ALEA, Lat. Am. J. Probab. Math. Stat. 18, 1643–1667 (2021)
DOI: 10.30757/ALEA.v18-60

Consistency, integrability and asymptotic normality for some intermittent estimators

Gusztáv Morvai and Benjamin Weiss

Alfréd Rényi Institute of Mathematics, 13-15 Reáltanoda utca, H-1053, Budapest, Hungary, and, MTA-BME Stochastics Research Group,, Egry József utca 1, Building H, Budapest, 1111 Hungary
E-mail address: morvai@math.bme.hu
URL: http://math.bme.hu/∼morvai

Hebrew University of Jerusalem, Institute of Mathematics,, Jerusalem 91904 Israel
E-mail address: weiss@math.huji.ac.il
URL: http://mathematics.huji.ac.il/people/benjamin-weiss

Abstract. In this paper we consider universal nonparametric estimators for the conditional expectations of the output of a stationary process at carefully selected time instances (intermittent estimation). These estimators are based on the information provided by the random outputs at past times. Pointwise consistency, integrability of various suprema and asymptotic normality will be established for these nonparametric intermittent estimators.

1. Introduction

The study of statistical estimators has several components. In this paper we would like to present a few results about consistency, integrability and the asymptotic distribution of the estimators themselves.

To describe these results let us begin by reviewing the estimation problem. In a short communication that appeared in the Proceedings of the First International IEEE-USSR Information Workshop (Cover, 1976), Tom Cover formulated a number of problems that have generated a substantial literature.

Cover's first problem was on forward estimation of the conditional probability.

Problem 1 *Is there an estimation scheme $\hat{p}_n$ for the value $P(X_{n+1} = 1|X_0, \dots, X_n)$ such that $\hat{p}_n$ depends solely on the data segment $(X_0, \dots, X_n)$ and*

$$\lim_{n\to\infty} |\hat{p}_n(X_0, \dots, X_n) - P(X_{n+1} = 1|X_0, \dots, X_n)| = 0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

Received by the editors May 13th, 2020; accepted June 4th, 2021.

2010 *Mathematics Subject Classification.* 62G05, 60G25, 60G10.

Key words and phrases. Nonparametric estimation, stationary processes.

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

Notice that in the first problem of Cover the data segment $(X_0,\dots,X_n)$ grows in the positive, forward direction and the goal is to estimate the conditional probability for the next ever changing random variable X_{n+1}.

Cover's second problem was on backward estimation of the conditional probability.

Problem 2 *Is there an estimation scheme $\hat{p}_n$ for the value $P(X_1=1|X_{-n},\dots,X_0)$ such that $\hat{p}_n$ depends solely on the observed data segment $(X_{-n},\dots,X_0)$ and*

$$\lim_{n\to\infty}|\hat{p}_n(X_{-n},\dots,X_0)-P(X_1=1|X_{-n},\dots,X_0)|=0$$

almost surely for all stationary and ergodic binary time series $\{X_n\}_{n=-\infty}^{\infty}$?

Note that in Cover's second problem the data segment $(X_{-n},\dots,X_0)$ grows in the negative, backward direction and the goal is to estimate the conditional probability for a fix random variable X_1.

Observe that we are dealing with pointwise convergence. In the forward estimation problem the target itself, $P(X_{n+1}=1|X_0,\dots,X_n)$, does not converge in the pointwise sense in general, while in the backward estimation problem the target $P(X_1=1|X_{-n},\dots,X_0)$ does converge to $P(X_1=1|\dots,X_{-2},X_{-1},X_0)$ almost surely. It turns out that there is a big difference between these two problems and the answers to the forward and backward problems are far from being the same. Ornstein (1978) gave a rather complicated algorithm for the backward prediction problem whereas Bailey provided a proof for the nonexistence of a universal algorithm guaranteeing almost sure convergence in the forward estimation problem. To do this, Bailey (1976), assuming the existence of a universal algorithm, used Ornstein's technique of cutting and stacking Ornstein (1974) for the construction of a counterexample process for which the algorithm fails to converge (see Shields, 1991 for more details on this method).

The problem came to life again in the late eighties with the work of B. Ryabko (Ryabko, 1988). He used a simpler technique, namely - relabelling a countable state Markov chain, in order to prove the nonexistence of a universal estimator for Cover's first problem (cf. also Györfi et al., 1998, Morvai and Weiss, 2005c and Takahashi, 2011). In addition there was a growing interest in universal algorithms of various kinds in information theory and elsewhere, see Merhav and Feder (1998) and Morvai and Weiss (2007b) for a survey.

Three approaches evolved in an attempt to obtain positive results for the problem of forward estimation in the face of Bailey's theorem.

The first modifies the almost sure convergence to convergence in probability or almost sure convergence of the Cesaro averages. This was done already by Bailey in his thesis. Cf. Algoet (1992, 1994, 1999), Morvai et al. (1996, 1997), Nobel (2003), Györfi and Ottucsák (2007), Györfi et al. (2012) Morvai and Weiss (2011), Morvai and Weiss (2020a), Weiss (2000), Felber et al. (2013), Jones et al. (2012).

The second gives up on trying to estimate the distribution of the next output at all time moments n, and concentrates on guaranteeing prediction only at certain stopping times (intermittent estimation) cf. Morvai (2003), while the third restricts the class of processes for which the scheme is shown to succeed cf. Schäfer (2002) and Morvai and Weiss (2005a).

One may apply more than one of these restrictions cf. Györfi and Lugosi (2002), Morvai and Weiss (2004), Morvai and Weiss (2003), Morvai and Weiss (2005b), Morvai and Weiss (2007a), Morvai and Weiss (2012), Morvai and Weiss (2020b) and Molnár-Sáska and Morvai (2010).

For further reading see Suzuki (2003), D. Ryabko and B. Ryabko (Ryabko and Ryabko, 2015), B. Ryabko and Monarev (Ryabko and Monarev, 2005), B. Ryabko (Ryabko, 2008), D. Ryabko (Ryabko, 2019) and Morvai and Weiss (2021).

In this paper, we will follow the second way, we will estimate only on a sequence of stopping times, i.e. we will estimate in an intermittent way. We will assume that the process is stationary but we will not assume ergodicity. Even though the ergodic decomposition guarantees that almost every sample path will be coming from some ergodic process the probabilities, expectations, et al. that we will be concerned with involve of course the non-ergodic process itself. This means that the general case does not simply reduce to the ergodic case as for example in the pointwise ergodic theorem.

Let $\mathcal{X}$ be a finite or countable alphabet. Let $\{X_n\}_{n=0}^{\infty}$ be a stationary process taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a function that assigns to any letter $x \in \mathcal{X}$ a real number. Assume that $E(|f(X_1)|) < \infty$.

In section 2 we will define increasing stopping times $0 \le \lambda_0 < \lambda_1 < \lambda_2 < \dots$ and estimator m_n depending solely on the data segment $(X_0, \dots, X_{\lambda_n})$ for the value $E(f(X_{\lambda_n+1})|X_0, \dots, X_{\lambda_n})$. It will be shown (mainly using the ideas and techniques in Morvai et al., 1996, Algoet, 1999, Morvai, 2003 and Morvai and Weiss, 2005b) that the estimator is pointwise consistent along these stopping times. Thus, though Bailey (1976) proved that one can not estimate for all n in a pointwise consistent way, at least one can estimate in an intermittent way under very weak conditions. Further assume that $E(|f(X_1)| \log^+(|f(X_1)|)) < \infty$. We will show that both the supremum of the estimator and the supremum of the error are integrable. These results give information about the magnitude of the estimator and the error.

However, our main results will be on limit distributions. Assuming some further conditions we will prove, among others, that the normalized error has asymptotically normal distribution.

Throughout the paper we will give illuminating examples. We put the proofs of the theorems in a separate section (section 3) and we put the auxiliary lemmas in an Appendix.

2. Results

Let $\mathcal{X}$ be a finite or countably infinite alphabet. Let $\{X_n\}_{n=0}^{\infty}$ be a stationary process taking values from $\mathcal{X}$. A one-sided stationary time series $\{X_n\}_{n=0}^{\infty}$ can always be considered to be a two-sided stationary time series $\{X_n\}_{n=-\infty}^{\infty}$.

Now we introduce our algorithm. For notational convenience, let $X_m^n = (X_m, \dots, X_n)$, where $m \le n$.

We define the increasing sequence of stopping times $\{\lambda_n\}_{n=0}^{\infty}$ along which we will estimate. (Cf. Morvai, 2003 and Morvai and Weiss, 2005b.) Set $\lambda_0 = 0$ and define

$$\lambda_1 = \min\{t > 0 : X_t = X_0\}.$$

In general, for $n \ge 2$ define λ_n as

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

Note that λ_n is finite with probability one by the Poincaré recurrence theorem for discrete stationary processes, $\lambda_n \ge n$ by construction and λ_n is a stopping time on X_0^{∞}.

Remark 2.1. Notice that $X_{\lambda_{n+1}-\lambda_n}^{\lambda_{n+1}} = X_0^{\lambda_n}$. This fact will allow us to construct an auxiliary process $\{\tilde{X}_n\}_{n=-\infty}^{0}$ later in (2.1) which will play a key role in the proofs. This process $\{\tilde{X}_n\}_{n=-\infty}^{0}$ will allow us to use backward going techniques (cf. Algoet, 1999) in our forward intermittent case. (Cf. Morvai, 2003 and Morvai and Weiss, 2005b also.)

Let $f : \mathcal{X} \to \Re$ denote a function that assigns to any letter $x \in \mathcal{X}$ a real number. Assume that $E(|f(X_1)|) < \infty$. The goal is to estimate $E(f(X_{\lambda_n+1})|X_0^{\lambda_n})$ from samples $X_0^{\lambda_n}$. For $n \ge 1$ the n-th estimate m_n is defined as

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

Observe that m_n depends solely on $X_0^{\lambda_n}$. (The estimator m_n depends on the values $f(X_{\lambda_j+1})$ for $j=0,\dots,n-1$, but the stopping time λ_j itself depends on $X_0^{\lambda_j}$. We say m_n depends solely on $X_0^{\lambda_n}$ meaning that we can evaluate m_n from the observations $X_0^{\lambda_n}$.)

Example 2.2. Let $\mathcal{X}=\{a,b\}$ and let $f(a)=0$ and $f(b)=1$. Consider

$$X_0^6=(X_0,X_1,\dots,X_5,X_6)=abaabab.$$

The λ's are:

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

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

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

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

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

The m's are:

$$\begin{aligned}m_1 &= \frac{1}{1}\sum_{j=0}^{0} f(X_{\lambda_j+1})=\frac{1}{1}=1\\ m_2 &= \frac{1}{2}\sum_{j=0}^{1} f(X_{\lambda_j+1})=\frac{1+0}{2}=\frac{1}{2}\\ m_3 &= \frac{1}{3}\sum_{j=0}^{2} f(X_{\lambda_j+1})=\frac{1+0+1}{3}=\frac{2}{3}.\end{aligned}$$

Define the auxiliary time series $\{\tilde{X}_n\}_{n=-\infty}^0$ as follows. Set $\tilde{X}_0=X_0$ and for $j=1,2,\dots$ and $\lambda_{j-1}+1\le n\le \lambda_j$ define

$$\tilde{X}_{-n}=X_{\lambda_j-n}. \tag{2.1}$$

Since $X_{\lambda_{j+1}-\lambda_j}^{\lambda_{j+1}}=X_0^{\lambda_j}$, the process $\{\tilde{X}_n\}_{n=-\infty}^0$ is well defined and it is immediate that

$$\tilde{X}_{-n}=X_{\lambda_j-n} \text{ for any } \ j\ge n\ge 0.$$

By Lemma 4.3 in the Appendix, the time series $\{\tilde{X}_n\}_{n=-\infty}^0$ has the same distribution as $\{X_n\}_{n=-\infty}^0$ and $\{\tilde{X}_n\}_{n=-\infty}^0$ is stationary, since $\{X_n\}_{n=-\infty}^0$ is stationary. Thus the one sided time series $\{\tilde{X}_n\}_{n=-\infty}^0$ can be extended to be a two-sided time series $\{\tilde{X}_n\}_{n=-\infty}^\infty$. We use this fact for the purpose of defining the conditional expectation $E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0)$.

Example 2.3. Let $\mathcal{X} = \{a, b\}$. Consider

$$X_0^6 = (X_0, X_1, \ldots, X_5, X_6) = aabaaaa.$$

The λ's are:

$$\begin{aligned} \lambda_0 &= 0 \\ \lambda_1 &= 1 \\ \lambda_2 &= 4. \end{aligned}$$

Then

$$\begin{aligned} \tilde{X}_0 &= a \\ \tilde{X}_{-1} &= a \\ \tilde{X}_{-2} &= b \\ \tilde{X}_{-3} &= a \\ \tilde{X}_{-4} &= a \end{aligned}$$

that is

$$\tilde{X}_{-4}^0 = \tilde{X}_{-4}, \ldots, \tilde{X}_{-1}, \tilde{X}_0 = aabaa.$$

The first theorem is on consistency.

Theorem 2.4. *Let $\mathcal{X}$ be a finite or countable alphabet, let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary process taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a real valued function with $E|f(X_1)| < \infty$. Then*

$$\lim_{n\to\infty} m_n = \lim_{n\to\infty} E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) = E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0) \tag{2.2}$$

almost surely,

$$\lim_{n\to\infty} |m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})| = 0 \tag{2.3}$$

almost surely,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} E(f(X_{\lambda_j+1}) I_{\{|f(X_{\lambda_j+1})| \le j\}} | X_{-\infty}^{\lambda_j}) = E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0) \tag{2.4}$$

almost surely and

$$\lim_{n\to\infty} \left| m_n - \left(\frac{1}{n+1} \sum_{j=0}^{n} E(f(X_{\lambda_j+1}) I_{\{|f(X_{\lambda_j+1})| \le j\}} | X_{-\infty}^{\lambda_j}) \right) \right| = 0 \quad \textit{almost surely.} \tag{2.5}$$

Moreover if $E(|f(X_1)| \log^+(|f(X_1)|)) < \infty$ then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) = E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0) \tag{2.6}$$

almost surely,

$$\lim_{n\to\infty} \left| m_n - \left(\frac{1}{n+1} \sum_{j=0}^{n} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) \right) \right| = 0 \tag{2.7}$$

almost surely and

$$\lim_{n\to\infty} \left| E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) - \left(\frac{1}{n+1} \sum_{j=0}^{n} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) \right) \right| = 0 \tag{2.8}$$

almost surely.

Remark 2.5. Note that one can not estimate for all time instances even if it is known that the process is first order Markov with state space in a countable and bounded subset of the real numbers. More precisely, consider the countable and bounded alphabet $\mathcal{X} = \{0, 1, 2^{-s}, 1 + 2^{-s} \quad \text{for} \quad s = 1, 2, \dots\}$ in Györfi et al. (1998). Then for any sequence of estimators $\hat{e}_n : \mathcal{X}^n \to \Re$ there is a stationary and ergodic first order Markov chain $\{X_n\}_{n=0}^{\infty}$ taking values in $\mathcal{X}$ such that

$$P(\limsup_{n\to\infty}\{|\hat{e}_n(X_0^{n-1}) - E(X_n|X_{n-1})| \geq 1/8\}) \geq \frac{1}{8}.$$

Note that $|X_0| \leq 2$. Intuitively, when a new letter x appears at the first time at time n, one may have no clue about the value of the conditional expectation $E(X_n|X_{n-1} = x)$ from the observations X_0^{n-1} (in which the letter x can not be found) and since the alphabet is infinite, this situation happens infinitely often, cf. B. Ryabko (Ryabko, 1988) and Györfi, Morvai, and Yakowitz (Györfi et al., 1998). In the intermittent case we may skip these time instances. Of course, for restricted class of processes one can estimate for all time instances. For example, consider the class of independent identically distributed zero/one valued random variables. Then the trivial estimator $\hat{e}_n = \frac{1}{n}\sum_{j=0}^{n-1} X_j$ works, that is, $|\hat{e}_n - E(X_n|X_0^{n-1})| \to 0$ almost surely. But it obviously does not work, not even on our stopping time sequence, for the periodic binary Markov chain which alternates between zero and one and yields a stationary and ergodic process. Indeed, if $X_{\lambda_n} = 0$ then $E(X_{\lambda_n+1}|X_0^{\lambda_n}) = 1$ and if $X_{\lambda_n} = 1$ then $E(X_{\lambda_n+1}|X_0^{\lambda_n}) = 0$. But $\hat{e}_n \to 0.5$ almost surely.

Remark 2.6. The key to the proof of Theorem 2.4 is that $E(f(X_{\lambda_n+1})|X_0^{\lambda_n})$ converges almost surely as $n \to \infty$ as will be proven in (3.4) in the proof of Theorem 2.4.

Remark 2.7. The main arguments and ideas for proving the pointwise consistency of the forward going intermittent estimator in (2.3) can already be found in the proof of the pointwise consistency of the backward going estimator in Algoet (1999). (Cf. Morvai et al., 1996 also.) (2.2) and (2.3) were already proved for binary random variables in Morvai (2003) and for unbounded real valued random variables with finite second moment in Morvai and Weiss (2005b).

Remark 2.8. Note that these stopping times λ_n and estimators m_n are the special forward going versions of the backward going stopping times and estimators in Morvai et al. (1996) and Algoet (1999). (Cf. Morvai, 2003 and Morvai and Weiss, 2005b also.) Note also that while the backward going estimators in Morvai et al. (1996) and Algoet (1999) work for general real valued stationary processes (uncountable alphabet), the forward going versions of those estimators do not automatically work in that generality. Morvai et al. (1996) and Algoet (1999) used nested sequence of partitions $\{\mathcal{P}_k\}_{k=0}^{\infty}$ of the real line in their algorithms. E.g. the following partition was admissible for the general backward going estimator in Algoet (1999)

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

Let $x \to [x]^k$ denote a quantizer that assigns to any real number x the unique interval in $\mathcal{P}_k$ that contains x. Let $[X_m^n]^k = ([X_m]^k, \dots, [X_n]^k)$. In the spirit of Algoet (1999) the forward going stopping times would be as follows. $\lambda_0 = 0$ and for $n = 1, 2, \dots,$

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

Notice that now matching is required only up to the precision of the quantizer (non exact matchings). Now the n-th estimate m_n would be $m_n = \frac{1}{n}\sum_{j=0}^{n-1} X_{\lambda_j+1}$. However, Morvai and Weiss (2005b) proved that in this case there is a stationary and ergodic first order Markov chain $\{X_n\}_{n=0}^{\infty}$ taking values from a countable subset of the unit interval such that

$$P\left(\limsup_{n\to\infty}\left|m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n})\right| > 0\right) > 0.$$

Thus there is some difference between the general backward case (in Morvai et al., 1996 and Algoet (1999)) and the forward intermittent case. That is, it is very important that no quantizers are used in our scheme in the forward intermittent case. We insist on exact matchings in defining our stopping times. For more details cf. Morvai and Weiss (2005b).

Remark 2.9. Note that this stopping time sequence $\{\lambda_n\}_{n=0}^{\infty}$ is the only so far known sequence of stopping times for which there is a known pointwise consistent estimator in such a generality as in Theorem 2.4. It is an open problem if there are other stopping time sequences with estimators which are pointwise consistent under such mild assumptions as in Theorem 2.4.

Remark 2.10. The growth of our stopping time sequence $\{\lambda_n\}_{n=0}^{\infty}$ is faster than exponential, even in the case of binary alphabet, if the entropy of the process is positive, cf. Morvai (2003). Thus we can not hope for rates of convergence for this forward going intermittent estimator. The main value of the construction of these almost sure finite stopping times is not their practicality but their mere existence, that is, it is possible to estimate in a pointwise sense along these almost surely finite stopping times in a forward going intermittent way. Now one may try to find better sequence of stopping times.

Remark 2.11. If we make restrictions on the class of processes then one can construct better sequence of stopping times. In Morvai and Weiss (2020b) binary renewal processes were considered and even rates of convergence were proved. In Morvai and Weiss (2004), Morvai and Weiss (2003), Morvai and Weiss (2005b) and Morvai and Weiss (2012) continuity of the conditional expectation $E(f(X_1)|X_{-\infty}^0)$ on a set with probability one was assumed. Particularly, Morvai and Weiss (2012) considered stopping times as follows. For $k \geq 1$, let $1 \leq l_k \leq k$ be a nondecreasing unbounded sequence of integers, that is, $1 = l_1 \leq l_2 \dots$ and $\lim_{k\to\infty} l_k = \infty$. Set $\zeta_0 = 0$ and for $n = 1, 2, \dots$, let

$$\zeta_n = \min\{t > 0 : X_{t+\zeta_{n-1}-(l_n-1)}^{t+\zeta_{n-1}} = X_{\zeta_{n-1}-(l_n-1)}^{\zeta_{n-1}}\} + \zeta_{n-1}.$$

Morvai and Weiss (2012) showed that one can choose l_n's in such a way that the growth of the stopping times ζ_n will not be faster than polynomial in n.

Remark 2.12. Note that in (2.7) and (2.8), for $i, j \geq 0$ if $j \neq i$ then it is not true generally that

$$E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) = E(f(X_{\lambda_i+1})|X_{-\infty}^{\lambda_i})$$

almost surely even though the process $\{X_n\}_{n=-\infty}^{\infty}$ is stationary since the conditional expectations are evaluated at different values.

Remark 2.13. Note that (2.7) holds even though the estimator m_n depends solely on $(X_0, \dots, X_{\lambda_{n-1}})$ and it will never observe the values $(\dots, X_{-2}, X_{-1})$.

The second theorem is on integrability of various suprema. These results will give information about the magnitude of the estimators and the error.

Theorem 2.14. *Let $\mathcal{X}$ be a finite or countable alphabet, let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary process taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a real valued function with $E(|f(X_1)| \log^+(|f(X_1)|)) < \infty$. Then*

$$E(\sup_{n\ge 1} |m_n|) < \infty, \tag{2.9}$$

$$E(\sup_{n\ge 1} |E(f(X_{\lambda_n+1})|X_0^{\lambda_n})|) < \infty, \tag{2.10}$$

$$E(\sup_{n\ge 1} |m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})|) < \infty, \tag{2.11}$$

$$E(\sup_{n\ge 1} |\frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})|) < \infty \tag{2.12}$$

and

$$E\left(\sup_{n\ge 1}\left| m_n - \left(\frac{1}{n+1}\sum_{j=0}^{n} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})\right)\right|\right) < \infty. \tag{2.13}$$

Remark 2.15. For the backward estimator pointwise convergence was proved in Algoet (1999) and integrability of the supremum was proved in Morvai and Weiss (2011, 2020a). These are the conditions of Breiman's generalized ergodic theorem which is the main tool for proving convergence in Cesaro mean. For more details cf. Morvai and Weiss (2011, 2020a).

The third and main theorem is on limit distributions.

Theorem 2.16. *Let $\mathcal{X}$ be a finite or countably infinite alphabet. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary process taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a real valued function. Assume that for some $\alpha > 0$ $E\left(|f(X_1)|^{2+2\alpha}\right) < \infty$. Let $\epsilon > 0$ be arbitrary. Assume that there exist an almost surely finite real valued function $C(X_{-\infty}^0) < \infty$ and an integer valued function $0 < N(X_{-\infty}^0) < \infty$ such that for all $n \ge N(X_{-\infty}^0)$ and $m > n$, almost surely,*

$$|E(f(X_1)|X_{-n}^0) - E(f(X_1)|X_{-m}^0)| \le \frac{C(X_{-\infty}^0)}{n^{0.5+\epsilon}}. \tag{2.14}$$

Let

$$\eta = \sqrt{(E(f^2(X_1)|X_{-\infty}^0) - E(f(X_1)|X_{-\infty}^0)^2)}.$$

Then there is a random variable Z such that for all continuity points t of $P(Z < t)$,

$$\lim_{n\to\infty} P(\sqrt{n}(m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n})) < t) = P(Z < t) \tag{2.15}$$

where the random variable Z has characteristic function $E(e^{-\frac{1}{2}\eta^2 t^2})$.
If in addition $\eta > 0$ almost surely then for all t, and for all measurable set $H \in \sigma(X_{-\infty}^{\infty})$

$$P\left(\left\{\frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} < t\right\}\bigcap H\right) \to \Phi(t)P(X_{-\infty}^{\infty} \in H). \tag{2.16}$$

Particularly, (for $H = \Omega$), for all t,

$$P\left(\frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} < t\right) \to \Phi(t). \tag{2.17}$$

Remark 2.17. The key variable η that enters the results in Theorem 2.16 is the conditional variance of X_0 given the past and the condition that it be greater than zero is a natural one in the presence of positive entropy.

Note that the conditions of Theorem 2.16 do not imply that the process is mixing as the following example shows.

Example 2.18. Consider the Markov chain with states $\{0,1\}$ which alternates between the two states. This periodic Markov chain yields a stationary and ergodic process $\{M_n\}_{n=-\infty}^{\infty}$ with marginal distribution uniformly distributed on the two states. Let $\{U_n\}_{n=-\infty}^{\infty}$ be a sequence of independent and identically distributed random variables with $P(U_0=-100)=P(U_0=100)=0.5$ and let the $\{U_n\}_{n=-\infty}^{\infty}$ process be independent of the $\{M_n\}_{n=-\infty}^{\infty}$ process. Now let $X_n=M_n+U_n$. Clearly $\{X_n\}_{n=-\infty}^{\infty}$ is a stationary and ergodic process which is a first order Markov chain and satisfies (with the choice of f being the identity function) all the conditions of Theorem 2.16 (even η is positive almost surely) but it is not mixing.

Note that the conditions of Theorem 2.16 do not imply that the process is ergodic as the following example shows.

Example 2.19. Consider the Markov chain with four states $\{-9,-1,2,8\}$ and transitions $P(M_1=-1|M_0=-9)=P(M_1=-9|M_0=-1)=P(M_1=2|M_0=8)=P(M_1=8|M_0=2)=1$. Choosing the uniform distribution on the four states, we get a stationary but nonergodic process $\{M_n\}_{n=-\infty}^{\infty}$. Let $\{U_n\}_{n=-\infty}^{\infty}$ be a sequence of independent and identically distributed random variables with $P(U_0=-100)=P(U_0=100)=0.5$ and let the $\{U_n\}_{n=-\infty}^{\infty}$ process be independent of the $\{M_n\}_{n=-\infty}^{\infty}$ process. Now let $X_n=M_n+U_n$. Clearly $\{X_n\}_{n=-\infty}^{\infty}$ is a stationary process which is a first order Markov chain and satisfies (with the choice of f being the identity function) all the conditions of Theorem 2.16 (even η is positive almost surely) but it is not ergodic.

Remark 2.20. Note that already $|E(f(X_1))|<\infty$ implies that

$$E(f(X_1)|X_{-n}^0)\to E(f(X_1)|X_{-\infty}^0) \ \ \textit{almost surely}$$

and so the condition in (2.14) is on the speed of convergence.

Remark 2.21. Note that the condition in (2.14) will be used in the proof of Theorem 2.16 to show that

$$\sum_{j=0}^{n-1}\frac{(E(f(X_{\lambda_j+1})|X_0^{\lambda_j})-E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))}{\sqrt{n}}\to 0$$

almost surely.

Now we will construct a stationary nonergodic binary process $\{X_n\}_{n=-\infty}^{\infty}$ which satisfies all the conditions in Theorem 2.16 but

$$E(X_1|X_{-\infty}^0)\neq E(X_1|X_{-n}^0)$$

almost surely for $0\le n<\infty$, even though

$$E(X_1|X_{-n}^0)\to E(X_1|X_{-\infty}^0)$$

almost surely.

Example 2.22. Consider two processes A and B. Process A is independent and identically distributed $\{0,1\}$ valued and in process A the probability of one is ß0.9. Process B is independent and identically distributed $\{0,1\}$ valued and in the case of process B the probability of one is 0.1. We choose process A with probability 0.5 and process B with probability 0.5. So we will observe process A and process B with equal probabilities. The resulting process $\{X_n\}_{n=-\infty}^{\infty}$ is binary stationary but it is neither Markov of any order nor ergodic. It has two ergodic modes A and B. Now the infinite past determines the ergodic mode. Thus $E(X_1|X_{-\infty}^0)=0.9$ if the infinite past $X_{-\infty}^0$ determines process A and $E(X_1|X_{-\infty}^0)=0.1$ if the infinite past $X_{-\infty}^0$ determines process B. Let r denote the ratio

$$r=\frac{P(X_{-n}^0=x_{-n}^0|A)}{P(X_{-n}^0=x_{-n}^0|B)}.$$

Now for almost every $X_{-\infty}^0$ there will be such an $N(X_{-\infty}^0)$ so that either for all $n\ge N$ the average of X_{-n}^0 is greater than 0.8 or for all $n\ge N$ the average of X_{-n}^0 is less than 0.2, by the strong law of large numbers. Suppose the former case. An easy calculation yields that for all $n\ge N(x_{-\infty}^0)$

$$r=\frac{P(X_{-n}^0=x_{-n}^0|A)}{P(X_{-n}^0=x_{-n}^0|B)}\ge 9^{0.6n}.$$

Now basic calculation yields that for all $n\ge N(X_{-\infty}^0)$

$$0<P(X_1=1|A)-P(X_1=1|X_{-n}^0)=0.9-P(X_1=1|X_{-n}^0)\le\frac{1}{r}\le 9^{-0.6n}.$$

Thus the difference tends to zero exponentially fast. The other case goes similarly. It is clear that $\eta>0$ almost surely. Thus the process $\{X_n\}_{n=-\infty}^{\infty}$ satisfies all the conditions in Theorem 2.16.

Now we will apply Theorem 2.16 in a special case.

Just to remind, $X_m^n=(X_m,\dots,X_n)$, where $m\le n$ and if $m>n$ then, by convention, let X_m^n be the empty word which will be denoted by $\emptyset$. Let

$$\mathcal{X}^*=\bigcup_{k=0}^{\infty}\mathcal{X}^k$$

where $\mathcal{X}^0$ is a set that contains exactly the empty word $\emptyset$.

For convenience let $p(x_{-k+1}^0)$ and $p(y|x_{-k+1}^0)$ denote the distribution $P(X_{-k+1}^0=x_{-k+1}^0)$ and the conditional distribution $P(X_1=y|X_{-k+1}^0=x_{-k+1}^0)$, respectively. Note that $P(X_{t+1}^t=\emptyset)=1$, $P(X_1=y|X_1^0=\emptyset)=P(X_1=y)$.

Definition 2.23. For some $0\le k$ and $w_{-k+1}^0\in\mathcal{X}^k$ we say that w_{-k+1}^0 is a memory word if $p(w_{-k+1}^0)>0$ and for all $i\ge 1$, all $y\in\mathcal{X}$, all $z_{-k-i+1}^{-k}\in\mathcal{X}^i$

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

provided $p(z_{-k-i+1}^{-k},w_{-k+1}^0,y)>0$. If no proper suffix of w is a memory word then w is called a minimal memory word.

Define the set $\mathcal{W}_k$ of those memory words w_{-k+1}^0 with length k, that is,

$$\mathcal{W}_k=\{w_{-k+1}^0\in\mathcal{X}^k:\ w_{-k+1}^0\text{ is a memory word}\}.$$

Let

$$\mathcal{W}^*=\bigcup_{k=0}^{\infty}\mathcal{W}_k.$$

Example 2.24. Consider an independent and identically distributed process $\{X_n\}_{n=-\infty}^{\infty}$ on a countable alphabet. Then the empty word is a memory word and it is the only minimal memory word.

Example 2.25. Consider the Markov chain 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 $\{M_n\}_{n=-\infty}^{\infty}$. Define

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

Then $\{Z_n\}_{n=-\infty}^{\infty}$ is a stationary and ergodic binary Markov chain with order 2. The minimal memory words of the process $\{Z_n\}_{n=-\infty}^{\infty}$ are the '1', the '10' and the '00'.

Example 2.26. Consider the Markov chain 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 $\{M_n\}_{n=-\infty}^{\infty}$. Define $Z_n = I_{\{M_n \neq 0\}}$. Then $\{Z_n\}_{n=-\infty}^{\infty}$ is a stationary and ergodic binary renewal process with renewal state '0'. The minimal memory words of the process $\{Z_n\}_{n=-\infty}^{\infty}$ are '0', '01', '011', '0111', '01111', ...

Example 2.27. Consider a stationary and ergodic binary renewal process with renewal state '0'. Then any word with positive probability which contains at least one '0' is a memory word, though not necessarily minimal. Any word which contains more than one '0" can not be a minimal memory word.

Definition 2.28. For a stationary time series $\{X_n\}_{n=-\infty}^{\infty}$ the (random) length $K(X_{-\infty}^0)$ of the memory of the sample path $X_{-\infty}^0$ is the smallest possible $0 \le K < \infty$ such that for all $i \ge 1$, 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.

Remark 2.29. For stationary time series $\{X_n\}_{n=-\infty}^{\infty}$, $K(x_{-\infty}^0)$ is the smallest $k \ge 0$ such that $x_{-k+1}^0 \in \mathcal{W}_k$ and $K(x_{-\infty}^0) = \infty$ if there is no such k.

Example 2.30. Consider an independent and identically distributed process $\{X_n\}_{n=-\infty}^{\infty}$ on a countable alphabet.Then

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

almost surely.

Example 2.31. Consider a stationary and ergodic first order finite or countably infinite Markov chain $\{X_n\}_{n=-\infty}^{\infty}$. Then

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

almost surely.

Example 2.32. Consider the stationary and ergodic binary second order Markov chain $\{Z_n\}_{n=-\infty}^{\infty}$ in Example 2.25. Then

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

almost surely.

Example 2.33. Consider a stationary and ergodic second order finite or countably infinite Markov chain $\{X_n\}_{n=-\infty}^{\infty}$. Then

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

almost surely.

Example 2.34. Consider a stationary and ergodic binary renewal process $\{X_n\}_{n=-\infty}^{\infty}$ with renewal state '0'. Let $\tau(X_{-\infty}^0)$ be the smallest $t \ge 0$ such that $X_{-t} = 0$ and $X_i = 1$ for all $-t < i \le 0$. Then

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

almost surely. Consider the stationary and ergodic binary renewal process $\{Z_n\}_{n=-\infty}^{\infty}$ in Example 2.26. Then

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

almost surely.

Definition 2.35. A stationary time series $\{X_n\}_{n=-\infty}^{\infty}$ is called finitarily Markovian if the set of memory words has probability one, that is,

$$P\left(\bigcup_{n=0}^{\infty} \{X_{-n+1}^0 \in \mathcal{W}_n\}\right) = 1.$$

Remark 2.36. The stationary time series $\{X_n\}_{n=-\infty}^{\infty}$ is finitarily Markovian if and only if $K(X_{-\infty}^0)$ is finite (though not necessarily bounded) almost surely.

This class includes of course all finite order Markov chains but also many other processes such as the finitarily determined processes of Kalikow, Katznelson and Weiss (Kalikow et al., 1992), which serve to represent all isomorphism classes of zero entropy processes. For some concrete examples that are not Markovian consider the following example:

Example 2.37. Let $\{M_n\}_{n=-\infty}^{\infty}$ be any stationary 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 binary time series $\{X_n\}_{n=-\infty}^{\infty}$ is stationary also. 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 $\{X_n\}_{n=-\infty}^{\infty}$ 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 yields a stationary Markov chain $\{M_n\}_{n=-\infty}^{\infty}$. Clearly, 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 even or odd number of zeros.

We note that Morvai and Weiss (2005d) proved that there is no classification rule for discriminating the class of finitarily Markovian processes from other stationary processes.

The fourth theorem is on limit distributions for finitarily Markovian processes.

Theorem 2.38. *Let $\mathcal{X}$ be a finite or countably infinite alphabet. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary finitarily Markovian (FM) process taking values from $\mathcal{X}$. Assume that for some $\alpha > 0$ $E|f(X_1)|^{2+2\alpha} < \infty$. Let $\epsilon > 0$ arbitrary. Let V be the set of minimal memory words. If $w_{-k+1}^0 \in V$ then let*

$$\eta(w) = \sqrt{(E(f^2(X_1)|X_{-k+1}^0 = w_{-k+1}^0) - E(f(X_1)|X_{-k+1}^0 = w_{-k+1}^0)^2)}.$$

For all t,

$$\lim_{n\to\infty} P(\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})) < t) = \sum_{w\in V} F_w(t)p(w) \tag{2.18}$$

where $F_w(t)$ is a distribution function of a normal distribution with zero mean and variance $\eta(w)$. That is, the limiting distribution is a mixture of normal distributions. If in addition, for all $w \in V$ $\eta(w) > 0$ then for all $w \in V$ and t,

$$P\left(\frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X_{\lambda_n-k+1}^{\lambda_n} = w_{-k+1}^0))}{\sqrt{(E(f^2(X_{\lambda_n+1})|X_{\lambda_n-k+1}^{\lambda_n} = w_{-k+1}^0) - E(f(X_{\lambda_n+1})|X_{\lambda_n-k+1}^{\lambda_n} = w_{-k+1}^0)^2)}} < t \middle| H_w \right) \to \Phi(t) \tag{2.19}$$

where $H_w = \{X_{\lambda_k-k+1}^{\lambda_k} = w_{-k+1}^0\}$.

Remark 2.39. For a stationary, nonergodic, non finitarily Markovian, binary process satisfying all the conditions in Theorem 2.16 cf. Example 2.22.

For further reading on memory words see Morvai and Weiss (2007a), Morvai and Weiss (2008) and Morvai and Weiss (2005e).

3. Proofs

For an arbitrary $\mathcal{X}$-valued stationary time series $\{Y_n\}_{n=-\infty}^{0}$, let $\hat\lambda_0(Y_{-\infty}^0) = 0$ and for $n \ge 1$ define

$$\hat\lambda_n(Y_{-\infty}^0) = \hat\lambda_{n-1}(Y_{-\infty}^0) - \min\{t > 0 : Y_{\hat\lambda_{n-1}-t}^{-t} = Y_{\hat\lambda_{n-1}}^0\}. \tag{3.1}$$

Let T denote the left shift operator, that is, $(Tx_{-\infty}^{\infty})_i = x_{i+1}$. It is easy to see that if $\lambda_n(x_{-\infty}^{\infty}) = l$ then $\hat\lambda_n(T^l x_{-\infty}^{\infty}) = -l$.

It is immediate that

$$\tilde X_{\hat\lambda_n}^0 = X_0^{\lambda_n} \quad \text{for all } n \ge 0$$

since $X_{\lambda_{n+1}-\lambda_n}^{\lambda_{n+1}} = X_0^{\lambda_n}$.

3.1. **Proof of Theorem 2.4.** We will use the decomposition and reasoning as in the case of the backward going estimator in Algoet (1999) (in the proof of Theorem 4 in Algoet (1999)). (Cf. Morvai et al., 1996 also.) We want to truncate the potentially unbounded function f (as in Algoet, 1999) so that classic convergence theorems can be applied. This involves writing m_n as a sum of three terms (cf. the proof of Theorem 4 in Algoet, 1999). The first term handles the error made by the truncation. The second term is also negligible by results on the convergence of martingales. Finally the third term is shown to converge to the desired limit. The technical details are relegated to the Appendix. Consider

$$\begin{aligned} m_n &= \frac{1}{n}\sum_{j=0}^{n-1}\left(f(X_{\lambda_j+1}) - f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}\right) \\ &+ \frac{1}{n}\sum_{j=0}^{n-1}\left(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}} - E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}|X_0^{\lambda_j})\right) \\ &+ \frac{1}{n}\sum_{j=0}^{n-1}E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}|X_0^{\lambda_j}) \\ &= A_n + B_n + C_n. \end{aligned}$$

By Lemma 4.1 in the Appendix, X_{λ_j+1} has the same distribution as X_1 and the finite expectation of $f(X_1)$ implies that

$$\sum_{j=0}^{\infty}P(|f(X_{\lambda_j+1})| > j) = \sum_{j=0}^{\infty}P(|f(X_1)| > j) < \infty$$

and by the Borel-Cantelli lemma

$$I_{\{|f(X_{\lambda_j+1})|\le j\}} = 1$$

eventually almost surely. Thus $A_n \to 0$ almost surely.
Applying Lemma 4.5 (with $W_{j+1} = X_{\lambda_j+1}$ and $G_j = \sigma(X_0^{\lambda_j})$) we get $B_n \to 0$ almost surely.
Now we deal with the last term. By Lemma 4.1, Lemma 4.2 and Lemma 4.3 in the Appendix

$$E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}|X_0^{\lambda_j}) = E(f(\tilde{X}_1)I_{\{|f(\tilde{X}_1)|\le j\}}|\tilde{X}^0_{\hat{\lambda}_j(\tilde{X}^0_{-\infty})}).$$

Since

$$\sigma(\tilde{X}^0_{\hat{\lambda}_j(\tilde{X}^0_{-\infty})}) \uparrow \sigma(\tilde{X}^0_{-\infty}),$$

$$f(\tilde{X}_1)I_{\{|f(\tilde{X}_1)|\le j\}} \to f(\tilde{X}_1) \text{ almost surely,}$$

$$\sup_{j\ge 1}|f(\tilde{X}_1)I_{\{|f(\tilde{X}_1)|\le j\}}| \le |f(\tilde{X}_1)|$$

and

$$E(|f(\tilde{X}_1)|) < \infty,$$

by Corollary 1 pp. 237-238 in Chow and Teicher (1978) (Lemma 3 in Algoet Algoet, 1999) we get that almost surely

$$E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}|X_0^{\lambda_j}) = E(f(\tilde{X}_1)I_{\{|f(\tilde{X}_1)|\le j\}}|\tilde{X}^0_{\hat{\lambda}_j(\tilde{X}^0_{-\infty})}) \to E(f(\tilde{X}_1)|\tilde{X}^0_{-\infty}). \tag{3.2}$$

In turn,

$$\lim_{n\to\infty} m_n = E(f(\tilde{X}_1)|\tilde{X}^0_{-\infty}) \tag{3.3}$$

almost surely. Similarly,

$$E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) = E(f(\tilde{X}_1)|\tilde{X}^0_{\hat{\lambda}_j(\tilde{X}^0_{-\infty})}) \to E(f(\tilde{X}_1)|\tilde{X}^0_{-\infty}) \tag{3.4}$$

almost surely. We have proved (2.2) and (2.3).
Now we prove (2.4). Let

$$D_n = \frac{1}{n}\sum_{j=0}^{n-1}\left(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}} - E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\le j\}}|X_{-\infty}^{\lambda_j})\right).$$

Applying Lemma 4.5 (with $W_{j+1} = X_{\lambda_j+1}$ and $G_j = \sigma(X_{-\infty}^{\lambda_j})$) we get $D_n \to 0$ almost surely. Since

$$\frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})I_{\{|f(X_{\lambda_j+1})|\leq j\}}|X_{-\infty}^{\lambda_j}) = m_n - A_n - D_n$$

and we have already seen that $A_n \to 0$, $D_n \to 0$ and $m_n \to E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0)$ almost surely, we get (2.4). By (3.3) and (2.4) we get (2.5).

Now we assume $E(|f(X_1)|\log^+(|f(X_1)|)) < \infty$. By Lemma 4.1 in the Appendix, X_{λ_j+1} has the same distribution as X_1 and by applying Lemma 4.6 (with $W_{j+1} = X_{\lambda_j+1}$ and $G_j = \sigma(X_{-\infty}^{\lambda_j})$) we get

$$\left(m_n - \frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})\right) = \frac{1}{n}\sum_{j=0}^{n-1}\Big(f(X_{\lambda_j+1}) - E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})\Big) \to 0$$

almost surely. (This also follows from Elton, 1981, since the martingale differences

$$Z_j = f(X_{\lambda_j+1}) - E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})$$

are identically distributed by Lemma 4.1 in the Appendix.) Now by (3.3) m_n converges to $E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0)$ almost surely, and in turn

$$\frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) \to E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0)$$

almost surely also. We have proved (2.6) and (2.7). By (3.4) we get (2.8). The proof of Theorem 2.4 is complete.

3.2. **Proof of Theorem 2.14.**

For this theorem we have a stronger assumption on the function f beyond mere integrability. The stronger assumption on f enables us to use martingale techniques e.g. the classic result of Doob. Here too the technical details are in the Appendix. For the proof m_n will be decomposed into two terms and dealt with separately. Consider

$$\begin{aligned} m_n &= \frac{1}{n}\sum_{j=0}^{n-1}\Big(f(X_{\lambda_j+1}) - E(f(X_{\lambda_j+1})|X_0^{\lambda_j})\Big) \\ &+ \frac{1}{n}\sum_{j=0}^{n-1} E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) \\ &= A_n + B_n. \end{aligned}$$

By Lemma 4.1 in the Appendix, X_{λ_j+1} has the same distribution as X_1 and by applying Lemma 4.6 (with $W_j = X_{\lambda_{j-1}+1}$ and $G_{j-1} = \sigma(X_0^{\lambda_{j-1}})$) we get that $E(\sup_{n\geq 1}|A_n|) < \infty$. (Note that this does not follow from Elton (1981), since the martingale differences

$$U_j = f(X_{\lambda_j+1}) - E(f(X_{\lambda_j+1})|X_0^{\lambda_j})$$

are not identically distributed in general.)

By Lemma 4.1, Lemma 4.2 and Lemma 4.3 in the Appendix

$$E(f(X_{\lambda_j+1})|X_0^{\lambda_j}) = E(f(\tilde{X}_1)|\tilde{X}^0_{\hat{\lambda}_j(\tilde{X}_{-\infty}^0)})$$

almost surely and by Doob's inequality we get (2.10). Now by (2.10), $E(\sup_{n\ge 1}|B_n|)<\infty$ and we get (2.9). By (2.10) and (2.9) we get (2.11). Now

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{n-1}E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j}) &= -\frac{1}{n}\sum_{j=0}^{n-1}\Big(f(X_{\lambda_j+1})-E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})\Big)\\ &+ m_n.\end{aligned}$$

By Lemma 4.1 in the Appendix, X_{λ_j+1} has the same distribution as X_1 and by applying Lemma 4.6 (with $W_j=X_{\lambda_{j-1}+1}$ and $G_{j-1}=\sigma(X_{-\infty}^{\lambda_{j-1}})$) we get that

$$E(\sup_{n\ge 1}|\frac{1}{n}\sum_{j=0}^{n-1}\Big(f(X_{\lambda_j+1})-E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})\Big)|)<\infty.$$

(This also follows from Elton Elton (1981), since the martingale differences

$$Z_j=f(X_{\lambda_j+1})-E(f(X_{\lambda_j+1})|X_{-\infty}^{\lambda_j})$$

are identically distributed by Lemma 4.1 in the Appendix.) Now by (2.9) we get (2.12). By (2.12) and (2.9) we get (2.13). The proof of Theorem 2.14 is complete.

3.3. **Proof of Theorem 2.16.** Our proof of this theorem is based on a quite general central limit theorem for triangular arrays of martingales that can be found in the book of Hall and Heyde (1980). The bulk of what we do here is to show how to verify the conditions of that theorem. As to be expected the details are quite technical. First we prove (2.15).

$$\begin{aligned}&\sqrt{n}(m_n-E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))\\ &= \sqrt{n}(\frac{1}{n}\sum_{j=0}^{n-1}f(X_{\lambda_j+1})-E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))\\ &= \sum_{j=0}^{n-1}\frac{(f(X_{\lambda_j+1})-E(f(X_{\lambda_j+1})|X_0^{\lambda_j}))}{\sqrt{n}}\\ &+ \sum_{j=0}^{n-1}\frac{(E(f(X_{\lambda_j+1})|X_0^{\lambda_j})-E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))}{\sqrt{n}}\\ &= A_n+B_n.\end{aligned}$$

First we will deal with B_n. Since by definition $\tilde{X}_{\hat{\lambda}_n}^0=X_0^{\lambda_n}$ for all $n\ge 0$ and by Lemma 4.3 in the Appendix the processes $\{\tilde{X}_n\}$ and $\{X_n\}$ have the same distribution we have that

$$\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}(E(f(X_{\lambda_j+1})|X_0^{\lambda_j})-E(f(X_{\lambda_n+1})|X_0^{\lambda_n}))=\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}(E(f(\tilde{X}_1)|\tilde{X}_{\hat{\lambda}_j}^0)-E(f(\tilde{X}_1)|\tilde{X}_{\hat{\lambda}_n}^0).$$

Now by assumption there exists $C(\tilde{X}_{-\infty}^0)$ and $0<N(\tilde{X}_{-\infty}^0)<\infty$ such that for all $n\ge N(\tilde{X}_{-\infty}^0)$ and $m>n$, almost surely,

$$|E(f(\tilde{X}_1)|\tilde{X}_{-n}^0)-E(f(\tilde{X}_1)|\tilde{X}_{-m}^0)|\le\frac{C(\tilde{X}_{-\infty}^0)}{n^{0.5+\epsilon}}.$$

Thus B_n tends to zero almost surely.

Now we deal with A_n. For $n \geq 1$, $0 \leq i \leq n-1$, $W_{n,i}$ with $Z_i = f(X_{\lambda_i+1}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})$, $W_{n,i} = \frac{Z_i}{\sqrt{n}}$ is a sequence of martingale arrays and we will verify the conditions in Corollary 3.1 in Hall and Heyde (1980). First we show that for all $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} E\left((\frac{Z_i}{\sqrt{n}})^2 I_{\{|\frac{Z_i}{\sqrt{n}}|>\epsilon\}} |X_0^{\lambda_i} \right) = 0 \tag{3.5}$$

in probability. It will be enough to show that

$$E\left(\sum_{i=0}^{n-1} E\left((\frac{Z_i}{\sqrt{n}})^2 I_{\{|\frac{Z_i}{\sqrt{n}}|>\epsilon\}} |X_0^{\lambda_i} \right) \right) \to 0.$$

An easy calculation yields that

$$\begin{aligned}
&\sum_{i=0}^{n-1} E\left(E\left((\frac{Z_i}{\sqrt{n}})^2 I_{\{|\frac{Z_i}{\sqrt{n}}|>\epsilon\}} |X_0^{\lambda_i} \right) \right) \\
&= n^{-1-\alpha} \sum_{i=0}^{n-1} E\left((Z_i)^2 n^\alpha I_{\{\frac{|Z_i|^{2\alpha}}{\epsilon^{2\alpha}} > n^\alpha\}} \right) \\
&\leq n^{-1-\alpha} \sum_{i=0}^{n-1} E\left(|Z_i|^2 \frac{|Z_i|^{2\alpha}}{\epsilon^{2\alpha}} \right) \\
&= n^{-\alpha} \left(\epsilon^{-2\alpha} \frac{1}{n} \sum_{i=0}^{n-1} E\left(|f(X_{\lambda_i+1}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})|^{2+2\alpha} \right) \right) \\
&\leq n^{-\alpha} \left(\epsilon^{-2\alpha} \frac{1}{n} \sum_{i=0}^{n-1} E\left(|2f(X_{\lambda_i+1})|^{2+2\alpha} + |2E(f(X_{\lambda_i+1})|X_0^{\lambda_i})|^{2+2\alpha} \right) \right).
\end{aligned}$$

By Jensen's inequality and since by Lemma 4.1 in the Appendix, X_{λ_j+1} has the same distribution as X_1 we get

$$\begin{aligned}
&n^{-\alpha} \left(\epsilon^{-2\alpha} \frac{1}{n} \sum_{i=0}^{n-1} E\left(|2f(X_{\lambda_i+1})|^{2+2\alpha} + |2E(f(X_{\lambda_i+1})|X_0^{\lambda_i})|^{2+2\alpha} \right) \right) \\
&\leq n^{-\alpha} \left(\epsilon^{-2\alpha} 2^{2+2\alpha} \frac{1}{n} \sum_{i=0}^{n-1} E\left(|f(X_{\lambda_i+1})|^{2+2\alpha} + E(|f(X_{\lambda_i+1})|^{2+2\alpha}|X_0^{\lambda_i}) \right) \right) \\
&= n^{-\alpha} \left(\epsilon^{-2\alpha} \frac{1}{n} n 2^{2+2\alpha+1} E\left(|f(X_1)|^{2+2\alpha} \right) \right) \\
&= n^{-\alpha} \left(\epsilon^{-2\alpha} 2^{2+2\alpha+1} E\left(|f(X_1)|^{2+2\alpha} \right) \right) \\
&\to 0.
\end{aligned}$$

We have proved the first condition in Corollary 3.1 in Hall and Heyde (1980). Now we deal with the second condition in Corollary 3.1 in Hall and Heyde (1980). It will be enought to prove that

$$\lim_{n\to\infty} \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)} = \tilde{\eta} \tag{3.6}$$

almost surely where

$$\tilde{\eta} = \sqrt{(E(f^2(\tilde{X}_1)|\tilde{X}_{-\infty}^0) - E(f(\tilde{X}_1)|\tilde{X}_{-\infty}^0)^2)}.$$

By Lemma 4.2, Lemma 4.3 and the martingale convergence theorem for conditional expectations, almost surely,

$$\lim_{n\to\infty} E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) = \lim_{n\to\infty} E(f(\tilde{X}_1)|\tilde{X}^0_{\hat{\lambda}_n(\tilde{X}^0_{-\infty})}) = E(f(\tilde{X}_1)|\tilde{X}^0_{-\infty})$$

and

$$\lim_{n\to\infty} E(f^2(X_{\lambda_n+1})|X_0^{\lambda_n}) = \lim_{n\to\infty} E(f^2(\tilde{X}_1)|\tilde{X}^0_{\hat{\lambda}_n(\tilde{X}^0_{-\infty})}) = E(f^2(\tilde{X}_1)|\tilde{X}^0_{-\infty}).$$

We have proved the second condition in Corollary 3.1 in Hall and Heyde (1980). Since for $n \geq 1$, $0 \leq i \leq n-1$, $W_{n,i}$ is a sequence of martingale arrays and all the conditions of Corollary 3.1 in Hall and Heyde (1980) are satisfied and so by Corollary 3.1 in Hall and Heyde (1980) we get

$$\lim_{n\to\infty} P(\sqrt{n}(m_n - E(X_{\lambda_n+1}|X_0^{\lambda_n})) < t) = P(Z < t)$$

where the random variable Z has characteristic function $E(e^{-\frac{1}{2}\tilde{\eta}^2 t^2})$. Since by Lemma 4.3 the distributions of $\{\tilde{X}_n\}_{n=-\infty}^{\infty}$ and $\{X_n\}_{n=-\infty}^{\infty}$ are the same we get (2.15).

Now we prove (2.16).

$$\begin{aligned}
&\sqrt{n}\frac{m_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} \\
= \quad & \sqrt{n}\frac{\frac{1}{n}\sum_{j=0}^{n-1} f(X_{\lambda_j+1}) - E(f(X_{\lambda_n+1})|X_0^{\lambda_n})}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} \\
= \quad & \frac{A_n}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} \\
+ \quad & \frac{B_n}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X_0^{\lambda_i}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})^2)}} \\
= \quad & C_n + D_n.
\end{aligned}$$

First we deal with D_n. We have already proved that B_n tends to zero almost surely. The denominator tends to an almost surely strictly positive random variable by (3.6), the assumption that $\eta > 0$ almost surely and since $\tilde{\eta}$ and η has the same distribution. Thus D_n tends to zero almost surely. Now we deal with C_n. For $n \geq 1$, $0 \leq i \leq n-1$ with $Z_i = f(X_{\lambda_i+1}) - E(f(X_{\lambda_i+1})|X_0^{\lambda_i})$, $W_{n,i} = \frac{Z_i}{\sqrt{n}}$ is a sequence of martingale arrays and since by assumption $\eta > 0$ almost surely, by (3.5) and (3.6) the conditions of Corollary 3.2 in Hall and Heyde (1980) are satisfied and so C_n tends in distribution to a standard normal distribution. The proof of Theorem 2.16 is complete.

3.4. *Proof of Theorem 2.38*. Since the distribution function F_w has characteristic function $e^{-\frac{1}{2}\eta(w)^2 t^2}$, the mixture of distribution functions has characteristic function as the mixture of the characteristic functions we get that the characteristic function of $\sum_{w\in V} F_w(t)p(w)$ is

$$\sum_{w\in V} p(w)e^{-\frac{1}{2}\eta(w)^2 t^2} = E(e^{-\frac{1}{2}\eta^2 t^2}).$$

Since by Lemma 4.7 in the Appendix condition (2.14) in Theorem 2.16 is satisfied, by (2.15) in Theorem 2.16, the limit distribution has characteristic function $E(e^{-\frac{1}{2}\eta^2 t^2})$ this must be the mixture $\sum_{w\in V} F_w p(w)$. (The characteristic function determines the distribution.) We have proven (2.18). Now we prove (2.19).

Now on the set H_w, for $n > k$,

$$\frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1}))}{\sqrt{(E(f^2(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1}) - E(f(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1})^2)}}$$
$$= \frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X^{\lambda_n}_0))}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X^{\lambda_i}_0) - E(f(X_{\lambda_i+1})|X^{\lambda_i}_0)^2)}}$$
$$\cdot \frac{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X^{\lambda_i}_0) - E(f(X_{\lambda_i+1})|X^{\lambda_i}_0)^2)}}{\sqrt{(E(f^2(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1}) - E(f(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1})^2)}}$$

and

$$\frac{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X^{\lambda_i}_0) - E(f(X_{\lambda_i+1})|X^{\lambda_i}_0)^2)}}{\sqrt{(E(f^2(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1}) - E(f(X_{\lambda_n+1})|X^{\lambda_n}_{\lambda_n-k+1} = w^0_{-k+1})^2)}} \to 1$$

almost surely on H_w. Now (2.19) follows from (2.16) in Theorem 2.16

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(m_n - E(f(X_{\lambda_n+1})|X^{\lambda_n}_0))}{\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(E(f^2(X_{\lambda_i+1})|X^{\lambda_i}_0) - E(f(X_{\lambda_i+1})|X^{\lambda_i}_0)^2)}} < t \middle| H_w\right) = \Phi(t).$$

The proof of Theorem 2.38 is complete.

4. Appendix

It will be useful to define other processes $\{\hat{X}^{(k)}_n\}_{n=-\infty}^{\infty}$ for $k \ge 0$ as follows. Let

$$\hat{X}^{(k)}_{-n} = X_{\lambda_k - n} \text{ for } -\infty < n < \infty.$$

Note that process $\{\hat{X}^{(k)}_n\}_{n=-\infty}^{\infty}$ is just $\{X_n\}_{n=-\infty}^{\infty}$ translated by the random amount of steps λ_k.

Lemma 4.1. *Let $\mathcal{X}$ denote a finite or countably infinite alphabet. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary process taking values from $\mathcal{X}$. Then for arbitrary $k \ge 0$, the time series $\{\hat{X}^{(k)}_n\}_{n=-\infty}^{\infty}$ and $\{X_n\}_{n=-\infty}^{\infty}$ have identical distribution.*

Proof: It is enough to show that for all $k \ge 0$, $m \ge n \ge 0$, and $x_0^n \in \mathcal{X}^{n+1}$,

$$P((\hat{X}^{(k)}_{m-n}, \ldots, \hat{X}^{(k)}_m) = x_0^n) = P(X^m_{m-n} = x_0^n).$$

This is immediate by stationarity of $\{X_n\}_{n=-\infty}^{\infty}$ and by the fact that for all $k \ge 0$, $m \ge n \ge 0$, $l \ge 0$, $x_0^n \in \mathcal{X}^{n+1}$,

$$T^l\{X^{\lambda_k+m}_{\lambda_k+m-n} = x_0^n, \lambda_k = l\} = \{X^m_{m-n} = x_0^n, \hat{\lambda}_k(X^0_{-\infty}) = -l\}.$$

(For the definition of $\hat{\lambda}$ cf. (3.1).) The proof of Lemma 4.1 is complete.

In the next lemma we will use $\hat{\lambda}$. For the definition of $\hat{\lambda}$ cf. (3.1).

Lemma 4.2. *Let $\mathcal{X}$ denote a finite or countably infinite alphabet. Let $\{X_n\}_{n=0}^{\infty}$ be a stationary process taking values from $\mathcal{X}$.Then for $k \ge 0$, almost surely,*

$$\hat{\lambda}_k(\ldots, \hat{X}^{(k)}_{-1}, \hat{X}^{(k)}_0) = \hat{\lambda}_k(\tilde{X}^0_{-\infty})$$

and

$$\tilde{X}^0_{\hat{\lambda}_k(\tilde{X}^0_{-\infty})} = \hat{X}^{(k)}_{\hat{\lambda}_k(\ldots,\hat{X}^{(k)}_{-1},\hat{X}^{(k)}_0)}, \ldots, \hat{X}^{(k)}_0.$$

Proof: Observe that for any $i \geq 0$ and for all $j \geq i$, almost surely, $\tilde{X}_{-i} = \hat{X}^{(j)}_{-i}$. (Note that $\lambda_j(X_0^\infty) - j \geq 0$.) The proof of Lemma 4.2 is complete.

Lemma 4.3. *Let $\mathcal{X}$ denote a finite or countably infinite alphabet. Let $\{X_n\}_{n=0}^\infty$ be a stationary process taking values from $\mathcal{X}$. Then the distributions of $\{\tilde{X}_n\}_{n=-\infty}^0$ and $\{X_n\}_{n=-\infty}^0$ are the same.*

Proof: This is immediate from Lemma 4.1 and Lemma 4.2. The proof of Lemma 4.3 is complete.

The main objective of Lemma 4.4 is to prove Lemma 4.5 and Lemma 4.6. (Cf. Elton, 1981 and Morvai and Weiss, 2011, Morvai and Weiss (2020a) also.)

Lemma 4.4. *Let $\mathcal{X}$ be a finite or countable alphabet, for $n = 0, 1, \ldots$ let W_n be identically distributed random objects taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a real valued function with $E(|f(W_1)|) < \infty$. Let $\mathcal{G}_n$ be an increasing sequence of σ-algebras such that $f(W_n)$ is measurable with respect to $\mathcal{G}_n$. Then*

$$E\left(\sup_{1\leq n}\left|\sum_{i=1}^n \frac{f(W_i)I_{\{|f(W_i)|\leq i\}} - E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1})}{i}\right|\right) < \infty. \tag{4.1}$$

If in addition

$$E(|f(W_0)|\log^+(|f(W_0)|)) < \infty$$

then

$$E\left(\sup_{1\leq n}\left|\sum_{i=1}^n \frac{f(W_i) - E(f(W_i)|\mathcal{G}_{i-1})}{i}\right|\right) < \infty. \tag{4.2}$$

Proof: By Davis' inequality (valid for all martingale differences cf. e.g. Shiryayev, 1984 p. 470), for some constatnt $B > 0$ we get

$$\begin{aligned}
& E\left(\sup_{1\leq n}\left|\sum_{i=1}^n \frac{f(W_i)I_{\{|f(W_i)|\leq i\}} - E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1})}{i}\right|\right) \\
\leq\ & BE\left[\left(\sum_{i=1}^\infty \frac{(f(W_i)I_{\{|f(W_i)|\leq i\}} - E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1}))^2}{i^2}\right)^{0.5}\right] \\
\leq\ & B\left[\sum_{i=1}^\infty \frac{E\left((f(W_i)I_{\{|f(W_i)|\leq i\}} - E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1}))^2\right)}{i^2}\right]^{0.5}
\end{aligned}$$

where we used Jensen's inequality. Now

$$\begin{aligned}
& E\left((f(W_i)I_{\{|f(W_i)|\leq i\}} - E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1}))^2\right) \\
=\ & E\left((f(W_i)I_{\{|f(W_i)|\leq i\}})^2\right) + E\left(E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1})^2\right) \\
-\ & 2E\left(f(W_i)I_{\{|f(W_i)|\leq i\}}E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1})\right) \\
=\ & E\left((f(W_i)I_{\{|f(W_i)|\leq i\}})^2\right) - E\left(E(f(W_i)I_{\{|f(W_i)|\leq i\}}|\mathcal{G}_{i-1})^2\right) \\
\leq\ & E\left((f(W_i)I_{\{|f(W_i)|\leq i\}})^2\right) \\
=\ & E\left((f(W_i))^2 I_{\{|f(W_i)|\leq i\}}\right).
\end{aligned}$$

Since the W_i's are identically distributed therefore

$$\begin{aligned}
\sum_{i=1}^{\infty}\frac{1}{i^2}E\left((|f(W_i)|)^2 I_{\{|f(W_i)|\le i\}}\right) &= \sum_{i=1}^{\infty}\frac{1}{i^2}E\left(|f(W_0)|^2 I_{\{|f(W_0)|\le i\}}\right)\\
&= \sum_{i=1}^{\infty}\frac{1}{i^2}\sum_{j=1}^{i}E\left(|f(W_0)|^2 I_{\{j-1<|f(W_0)|\le j\}}\right)\\
&= \sum_{i=1}^{\infty}\left(E\left(|f(W_0)|^2 I_{\{i-1<|f(W_0)|\le i\}}\right)\left(\sum_{j=i}^{\infty}\frac{1}{j^2}\right)\right)\\
&\le \sum_{i=1}^{\infty}\left(E\left(|f(W_0)|^2 I_{\{i-1<|f(W_0)|\le i\}}\right)\frac{2}{i}\right)\\
&\le 2\sum_{i=1}^{\infty}\left(E\left(|f(W_0)| I_{\{i-1<|f(W_0)|\le i\}}\right)\right)\\
&\le 2E(|f(W_0)|)\\
&< \infty
\end{aligned}$$

where we used that $\sum_{j=i}^{\infty} j^{-2}\le 2/i$. Combining all these we get (4.1), (cf. Theorem 2.19 in Hall and Heyde, 1980 also).
Now we assume that $E(|f(W_0)|\log^+(|f(W_0)|))<\infty$.

$$\begin{aligned}
E|f(W_n)I_{\{|f(W_n)|>n\}}-E(f(W_n)I_{\{|f(W_n)|>n\}}|\mathcal{G}_{n-1})| &\le 2E\left(|f(W_n)|I_{\{|f(W_n)|>n\}}\right)\\
&= 2E\left(|f(W_0)|I_{\{|f(W_0)|>n\}}\right)
\end{aligned}$$

since W_n's are identically distributed. Thus

$$\begin{aligned}
&E\left(\sum_{n=1}^{\infty}\frac{|f(W_n)I_{\{|f(W_n)|>n\}}-E(f(W_n)I_{\{|f(W_n)|>n\}}|\mathcal{G}_{n-1})|}{n}\right)\\
&\le 2\sum_{n=1}^{\infty}\frac{1}{n}E\left(|f(W_0)|I_{\{|f(W_0)|>n\}}\right).
\end{aligned}$$

Since $E((|f(W_0)|)\log^+(|f(W_0)|)<\infty$, Lemma 2 in Elton (1981) implies that

$$\sum_{n=1}^{\infty}\frac{1}{n}E\left(|f(W_0)|I_{\{|f(W_0)|>n\}}\right)<\infty$$

and so we get

$$E\left(\sum_{n=1}^{\infty}\frac{|f(W_n)I_{\{|f(W_n)|>n\}}-E(f(W_n)I_{\{|f(W_n)|>n\}}|\mathcal{G}_{n-1})|}{n}\right)<\infty. \tag{4.3}$$

Since

$$\begin{aligned}
&E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{f(W_i)I_{\{|f(W_i)|>i\}}-E(f(W_i)I_{\{|f(W_i)|>i\}}|\mathcal{G}_{i-1})}{i}\right|\right)\\
&\le E\left(\textstyle\sum_{n=1}^{\infty}\frac{|f(W_n)I_{\{|f(W_n)|>n\}}-E(f(W_n)I_{\{|f(W_n)|>n\}}|\mathcal{G}_{n-1})|}{n}\right),
\end{aligned}\tag{4.4}$$

by (4.1), (4.4) and (4.3) we get (4.2). The proof of Lemma 4.4 is complete.

The next result (with a different proof) can already be found in the proof of Theorem 2.19 in Hall and Heyde (1980). (Cf. Elton, 1981, Algoet, 1999 and Morvai and Weiss, 2011, Morvai and Weiss, 2020a also.)

Lemma 4.5. *(Cf. Theorem 2.19 in Hall and Heyde, 1980) Let $\mathcal{X}$ be a finite or countable alphabet. For $n = 1, 2, \dots$ let W_n be identically distributed random objects taking values from $\mathcal{X}$ and let $f : \mathcal{X} \to \Re$ denote a real valued function with $E(|f(W_1)|) < \infty$. For $n = 0, 1, \dots$ let $\mathcal{G}_n$ be an increasing sequence of σ-algebras. Assume that for $n = 1, 2, \dots$ $f(W_n)$ is measurable with respect to $\mathcal{G}_n$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(f(W_i) I_{\{|f(W_i)| \le i\}} - E(f(W_i) I_{\{|f(W_i)| \le i\}} | \mathcal{G}_{i-1}) \right) = 0 \tag{4.5}$$

almost surely.

Proof: Since by Lemma 4.4

$$U_n = \sum_{i=1}^{n} \frac{f(W_i) I_{\{|f(W_i)| \le i\}} - E(f(W_i) I_{\{|f(W_i)| \le i\}} | \mathcal{G}_{i-1})}{i}$$

is a martingale with

$$E\left(\sup_{1\le n} \left| \sum_{i=1}^{n} \frac{f(W_i) I_{\{|f(W_i)| \le i\}} - E(f(W_i) I_{\{|f(W_i)| \le i\}} | \mathcal{G}_{i-1})}{i} \right| \right) < \infty$$

by Doob's convergence theorem U_n converges almost surely. Then Kronecker's lemma (cf. Shiryayev, 1984 p. 365) yields (4.5). The proof of Lemma 4.5 is complete.

The second statement (4.7) in the next lemma (with a different proof) can already be found in Theorem 2.19 in Hall and Heyde (1980). (Cf. Elton, 1981 and Morvai and Weiss, 2011, Morvai and Weiss, 2020a also.)

Lemma 4.6. *Let $\mathcal{X}$ be a finite or countable alphabet. For $n = 1, \dots$ let W_n be identically distributed random objects taking values from $\mathcal{X}$. Let $f : \mathcal{X} \to \Re$ denote a real valued function with $E(|f(W_0)| \log^+(|f(W_0)|)) < \infty$. For $n = 0, 1, \dots$ let $\mathcal{G}_n$ be an increasing sequence of σ-algebras. For $n = 1, 2, \dots$ assume that $f(W_n)$ is measurable with respect to $\mathcal{G}_n$. Then*

$$E\left(\sup_{1\le n} \left| \frac{1}{n} \sum_{i=1}^{n} (f(W_i) - E(f(W_i)|\mathcal{G}_{i-1})) \right| \right) < \infty \tag{4.6}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (f(W_i) - E(f(W_i)|\mathcal{G}_{i-1})) = 0 \tag{4.7}$$

almost surely.

Proof: Since by Lemma 4.4

$$U_n = \sum_{i=1}^{n} \frac{f(W_i) - E(f(W_i)|\mathcal{G}_{i-1})}{i}$$

is a martingale with

$$E(\sup_{1\le n} |U_n|) < \infty$$

and since for any sequence of real numbers $\{a_i\}$,

$$\sup_{1\le n} \frac{1}{n} \left| \sum_{i=1}^{n} a_i \right| \le 2 \left(\sup_{1\le n} \left| \sum_{i=1}^{n} \frac{1}{i} a_i \right| \right),$$

(cf. Lemma 7 in Elton, 1981), we get (4.6). By Doob's convergence theorem U_n converges almost surely. Then Kronecker's lemma (cf. Shiryayev, 1984, p. 365) yields (4.7). The proof of Lemma 4.6 is complete.

Lemma 4.7. *Let $\mathcal{X}$ denote a finite or countably infinite alphabet. Let $\{X_n\}_{n=0}^{\infty}$ be a stationary and ergodic finitarily Markovian (FM) process taking values from $\mathcal{X}$. Assume that $E|f(X_1)| < \infty$. Then for $n, m > K(X_{-\infty}^0)$,*

$$|E(f(X_1)|X_{-n}^0) - E(f(X_1)|X_{-m}^0)| = 0.$$

Proof: The statement follows immediately from the definitions. The proof of Lemma 4.7 is complete.

References

Algoet, P. Universal schemes for prediction, gambling and portfolio selection. *Ann. Probab.*, **20** (2), 901–941 (1992). MR1159579 with a correction at MR1330780.

Algoet, P. Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Trans. Inform. Theory*, **45** (4), 1165–1185 (1999). MR1686250.

Algoet, P. H. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory*, **40** (3), 609–633 (1994). MR1295308.

Bailey, D. H. *Sequential Schemes for Classifying and Predicting Ergodic Processes*. ProQuest LLC, Ann Arbor, MI (1976). Thesis (Ph.D.)–Stanford University. MR2626644.

Chow, Y. S. and Teicher, H. *Probability theory: Independence, interchangeability, martingales*. Springer-Verlag, New York-Heidelberg (1978). ISBN 0-387-90331-3. MR513230.

Cover, T. M. Open problems in information theory. In *Proceedings of the IEEE-USSR Joint Workshop in Information Theory (Moscow, 1975)*, pp. 35–36 (1976). MR0469507.

Elton, J. A law of large numbers for identically distributed martingale differences. *Ann. Probab.*, **9** (3), 405–412 (1981). MR614626.

Felber, T., Jones, D., Kohler, M., and Walk, H. 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** (10), 1689–1707 (2013). DOI: 10.1016/j.jspi.2013.06.002.

Györfi, L. and Lugosi, G. Strategies for sequential prediction of stationary time series. In *Modeling uncertainty*, volume 46 of *Internat. Ser. Oper. Res. Management Sci.*, pp. 225–248. Kluwer Acad. Publ., Boston, MA (2002). MR1893282.

Györfi, L., Morvai, G., and Yakowitz, S. J. Limits to consistent on-line forecasting for ergodic time series. *IEEE Trans. Inform. Theory*, **44** (2), 886–892 (1998). MR1607704.

Györfi, L. and Ottucsák, G. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, **53** (5), 1866–1872 (2007). MR2317147.

Györfi, L., Ottucsák, G., and Walk, H. Machine Learning for Financial Engineering. Imperial College Press, London (2012). DOI: 10.1142/p818.

Hall, P. and Heyde, C. C. *Martingale limit theory and its application*. Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London (1980). ISBN 0-12-319350-8. MR624435.

Jones, D., Kohler, M., and Walk, H. Weakly universally consistent forecasting of stationary and ergodic time series. *IEEE Trans. Inform. Theory*, **58** (2), 1191–1202 (2012). MR2918019.

Kalikow, S., Katznelson, Y., and Weiss, B. Finitarily deterministic generators for zero entropy systems. *Israel J. Math.*, **79** (1), 33–45 (1992). MR1195252.

Merhav, N. and Feder, M. Universal prediction. volume 44, pp. 2124–2147 (1998). Information theory: 1948–1998. MR1658815.

Molnár-Sáska, G. and Morvai, G. Intermittent estimation for Gaussian processes. *IEEE Trans. Inform. Theory*, **56** (6), 2778–2782 (2010). MR2683434.

Morvai, G. Guessing the output of a stationary binary time series. In *Foundations of statistical inference (Shoresh, 2000)*, Contrib. Statist., pp. 207–215. Physica, Heidelberg (2003). MR2017826.

Morvai, G. and Weiss, B. Forecasting for stationary binary time series. *Acta Appl. Math.*, **79** (1-2), 25–34 (2003). MR2021874.

Morvai, G. and Weiss, B. Intermittent estimation of stationary time series. *Test*, **13** (2), 525–542 (2004). MR2154012.

Morvai, G. and Weiss, B. Forward estimation for ergodic time series. *Ann. Inst. H. Poincaré Probab. Statist.*, **41** (5), 859–870 (2005a). MR2165254.

Morvai, G. and Weiss, B. Inferring the conditional mean. *Theory Stoch. Process.*, **11** (1-2), 112–120 (2005b). MR2327452.

Morvai, G. and Weiss, B. Limitations on intermittent forecasting. *Statist. Probab. Lett.*, **72** (4), 285–290 (2005c). MR2153125.

Morvai, G. and Weiss, B. On classifying processes. *Bernoulli*, **11** (3), 523–532 (2005d). MR2146893.

Morvai, G. and Weiss, B. Order estimation of Markov chains. *IEEE Trans. Inform. Theory*, **51** (4), 1496–1497 (2005e). MR2241507.

Morvai, G. and Weiss, B. On estimating the memory for finitarily Markovian processes. *Ann. Inst. H. Poincaré Probab. Statist.*, **43** (1), 15–30 (2007a). MR2288267.

Morvai, G. and Weiss, B. On sequential estimation and prediction for discrete time series. *Stoch. Dyn.*, **7** (4), 417–437 (2007b). MR2378577.

Morvai, G. and Weiss, B. Estimating the lengths of memory words. *IEEE Trans. Inform. Theory*, **54** (8), 3804–3807 (2008). MR2451043.

Morvai, G. and Weiss, B. Nonparametric sequential prediction for stationary processes. *Ann. Probab.*, **39** (3), 1137–1160 (2011). MR2789586.

Morvai, G. and Weiss, B. A note on prediction for discrete time series. *Kybernetika (Prague)*, **48** (4), 809–823 (2012). MR3013400.

Morvai, G. and Weiss, B. Estimating the conditional expectations for continuous time stationary processes. *Kybernetika (Prague)*, **56** (3), 410–431 (2020a). MR4131737.

Morvai, G. and Weiss, B. Universal rates for estimating the residual waiting time in an intermittent way. *Kybernetika (Prague)*, **56** (4), 601–616 (2020b). MR4168527.

Morvai, G. and Weiss, B. On universal algorithms for classifying and predicting stationary processes. *Probab. Surv.*, **18**, 77–131 (2021). MR4255241.

Morvai, G., Yakowitz, S., and Györfi, L. Nonparametric inference for ergodic, stationary time series. *Ann. Statist.*, **24** (1), 370–379 (1996). MR1389896.

Morvai, G., Yakowitz, S. J., and Algoet, P. Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inform. Theory*, **43** (2), 483–498 (1997). MR1447529.

Nobel, A. B. On optimal sequential prediction for general processes. *IEEE Trans. Inform. Theory*, **49** (1), 83–98 (2003). MR1965889.

Ornstein, D. S. *Ergodic theory, randomness, and dynamical systems.* Yale Mathematical Monographs, No. 5. Yale University Press, New Haven, Conn.-London (1974). MR0447525.

Ornstein, D. S. Guessing the next output of a stationary process. *Israel J. Math.*, **30** (3), 292–296 (1978). MR508271.

Ryabko, B. Prediction of random sequences and universal coding. *Problems of Inform. Trans.*, **24** (2), 87–96 (1988). MR955983.

Ryabko, B. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In *2008 IEEE Information Theory Workshop*, pp. 271–275 (2008). DOI: 10.1109/ITW.2008.4578667.

Ryabko, B. and Monarev, V. A. Experimental investigation of forecasting methods based on data compression algorithms. *Problems of Information Transmission*, **41** (1), 65–69 (2005). DOI: 10.1007/s11122-005-0011-9.

Ryabko, D. On asymptotic and finite-time optimality of Bayesian predictors. *J. Mach. Learn. Res.*, **20**, Paper No. 149, pp. 1–24 (2019). MR4030163.

Ryabko, D. and Ryabko, B. Predicting the outcomes of every process for which an asymptotically accurate stationary predictor exists is impossible. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1204–1206 (2015). DOI: 10.1109/ISIT.2015.7282646.

Schäfer, D. Strongly consistent online forecasting of centered Gaussian processes. *IEEE Trans. Inform. Theory*, **48** (3), 791–799 (2002). MR1889985.

Shields, P. C. Cutting and stacking: a method for constructing stationary processes. *IEEE Trans. Inform. Theory*, **37** (6), 1605–1617 (1991). MR1134300.

Shiryayev, A. N. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York (1984). ISBN 0-387-90898-6. MR737192.

Suzuki, J. Universal prediction and universal coding. *Systems and Computers in Japan*, **34** (6), 1–11 (2003). DOI: 10.1002/scj.10357.

Takahashi, H. Computational limits to nonparametric estimation for ergodic processes. *IEEE Trans. Inform. Theory*, **57** (10), 6995–6999 (2011). MR2882275.

Weiss, B. *Single orbit dynamics*, volume 95 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI (2000). ISBN 0-8218-0414-6. MR1727510.