The Annals of Statistics
1995, Vol. 23, No. 1, 132–143

EFFICIENCY OF EMPIRICAL ESTIMATORS FOR MARKOV CHAINS[1]

BY P. E. GREENWOOD AND W. WEFELMEYER

University of British Columbia and University of Siegen

Suppose we observe a uniformly ergodic Markov chain with unknown transition distribution. The *empirical estimator* for a linear functional of the (invariant) joint distribution of two successive observations is defined using the pairs of successive observations. Its efficiency is proved using a martingale approximation. As a corollary we show efficiency of the empirical joint distribution function in the sense of a functional convolution theorem.

1. Introduction. Before introducing empirical estimators for Markov chains let us recall what this term means if our observations $X_1, \ldots, X_n$ are i.i.d. from an unknown distribution P. The distribution P is determined by values $P(A)$ assigned to certain sets A. The *empirical estimator* for $P(A)$ is the proportion of observations falling into A,

$$J_n^1(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \in A),$$

which is the probability of A under the *empirical distribution*

$$J_n^1(dx) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(dx).$$

The estimator $J_n^1(A)$ is known to be efficient for $P(A)$ in nonparametric models. Here a nonparametric model is a model described by the family of all distributions on some measurable space, or by a family which is dense in this family in an appropriate sense. The result is due to Levit (1974) and Koshevnik and Levit (1976). See also the monographs by Pfanzagl and Wefelmeyer (1982) and Bickel, Klaassen, Ritov and Wellner (1993).

Suppose now that our observations $X_0, \ldots, X_n$ are from a stationary Markov chain with transition distribution $P(x, dy)$ and invariant probability measure $\pi(dx)$. Let $P_2(dx, dy) = \pi(dx)P(x, dy)$ denote the joint distribution of two successive observations. The distribution of the process is determined by conditional probabilities $P_2(A \times B)/\pi(A)$ for certain sets A and B. Since π is the marginal distribution of P_2, the distribution of the process is determined by probabilities $P_2(A \times B)$. The *empirical estimator* for

Received May 1992; revised January 1994.
[1]Work supported by NSERC, Canada.
AMS 1991 *subject classifications*. Primary 62G20, 62M05.
Key words and phrases. Efficiency, empirical estimator, Markov chain.

$P_2(A \times B)$ is the proportion of times it happens that an observation falls into A and the following observation falls into B,

$$J_n^2(A \times B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_{i-1} \in A, X_i \in B),$$

which is the probability of $A \times B$ under the *empirical distribution*

$$J_n^2(dx, dy) = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_{i-1}, X_i)}(dx, dy).$$

If, instead of being stationary, the process starts at X_0 with an arbitrary initial distribution, under mild assumptions the distribution of (X_{i-1}, X_i) will converge rapidly to the invariant joint distribution P_2, and J_n^2 will estimate P_2 consistently.

The *empirical estimator* for the expected value Ef under P_2 of a function $f(x, y)$ is

$$J_n^2(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_{i-1}, X_i).$$

For example, the *least squares estimator*

$$\frac{\sum_{i=1}^{n} X_{i-1} X_i}{\sum_{i=1}^{n} X_{i-1}^2}$$

estimates Ef/Eg, where $f(x, y) = xy$ and $g(x, y) = x^2$. Another example involving empirical estimators is a misspecified Markov chain model. Suppose we have specified a parametric family of transition distributions $P_\vartheta(x, dy)$, while in reality the transition distribution may be arbitrary. Assign to each transition distribution P the parameter which minimizes the Kullback–Leibler distance to the parametric model. This defines a functional of P and hence of P_2. It is estimated by the maximum likelihood estimator based on the parametric model. In fact, the maximum likelihood estimator is obtained by applying the functional to J_n^2. Hence efficiency of the maximum likelihood estimator under misspecification would follow from efficiency of J_n^2. See Greenwood and Wefelmeyer (1993).

The latter example was our original motivation for the question: Is the empirical distribution J_n^2 efficient? A question related to efficiency of J_n^2 is studied by Penev (1991), who proves for uniformly ergodic Markov chains that the usual empirical distribution J_n^1 efficiently estimates the invariant measure. Note that J_n^1 is the marginal of J_n^2. We prove efficiency of J_n^2 by a different and simpler approach. Compare Bickel (1993).

We use an efficiency concept based on a nonparametric version of Hájek's convolution theorem. It refers to a class of regular estimators. Regularity is defined in terms of a certain local model. The local model is chosen to fulfill two requirements. It is large enough for the variance bound to be attainable, namely, by the empirical estimator. It is small enough not to exclude reason-

able estimators as long as it is contained in the given model; see the paragraph on possible choices of nonparametric model in Section 2.

Our proof of efficiency extends easily to functions of $k+1$ or less arguments if the model includes Markov chains of kth order rather than only first order. Consider, for instance, a function $f(x, y, z)$ of three arguments. The empirical estimator

$$J_n^3(f) = \frac{1}{n-1} \sum_{i=2}^{n} f(X_{i-2}, X_{i-1}, X_i)$$

is efficient for Ef in the model consisting of second-order Markov chains by the version of our result for $k = 2$. The model consisting of *first*-order Markov chains is, however, a strict submodel, and we do not expect the empirical estimator $J_n^3(f)$ to remain efficient in the smaller model. This can be verified by calculating the variance bound for estimators of Ef in the model consisting of first-order Markov chains.

Let us compare the above with the situation in the i.i.d. case where only the empirical distribution J_n^1 is needed: $J_n^1 \times \cdots \times J_n^1$ is efficient for the joint distribution $P \times \cdots \times P$. For Markov chains, J_n^1 is efficient for $P_1 = \pi$, but the efficient estimator for P_2 is J_n^2, and J_n^3 is *not* efficient for the joint distribution P_3 of three successive observations.

2. Results. Let $X_0, \ldots, X_n$ be observations from a Markov chain with values in an arbitrary state space E with countably generated σ-field. Let $P(x, dy)$ denote the transition distribution, and $\mu(dx)$ the initial distribution. Suppose that the chain is ergodic and uniformly ergodic. Then there is a unique invariant probability measure $\pi(dx)$ and an a in $(0, 1)$ such that

$$\|P^k - \Pi\| \le a^k, \tag{2.1}$$

with $\Pi(x, dy) = \pi(dy)$ the invariant projection of P. Here the norm of a transition distribution P is the operator norm

$$\|P\| = \sup\{\|\mu P\|: \|\mu\| \le 1\},$$

with $\mu P(dy) = \int \mu(dx) P(x, dy)$, and $\|\mu\|$ the variation norm of the finite signed measure μ.

Consider the function space

$$B = \{f: E^2 \to \mathbb{R}, \text{ bounded, measurable}\}.$$

For functions f in B, the expectation under the invariant joint distribution of two successive observations is

$$Ef = \int\int \pi(dx) P(x, dy) f(x, y). \tag{2.2}$$

We use the following notation for conditional expectations of functions of more than one variable. For a function $f(x, y)$ and $k = 1, 2, \ldots,$

$$E_{x_0}^k f = \int \cdots \int P(x_0, dx_1) \cdots P(x_{k-1}, dx_k) f(x_{k-1}, x_k).$$

In particular, $E_x f = \int P(x, dy) f(x, y)$. For a function $f(x, y) = f(y)$ of one variable, $E_x^k f = \int P^k(x, dy) f(y)$. We introduce a subspace of B:

$$H = \{f \in B,\ E_x f = 0 \text{ for all } x \text{ in } E\}.$$

Let $f(x, y)$ be a function in B and consider the stochastic process

$$\sum_{i=1}^{n} (f(X_{i-1}, X_i) - Ef).$$

For unbounded f we refer to Greenwood and Wefelmeyer (1993). Our first assertion identifies a function Af in H such that one can replace $f - Ef$ by Af in this process with error uniformly of order $\log n$. The process thus created is a martingale since $E_x Af = 0$ for all x in E. A nonuniform version of this martingale approximation for arbitrary stationary sequences is due to Gordin (1969).

LEMMA 1. *The operator A defined by*

$$(Af)(x, y) = f(x, y) - E_x f + \sum_{k=1}^{\infty} \left(E_y^k f - E_x^{k+1} f\right)$$

is a linear operator mapping B into H, is the identity on H, and fulfills

$$\sum_{i=1}^{n} (f(x_{i-1}, x_i) - Ef - (Af)(x_{i-1}, x_i)) = O(\log n)$$

uniformly for any uniformly bounded set of functions f in B and uniformly over sequences $x_0, \ldots$.

Lemma 1 is proved in Section 3. Our application of Lemma 1 to the problem of estimating Ef efficiently uses only the rate $o(n^{1/2})$.

The efficiency argument begins with a version of *local asymptotic normality*. For h in H, let the transition distribution P_{nh} be defined by

$$P_{nh}(x, dy)/P(x, dy) = 1 + n^{-1/2} h(x, y).$$

Let P^n and P^{nh} denote the joint distributions of the first $n + 1$ observations $X_0, \ldots, X_n$ if P and P_{nh}, respectively, are true and the initial distribution is μ. Since the Markov chain is ergodic under the transition distribution P, we have $(1/n)\sum_{i=1}^{n} h^2(X_{i-1}, X_i) \to Eh^2$, and the log-likelihood ratio has the stochastic approximation

$$\log dP^{nh}/dP^n = n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) - \tfrac{1}{2} Eh^2 + o_{P^n}(1).$$

The process $\sum_{i-1}^{n} h(X_{i-1}, X_i)$ is a martingale because $E_x h = 0$. By a martingale central limit theorem [see, e.g., Jacod and Shiryaev (1987), page 448, Remark 3.77.2], the sum converges in distribution,

$$n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) \Rightarrow N_h, \tag{2.3}$$

where N_h is normal and has mean 0 and variance Eh^2. One can, alternatively, use a central limit theorem for Markov chains [e.g., Ibragimov and Linnik (1971), page 368], once having noticed that the variance given there reduces to Eh^2 for h in H. Local asymptotic normality for Markov chains is basically due to Roussas (1965); see also Roussas [(1972), pages 53ff.] and Penev (1991).

Now we show that, for f in B, the expectation Ef is a differentiable functional of the transition distribution P. We define $E_{nh}f$ as in (2.2), with P replaced by P_{nh}, and π replaced by the corresponding invariant probability measure π_{nh}.

LEMMA 2. *For f in B and h in H,*

$$n^{1/2}(E_{nh}f - Ef) \to E(hAf).$$

Lemma 2 is proved in Section 3. It says that the expectation Ef is *differentiable* at P with *gradient* Af in H. The gradient is unique. If we think of H as embedded in $L_2(P_2)$, then any function g in $L_2(P_2)$ with $g - Af$ orthogonal to H is also a gradient, and we would distinguish Af by calling it *canonical*.

With Lemma 2, the convolution theorem for regular estimators of Ef has the following version. We call an estimator T_n *regular* for Ef at P with *limit* L if, for all h in H,

$$n^{1/2}(T_n - E_{nh}f) \Rightarrow L \quad \text{under } P^{nh}.$$

Then

$$\left(n^{-1/2}\sum_{i=1}^{n}(Af)(X_{i-1}, X_i),\right.$$

$$\left. n^{1/2}(T_n - Ef) - n^{-1/2}\sum_{i=1}^{n}(Af)(X_{i-1}, X_i)\right) \tag{2.4}$$

$$\Rightarrow (N_{Af}, M) \quad \text{under } P^n,$$

with M independent of N_{Af}. In particular,

$$L = N_{Af} + M. \tag{2.5}$$

This justifies calling an estimator T_n *efficient* for Ef at P if

$$n^{1/2}(T_n - Ef) \Rightarrow N_{Af} \quad \text{under } P^n.$$

As another consequence of (2.4), an estimator T_n is regular and efficient for Ef at P if and only if it is *asymptotically linear* for Ef at P with *influence function* the gradient, Af,

$$n^{1/2}(T_n - Ef) = n^{-1/2}\sum_{i=1}^{n}(Af)(X_{i-1}, X_i) + o_{P^n}(1). \tag{2.6}$$

For estimators of the parameter in a one-dimensional model, relation (2.4) is due to Ibragimov and Has'minskii [(1981), page 155, Theorem 9.2]. It is a

refinement of the classical convolution theorem, relation (2.5), which is due to Hájek (1970). A version of (2.4) for differentiable functionals, including the characterization (2.6), may be found, for example, in Greenwood and Wefelmeyer [(1990), pages 359ff.].

The empirical estimator for Ef with f in B is

$$J_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_{i-1}, X_i).$$

In the Introduction we have written J_n^2 for J_n. With regularity and efficiency defined as above, our theorem now follows immediately from Lemma 1 and the characterization (2.6).

THEOREM. *For f in B, the empirical estimator $J_n(f)$ is a regular and efficient estimator of the linear functional Ef at P.*

From Lemma 1 and the central limit theorem for Markov chains referred to after (2.3), we obtain an explicit form of the asymptotic variance of $J_n(f)$ and hence of the variance bound $E(Af)^2$, namely,

$$E(Af)^2 = Ef^2 - (Ef)^2 + 2 \sum_{k=1}^{\infty} \left(Ef(X_0, X_1) f(X_k, X_{k+1}) - (Ef)^2 \right). \tag{2.7}$$

The proof of efficiency of empirical estimators given above is particularly simple in several respects.

1. Differentiability of Ef and the form of the gradient Af follow immediately from Lemma 1 and Le Cam's third lemma. See the proof of Lemma 2. A direct but more tedious proof can be obtained by expanding $\pi' = \pi_{nh}$ around π, using

$$\pi' = \pi + \pi(P' - P)R + o(\|P' - P\|)$$

with

$$R = (I - P + \Pi)^{-1} = I + \sum_{k=1}^{\infty} (P^k - \Pi).$$

See Kartashov [(1985a), page 87, or (1985b), page 251] for these expansions.

2. Using Lemmas 1 and 2, we saw that the empirical estimator $J_n(f)$ is efficient without first calculating the variance bound $E(Af)^2$. Then we obtained the explicit form of $E(Af)^2$ from Lemma 1 and the central limit theorem for Markov chains without calculation. A different, tedious, approach is to calculate $E(Af)^2$ and compare it with the asymptotic variance of $J_n(f)$.
3. Regularity of $J_n(f)$ follows from the characterization (2.6). A direct proof would include showing asymptotic normality of $J_n(f)$ under all sequences P^{nh} in the local model.

For a function f of one argument, $J_n(f)$ reduces to the expectation of f under the usual empirical distribution. For such estimators, efficiency was proved by Penev [(1991); see also (1990a, b)]. His treatment differs from ours in several respects, including the three aspects listed above. Penev takes the state space to be the unit interval. For general state space, see van der Vaart and Wellner (1989). They also prove a weaker version of our Lemma 1 for functions of one variable.

From Lemma 1 one obtains a functional version of the convolution theorem (2.4). Suppose we want to estimate a function of the form $f \to Ef$, with f running through some index class of uniformly bounded functions. For a functional version of (2.4) we need tightness of the sequence of processes indexed by f,

$$f \to n^{-1/2} \sum_{i=1}^{n} (Af)(X_{i-1}, X_i).$$

By Lemma 1 this holds if we have tightness for the corresponding empirical processes

$$f \to n^{-1/2} \sum_{i=1}^{n} (f(X_{i-1}, X_i) - Ef).$$

The latter is, in general, easier to check since the functions f are usually simpler than the functions Af.

Possible choices of nonparametric model. The results stated so far refer to a fixed uniformly ergodic transition distribution and a specific local model. Hence they are not only applicable to the model of *all* uniformly ergodic Markov chains. In fact, the results are true for *any* model containing the fixed transition distribution. However, the local model P^{nh} with h in H is chosen with "nonparametric" models in mind. In general, our local model may not be contained in the underlying model. In that case it will treat competing estimators unfairly by demanding that the be regular in directions h outside the model. We refer to Penev (1991) for conditions under which the local model lies in the underlying model.

Stationary Markov chains. To simplify the exposition, we have assumed that the initial distribution μ is fixed. The results remain true for all models in which the initial distribution is not assumed to vary too strongly with the transition distribution. More specifically, we have to exclude cases in which the first observation X_0 carries a nonnegligible amount of information as compared to the rest of the observations $X_1, \ldots, X_n$. When the initial distribution is *fixed*, this is trivially true: the initial distribution cancels in the likelihood ratio between P^{nh} and P^n. If the process is *stationary*, the initial distribution is the invariant probability measure. It depends continuously on the transition distribution [see Kartashov (1985a), page 74, Theorem 4, or

(1985b), page 251, Theorem B]. Hence the factor $d\pi_{nh}/d\pi$ in the likelihood ratio is asymptotically negligible.

The i.i.d. case. It is interesting to see how the argument presented here works in the i.i.d. case. Suppose $X_1, \ldots, X_n$ are i.i.d. with distribution $P(dx)$. Introduce a local model $dP_{nh}/dP = 1 + n^{-1/2}h$ with $h(x)$ bounded and $Eh = 0$. Then we have local asymptotic normality,

$$\log dP^{nh}/dP^n = n^{-1/2} \sum_{i=1}^{n} h(X_i) - \tfrac{1}{2}Eh^2 + o_{P^n}(1).$$

Let $f(x)$ be bounded, and consider Ef as a linear functional of the distribution P. The empirical estimator is $(1/n)\sum_{i=1}^{n} f(X_i)$. Lemmas 1 and 2 hold trivially with $Af = f - Ef$ and imply that the influence function of the empirical estimator is the gradient of the functional Ef, which means the estimator is regular and efficient for Ef.

Note that in the i.i.d. case we can have

$$n^{-1/2} \sum_{i=1}^{n} f(X_i) = o_{P^n}(1)$$

only if $f = 0$, while for a Markov chain we have

$$n^{-1/2} \sum_{i=1}^{n} f(X_{i-1}, X_i) = o_{P^n}(1)$$

if and only if $Ef = 0$ and $Af = 0$. This implies $f = 0$ if and only if f is in H. In the i.i.d. case the role of the local parameter space H is played by the bounded functions $h(x)$ with $Eh = 0$. The local parameters are characterized as the functions for which $\sum_{i=1}^{n} f(X_i)$ and $\sum_{i=1}^{n} f(X_{i-1}, X_i)$, respectively, are martingales. To get asymptotic linearity of an empirical estimator, we must replace the function $f - Ef$ by a function in the local parameter space. This is easy in the i.i.d. case: the function is already in this space.

EXAMPLE. Suppose the state space is finite-dimensional. The distribution function F for the joint distribution of (X_0, X_1) is defined by

$$F(s,t) = E\mathbf{1}(X_0 \le s, X_1 \le t).$$

The empirical distribution function F_n is defined by

$$F_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_{i-1} \le s, X_i \le t).$$

The convolution theorem (2.4) implies its own multivariate version, as shown, for example, in Greenwood and Wefelmeyer [(1990), page 362, Corollary 2.22]. From Lemma 1 and the multivariate version of (2.4) we obtain joint efficiency for finitely many pairs (s,t).

To prove that F_n is efficient for F in a *functional* sense, we need a functional version of the convolution theorem (2.5). We want to apply results for *stationary* φ-mixing sequences. Hence we assume that the Markov chain is stationary. Then (2.5) still holds, as noted in the paragraph on stationary Markov chains. Call an estimator T_n *regular* for F at P with *limit* L if $T_n(s,t)$ is regular for $F(s,t)$ with limit $L(s,t)$ for each pair (s,t), and $n^{1/2}(T_n - F)$ is tight under P^n. Since the Markov chain X_i is uniformly ergodic, the sequence (X_{i-1}, X_i) is φ-mixing with exponential rate. The process $n^{1/2}(F_n - F)$ is the (usual) empirical process for the sequence (X_{i-1}, X_i). Tightness of $n^{1/2}(F_n - F)$ is now implied by a functional central limit theorem for the empirical process for a multivariate stationary φ-mixing sequence. See Sen (1974) or Yoshihara (1974) for continuous distribution function F. For a general recent result see Arcones and Yu (1994). Alternatively, one could use, for example, the almost-sure invariance principle for the empirical process for a multivariate strongly mixing sequence in Philipp (1984). In particular, F_n is regular, and a functional version of (2.5) holds:

$$L = N + M,$$

where N is a Gaussian process with covariance function equal to the limiting covariance function of F_n, and M is independent of N. For the functional version of the convolution theorem, see Schick and Susarla (1990).

As a special case we obtain efficiency of the usual empirical distribution function for i.i.d. observations [Beran (1977) and Millar (1985)]. As noted above, Lemmas 1 and 2 are trivial in this case. Hence efficiency and regularity of the empirical distribution function are immediate.

3. Proofs of the lemmas.

PROOF OF LEMMA 1. The following relations hold uniformly for $|f| \le 1$ and sequences $x_0, \ldots$. By (2.1), for c sufficiently large,

$$\left| \sum_{k > c \log n} \sum_{i=1}^{n} \left(E_{x_i}^k f - E_{x_{i-1}}^{k+1} f \right) \right| \le 2n \sum_{k > c \log n} a^k \to 0.$$

Hence

$$\sum_{i=1}^{n} (Af)(x_{i-1}, x_i) = \sum_{i=1}^{n} \left(f(x_{i-1}, x_i) - E_{x_{i-1}} f \right) + \sum_{k \le c \log n} \sum_{i=1}^{n} \left(E_{x_i}^k f - E_{x_{i-1}}^{k+1} f \right) + o(1).$$

Rearranging the sums, we see that the conditional expectations $E_{x_i}^k$ cancel except for $i = 0$ and $i = n$. Hence the right-hand side equals

$$\sum_{i=1}^{n} \left(f(x_{i-1}, x_i) - E_{x_{i-1}}^{(c \log n)+1} f \right) + \sum_{k \le c \log n} \left(E_{x_n}^k f - E_{x_0}^k f \right).$$

The second sum is of order $\log n$ since f is bounded. The result now follows by replacing $E_{x_{i-1}}^{(c \log n)+1} f$ by Ef. The error is negligible by a second application of (2.1),

$$\left| \sum_{i=1}^{n} \left(E_{x_{i-1}}^{(c \log n)+1} f - Ef \right) \right| \leq n a^{(c \log n)+1} \to 0. \qquad \square$$

PROOF OF LEMMA 2. Assume for simplicity that $|h| \leq 1$ and $|f| \leq 1$. It will be convenient to consider local asymptotic normality with π_{nh} as initial distribution. We use the following stability result of Kartashov [(1985a), Theorem 6]. There exists an $\varepsilon > 0$ and c such that $\|P' - P\| < \varepsilon$ implies

$$\|\pi' - \pi\| \leq c\|P' - P\|, \qquad \|P'^k - P^k\| \leq c\|P' - P\|.$$

If ε is chosen small enough, we obtain from (2.1) that there exists $a_\varepsilon < 1$ such that

$$\|P'^k - \Pi'\| \leq a_\varepsilon^k. \tag{3.1}$$

By definition of the transition distribution P_{nh} we have

$$\|P_{nh} - P\| \leq n^{-1/2}.$$

In particular, $\|\pi_{nh} - \pi\| \to 0$, and local asymptotic normality remains true as written if the initial distribution is π_{nh}. The proof of Lemma 2 is based on a contiguity argument. By the Cramér–Wold theorem, $n^{-1/2}\sum_{i=1}^{n} h(X_{i-1}, X_i)$ and $n^{-1/2}\sum_{i=1}^{n} (Af)(X_{i-1}, X_i)$ are jointly asymptotically normal under P^n with variances Eh^2 and $E(Af)^2$ and covariance $E(hAf)$. Since we have local asymptotic normality, Le Cam's third lemma implies

$$n^{-1/2} \sum_{i=1}^{n} (Af)(X_{i-1}, X_i) \Rightarrow N_{Af} + E(hAf) \quad \text{under } P^{nh}, \tag{3.2}$$

where N_{Af} is normal with mean 0 and variance $E(Af)^2$. By local asymptotic normality, P^{nh} is contiguous to P^n. Hence the martingale approximation in Lemma 1 and (3.2) imply

$$n^{-1/2} \sum_{i=1}^{n} (f(X_{i-1}, X_i) - Ef) \Rightarrow N_{Af} + E(hAf) \quad \text{under } P^{nh}. \tag{3.3}$$

Write the random variable in (3.3) as

$$n^{-1/2} \sum_{i=1}^{n} (f(X_{i-1}, X_i) - E_{nh} f) + n^{1/2}(E_{nh} f - Ef).$$

Since the initial distribution is π_{nh}, the chain is stationary, and $n^{1/2}(E_{nh} f - Ef)$ is the mean of the above random variable. We want to show that this mean converges to the mean $E(hAf)$ of the limit distribution. This follows if $n^{-1/2}\sum_{i=1}^{n}(f(X_{i-1}, X_i) - E_{nh} f)$ is uniformly integrable under P^{nh}. Uniform

integrability follows using (3.1) and a moment inequality for φ-mixing sequences [Ibragimov and Linnik (1971), page 309]:

$$
\begin{aligned}
&E^{nh}\left(\left|n^{-1/2}\sum_{i=1}^{n}\left(f(X_{i-1},X_i)-E_{nh}f\right)\right|\right.\\
&\qquad\left.\times 1\left\{\left|n^{-1/2}\sum_{i=1}^{n}\left(f(X_{i-1},X_i)-E_{nh}f\right)\right|>c\right\}\right)\\
&\le c^{-1}E^{nh}\left(n^{-1/2}\sum_{i=1}^{n}\left(f(X_{i-1},X_i)-E_{nh}f\right)\right)^2\\
&= c^{-1}n^{-1}\sum_{i,j=1}^{n}E^{nh}\left(f(X_{i-1},X_i)-E_{nh}f\right)\left(f(X_{j-1},X_j)-E_{nh}f\right)\\
&\le c^{-1}\left(1+16\sum_{k=0}^{\infty}a_{\varepsilon}^{k}\right). \qquad \square
\end{aligned}
$$

Acknowledgments. The comments of a referee and an Associate Editor have led to considerable improvements in the presentation of the paper. They are greatly appreciated.

REFERENCES

ARCONES, M. A. and YU, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *J. Theoret. Probab.* **7** 47–71.

BERAN, R. (1977). Estimating a distribution function. *Ann. Statist.* **5** 400–404.

BICKEL, P. J. (1993). Estimation in semiparametric models. In *Multivariate Analysis: Future Directions* (C. R. Rao, ed.) 55–73. North-Holland, Amsterdam.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press.

GORDIN, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.* **10** 1174–1176.

GREENWOOD, P. E. and WEFELMEYER, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* **36** 353–370.

GREENWOOD, P. E. and WEFELMEYER, W. (1993). Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models. Unpublished manuscript.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation. Asymptotic Theory.* Springer, New York.

IBRAGIMOV, I. A. and LINNIK, YU. V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes.* Springer, New York.

KARTASHOV, N. V. (1985a). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* **30** 71–89.

KARTASHOV, N. V. (1985b). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I. *Theory Probab. Appl.* **30** 247–259.

KOSHEVNIK, YU. A. and LEVIT, B. YA. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.

LEVIT, B. YA. (1974). On optimality of some statistical estimates. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, ed.) **2** 215–238. Charles Univ., Prague.

MILLAR, P. W. (1985). Non-parametric applications of an infinite dimensional convolution theorem. *Z. Wahrsch. Verw. Gebiete* **68** 545–556.

PENEV, S. (1990a). Convolution theorem for estimating the stationary distribution of Markov chains. *Doklady Bolgarskoj Akademii Nauk* **43** 29–32.

PENEV, S. (1990b). Stability of non-parametric procedures against Markov dependence. Preprint, Inst. Appl. Math. Informatics, Tech. Univ. Sofia.

PENEV, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27** 105–123.

PFANZAGL, J. and WEFELMEYER, W. (1982). *Contributions to a General Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

PHILIPP, W. (1984). Invariance principles for sums of mixing random elements and the multivariate empirical process. In *Limit Theorems in Probability and Statistics* (P. Révész, ed.) **2** 843–873. *Colloq.* North-Holland, Amsterdam.

ROUSSAS, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.* **36** 978–992.

ROUSSAS, G. G. (1972). *Contiguity of Probability Measures.* Cambridge Univ. Press.

SCHICK, A. and SUSARLA, V. (1990). An infinite dimensional convolution theorem with applications to random censoring and missing data models. *J. Statist. Plann. Inference* **24** 13–23.

SEN, P. K. (1974). Weak convergence of multivariate empirical processes for stationary φ-mixing processes. *Ann. Probab.* **2** 147–154.

VAN DER VAART, A. W. and WELLNER, J. A. (1989). Prohorov and continuous mapping theorems in the Hoffmann–Jørgensen weak convergence theory, with applications to convolution and asymptotic minimax theorems. Unpublished manuscript.

YOSHIHARA, K. (1974). Extensions of Billingsley's theorems on weak convergence of empirical processes. *Z. Wahrsch. Verw. Gebiete* **29** 87–92.

DEPARTMENT OF MATHEMATICSS
UNIVERSITY OF BRITISH COLUMBIA
121–1984 MATHEMATICS ROAD
VANCOUVER, BRITISH COLUMBIA
CANADA V6T 1Z2

FB 6 MATHEMATIC
UNIVERSITY OF SIEGEN
HOELDERLIN STREET 3
57068 SIEGEN
GERMANY