] Ren, J., Bai, X., Lu, Y.Y., Tang,K., Wang, Y, Reiner, G., and Sun, F. (2018) Alignment-Free Sequence Analysis and Applications *Annual Review of Biomedical Data Science* **1** 93–114.

[89] Ryabko, B. (1988) Prediction of random sequences and universal coding. *Problems of Inform. Trans.* **24** 87–96. MR0955983

[90] Ryabko, B. (2009) Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. *IEEE Transactions on Information Theory* **55** 4309–4315. MR2582884

[91] Ryabko, B. and Astola, J. (2006) Universal codes as a basis for time series testing. *Stat. Methodol.* **3** 375–397. MR2252392

[92] Ryabko, D. (2006) Pattern recognition for conditionally independent data. *Journal of Machine Learning Research* **7** 645–664. MR2274382

[93] Ryabko, D. (2009) An impossibility result for process discrimination. *2009 IEEE International Symposium on Information Theory, VOLS 1- 4* **1-4** 1734–1738.

[94] Ryabko, D. (2010) Discrimination Between B-Processes is Impossible. *Journal of Theoretical Probability* **23** 565–575. MR2274382

[95] Ryabko, D. (2019) *Asymptotic nonparametric statistical analysis of stationary time series*, Springer International Publishing.

[96] Ryabko, D. (2019) On asymptotic and finite-time optimality of Bayesian predictors. *J. Mach. Learn. Res.* **20** 1–24. MR4030163

[97] Scarpellini, B. (1979) Predicting the future of functions on flows, *Math. Systems Theory*, **12**, 281–296. MR0529563

[98] Scarpellini, B. (1979) Entropy and nonlinear prediction, *Probability Theory and Related Fields*, **50**, (2), 165–178. MR0551610

[99] Scarpellini, B. (1981) Conditional expectations of stationary processes. *Z. Wahrsch. Verw. Gebiete* **56** no. 4, 427–441. MR0621658

[100] Schäfer, D. (2002) Strongly Consistent Online Forecasting of Centered Gaussian Processes. *IEEE Transactions on Information Theory* **48** 791–799. MR1889985

[101] Shannon,C.E. (1948) A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656. MR0026286

[102] Takahashi, H. (2011) Computational Limits to Nonparametric Estima-

tion for Ergodic Processes *IEEE Transactions on Information Theory* **57** 6995–6999. MR2882275

[103] Weiss,B. (2005) Some remarks on filtering and prediction of stationary processes. *Israel J. of Math.* **149** 345–360. MR2191220

[104] Weiss, B. (2000) *Single Orbit Dynamics*, American Mathematical Society. MR1727510

[105] Ziv, J. and Lempel, A. (1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* **23** 337–343. MR0530215

[106] Ziv, J. (1978) Coding theorems for individual sequences. *IEEE Transactions on Information Theory* **24** 405–412. MR0504371