The Annals of Probability
1994, Vol. 22, No. 2, 960–977

ENTROPY AND THE CONSISTENT ESTIMATION OF JOINT DISTRIBUTIONS

KATALIN MARTON[1] AND PAUL C. SHIELDS[2]

Hungarian Academy of Sciences, and University of Toledo and Eötvös Loránd University

The kth-order joint distribution for an ergodic finite-alphabet process can be estimated from a sample path of length n by sliding a window of length k along the sample path and counting frequencies of k-blocks. In this paper the problem of consistent estimation when $k = k(n)$ grows as a function of n is addressed. It is shown that the variational distance between the true $k(n)$-block distribution and the empirical $k(n)$-block distribution goes to 0 almost surely for the class of weak Bernoulli processes, provided $k(n) \leq (\log n)/(H + \varepsilon)$, where H is the entropy of the process. The weak Bernoulli class includes the i.i.d. processes, the aperiodic Markov chains and functions thereof and the aperiodic renewal processes. A similar result is also shown to hold for functions of irreducible Markov chains. This work sharpens prior results obtained for more general classes of processes by Ornstein and Weiss and by Ornstein and Shields, which used the $\bar{d}$-distance rather than the variational distance.

1. Introduction. The kth-order joint distribution for an ergodic finite-alphabet process can be estimated from a sample path of length n by sliding a window of length k along the sample path and counting frequencies of k-blocks. If k is fixed the procedure is consistent in that the resulting empirical k-block distribution will almost surely converge to the true distribution of k-blocks as $n \to \infty$, a fact guaranteed by the ergodic theorem. The consistency of such estimates is important when using training sequences, that is, finite sample paths, to design engineering systems. The empirical k-block distribution for a training sequence is used as the basis for design, after which the system is run on other, independently drawn sample paths. There are some situations, such as data compression, where it is good to make the block length as long as possible. Thus it would be desirable to have consistency results for the case when the block length function $k = k(n)$ grows as rapidly as possible, as a function of sample path length n. This is the problem addressed in this paper. (Rigorous definitions and terminology will be given in Section 2.)

A sequence $\{k(n)\}$ will be said to be *admissible* for a given ergodic process μ if the variational distance between the true distribution and the empirical distribution of $k(n)$-blocks converges almost surely to 0 as $n \to \infty$. Every ergodic

Received June 1992.

[1]Partially supported by Hungarian National Foundation for Scientific Research Grant OTKA 1906.

[2]Partially supported by NSF Grant DMS-90-24240.

AMS 1991 *subject classifications.* Primary 28D20, 60J05, 62B20; secondary 60G10, 94A17.

Key words and phrases. Empirical distribution, entropy, weak Bernoulli processes.

process had an admissible sequence such that $\lim_n k(n) = \infty$, by the ergodic theorem. It is also not hard to see that for any sequence $k(n) \to \infty$ there is an ergodic measure for which $\{k(n)\}$ is not admissible.

The problem addressed in this paper is whether it is possible to make a universal choice of $\{k(n)\}$, provided we restrict to some "nice" class of processes, such as i.i.d. processes or Markov chains. Here is where entropy enters the picture, for if k is large, then, with high probability, the probability of a k-block will be roughly 2^{-kH}; thus if $k(n) \geq (1+\varepsilon)(\log n)/H$, then we have no hopes that the empirical k-block distribution will be close to the true distribution. Consistent estimation also may not be possible for the choice $k(n) \sim (\log n)/H$. For example, in the unbiased coin-tossing case when $H = 1$, the choice $k(n) \sim \log n$ is not admissible, for it is easy to see that, with high probability, an approximate $(1 - e^{-1})$ fraction of the k-blocks will fail to appear in a given sample path of length n.

In this paper we consider the case when

$$k(n) \sim (1 - \varepsilon)\frac{\log n}{H}, \tag{1}$$

where μ is an ergodic process of entropy H. Our principal results may be (informally) stated as follows:

1. If μ is i.i.d., then the variational distance between the true distribution and the empirical distribution of $k(n)$-blocks converges to 0 almost surely, provided (1) holds.
2. The preceding result also holds for irreducible Markov chains, for functions of irreducible Markov chains, for ψ-mixing processes and for weak Bernoulli processes.

The ψ-mixing and weak Bernoulli concepts are generalizations of the property that past and distant future become asymptotically independent.

Our first motivation for this paper was the training sequence problem described in the opening paragraph. A second motivation was a desire to obtain a more classical version of the positive result obtained by Ornstein and Weiss, who used the $\bar{d}$-distance rather than the variational distance [Ornstein and Weiss (1990)]. They showed that if the process is finitely determined, then the $\bar{d}$-distance between the empirical $k(n)$-block distribution and the true $k(n)$-block distribution goes to 0, almost surely, provided $k(n) \sim (\log n)/H$. The finitely determined processes are just the "almost aperiodic Markov" processes in that thay are precisely the $\bar{d}$-limits of aperiodic multistep Markov chains. The $\bar{d}$-distance is bounded above by the variational distance; thus our results are a sharpening of the Ornstein–Weiss result for the case when $k \leq (1 - \varepsilon)(\log n)/H$ and the process satisfies strong enough forms of asymptotic independence. Furthermore, it can be shown that our results imply their $\bar{d}$-results even for the case $k(n) \sim (\log n)/H$.

A third motivation for this paper was the result about waiting times obtained by Wyner and Ziv (1989). They showed that if $W_n(x,y)$ is the waiting time until the first n terms of x appear in the sequence y, then, for ergodic Markov

chains, $(1/n)\log W_n(x,y)$ converges in probability to H, provided x and y are independently chosen sample paths. The results of this paper can be used to prove stronger versions of their theorem. These applications along with various counterexamples are presented in a separate paper [Shields (1993)].

Definitions and precise statements will be given in the next section. Proofs will be given in Section 3. Application of our results to the Ornstein–Weiss $\bar{d}$-estimation problem will be discussed in Section 4.

2. Definitions and statements of results. For our purposes a process is a shift-invariant Borel probability measure μ on the space A^∞ of sequences $x = \{x_n\}$, drawn from a finite alphabet A. The shift $T = T_A$ is defined by $(T_A x)_n = x_{n+1}, x \in A^\infty$, the sequence $a_m, a_{m+1}, \ldots, a_n$ will be denoted by a_m^n, the set of such a_m^n will be denoted by A_m^n, A^n will denote A_1^n, and $[a_m^n]$ will denote the cylinder set defined by a_m^n, that is,

$$[a_m^n] = \{x \in A^\infty\colon x_i = a_i, m \le i \le n\}.$$

The field generated by the cylinder sets $[a_m^n]$, for fixed $m \le n$, will be denoted by $\mathcal{F}_m^n$. The complement of the set B will be denoted by B^c.

The process μ defines a measure μ_k on A^k by the formula

$$\mu_k(a_1^k) = \mu([a_1^k]) = P(\{x\colon x_i = a_i, 1 \le i \le k\}).$$

We call μ_k the *true (or theoretical) distribution of k-blocks*. When k is understood we sometimes write μ instead of μ_k. Note that a process is i.i.d. if and only if each μ_k is the product measure defined by μ_1. A process μ with alphabet A is called a *function of a Markov chain*, or a finite-state process or hidden Markov chain, if there is a Markov chain ν with alphabet B and a function $f\colon B \mapsto A$ such that $\mu = \nu \circ F^{-1}$, where $F\colon B^\infty \mapsto A^\infty$ is defined by $y = F(x)$, where $y_n = f(x_n), n \ge 1$.

The entropy of a process μ will be denoted by H_μ or by H, if μ is understood, and is defined by the limit

$$H = \lim_n \frac{H_n}{n}, \qquad H_n = -\sum_{a_1^n} \mu_n(a_1^n)\log \mu_n(a_1^n),$$

where here and elsewhere in this paper base 2 logarithms will be used. (See Billingsley (1965), for a discussion of the entropy concept.) To avoid uninteresting cases, we make the following assumption throughout this paper, unless stated otherwise.

ASSUMPTION. Our processes are *ergodic and have positive entropy.*

Let $x_1^n \in A^n$ and let $k \le n$. For each $a_1^k \in A^k$ define

$$f(a_1^k \mid x_1^n) = \left|\{i \in [1, n-k+1]\colon x_i^{i+k-1} = a_1^k\}\right|,$$

that is, the number of occurrences of a_1^k in x_1^n. The *empirical k-block distribution* is the measure $\hat{\mu}_k(\cdot|x_1^n)$ on A^k defined by

$$\hat{\mu}_k(a_1^k \mid x_1^n) = \frac{f(a_1^k \mid x_1^n)}{n-k+1}, \qquad a_1^k \in A^k.$$

In cases where x_1^n is understood we use the simpler notation $\hat{\mu}_k$.

The variational distance between two measures p and q on A^k is defined by

$$|p-q| = \sum_{a_1^k} |p(a_1^k) - q(a_1^k)|.$$

A nondecreasing sequence $k(n) \leq n$ will be said to be *admissible* for the ergodic measure μ if

$$\lim_{n\to\infty} |\hat{\mu}_{k(n)}(\cdot|x_1^n) - \mu_{k(n)}| = 0, \quad \text{a.s.}$$

Our first two results can now be stated.

THEOREM 1. *If μ is ergodic with positive entropy and $k(n) \geq (\log n)/(H-\varepsilon)$, then $k(n)$ is not admissible for μ.*

THEOREM 2. *If μ is i.i.d. and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

Our extension to the Markov and related cases will be expressed in terms of an asymptotic independence property known as ψ-mixing. A process is *ψ-mixing* if there is a sequence $\psi(g)\downarrow 1$ such that, for all $m, n \geq 1$,

$$\mu(C \cap D) \leq \psi(g)\mu(C)\mu(D), \qquad C \in \mathcal{F}_1^m, \qquad D \in \mathcal{F}_{g+m+1}^{g+m+n}.$$

It is easy to see that aperiodic Markov chains, as well as functions thereof, are ψ-mixing.

THEOREM 3. *If μ is ψ-mixing and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

A modification of our techniques will enable us to extend Theorem 3 to certain nonmixing processes, such as ergodic Markov chains and functions of ergodic Markov chains; we state this as the following corollary.

COROLLARY 1. *If μ is a function of an irreducible Markov chain and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

The concept of admissibility does not specify the speed with which $|\hat{\mu}_{k(n)}(\cdot|x_1^n) - \mu_{k(n)}|$ goes to 0. With only a bit more effort our method will yield a speed of convergence result, at least for functions of Markov chains. Sharper results are

possible but we will prove only what is needed for the application to waiting-time problems given in Shields (1993). To state this corollary, we define the set

$$B_{k,n}(\delta) = \{x_1^n\colon\ |\hat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta\}.$$

COROLLARY 2. *If μ is a function of an irreducible Markov chain and $k(n) \leq (\log n)/(H+\varepsilon)$, then $\mu(B_{k(n),n}(1/k(n)^2))$ is summable in n, for each $\varepsilon > 0$.*

A weaker property than ψ-mixing is the weak Bernoulli property, a concept introduced in Friedman and Ornstein (1970) as part of the proof that aperiodic Markov chains are isomorphic to i.i.d. processes in the sense of ergodic theory. A process is *weak Bernoulli* if

$$\lim_{g\to\infty} \sum_{a_1^n,\, b_1^m} \left|\mu\Big(\big(T^{-(g+m)}[a_1^n]\big) \cap [b_1^m]\Big) - \mu\big([a_1^n]\big)\mu\big([b_1^m]\big)\Big|\right| = 0, \tag{2}$$

uniformly in n and m.

THEOREM 4. *If μ is weak Bernoulli and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

It is easy to see that ψ-mixing implies weak Bernoulli, so Theorem 4 includes Theorem 3. We have chosen to establish the ψ-mixing result separately, however, for two reasons. First, the ψ-mixing proof is simpler and shows clearly the basic ideas that are used to establish the weak Bernoulli result. Second, only simple modifications of the ψ-mixing proof are required for the rate result, Corollary 2.

Further notation and definitions will be introduced as needed. Proofs will be given in Section 3, and connections with the Ornstein–Weiss $\bar{d}$-versions will be discussed in Section 4.

3. Proofs of results. The set

$$T(k,\varepsilon) = \left\{x_1^k\colon\ 2^{-k(H+\varepsilon)} \leq \mu(x_1^k) \leq 2^{-k(H-\varepsilon)}\right\} \tag{3}$$

will play an important role in several of our proofs. The entropy theorem (also known as the Shannon–McMillan–Breiman theorem) guarantees that for every $\varepsilon > 0$ and for almost all x there is a $K = K(x,\varepsilon)$ such that $x_1^k \in T(k,\varepsilon)$ for $k \geq K$. In particular, $\mu(T(k,\varepsilon)) \to 1$ for any $\varepsilon > 0$. Furthermore, the lower bound on the probability yields the following upper bound on cardinality:

$$|T(k,\varepsilon)| \leq 2^{k(H+\varepsilon)}. \tag{4}$$

We think of $T(k,\varepsilon)$ as the "typical" k-blocks.

THEOREM 1. *If μ is ergodic with positive entropy and $k(n) \geq (\log n)/(H-\varepsilon)$, then $k(n)$ is not admissible for μ.*

PROOF. Assume $k \geq (\log n)/(H - \varepsilon)$ and note that there are no more than $2^{k(H-\varepsilon)}$ distinct k-blocks in x_1^n. The theorem follows from the fact that when k is large we have no hopes of seeing all the typical k-blocks. The precise proof goes as follows. Define

$$\mathcal{U}_k(x_1^n) = \{a_1^k: x_i^{i+k-1} = a_1^k \text{ for some } i \in [1, n-k+1]\}.$$

We think of $\mathcal{U}_k(x_1^n)$ as the empirical *universe* of k-blocks determined by x_1^n. Note that

$$\hat{\mu}_k(a_1^k \mid x_1^n) = 0, \qquad a_1^k \notin \mathcal{U}_k(x_1^n),$$

and hence

$$|\hat{\mu}_k - \mu_k| \geq \mu\Big(\big(\mathcal{U}_k(x_1^n)\big)^c\Big). \tag{5}$$

Our assumption $n \leq 2^{k(H-\varepsilon)}$ implies that $|\mathcal{U}_k(x_1^n)| \leq 2^{k(H-\varepsilon)}$, so that, since each member of $T_k(\varepsilon/2)$ has measure at most $2^{-k(H-\varepsilon/2)}$, the following holds:

$$\mu\Big(\mathcal{U}_k(x_1^n) \cap T_k(\varepsilon/2)\Big) \leq 2^{k(H-\varepsilon)}2^{-k(H-\varepsilon/2)} = 2^{-k\varepsilon/2}.$$

Since $\mu(T_k(\varepsilon/2))$ goes to 1, we therefore know that $\mu((\mathcal{U}_k(x_1^n))^c)$ also goes to 1, which implies the theorem, since (5) holds. Note that we actually proved that $|\hat{\mu}_{k(n)}(\cdot|x_1^n) - \mu_{k(n)}|$ does not even go to 0 in probability. □

Now we turn to the positive admissibility results. The key to these is the following lemma, which is essentially just a combination of a large deviations bound of Hoeffding (1965) and Sanov (1957) and an inequality of Pinsker (1960). We sketch a proof which is based on Lemmas 2.2 and 2.6 and Exercise 3.17 of Csiszár and Körner (1981).

LEMMA 1. *There is a positive constant c such that for any finite set A and i.i.d. process μ with alphabet A the following holds:*

$$\mu\big(\{x_1^n: |\hat{\mu}_1 - \mu_1| \geq \varepsilon\}\big) \leq (n+1)^{|A|}2^{-nc\varepsilon^2}. \tag{6}$$

PROOF. The first key to the lemma is the following reexpression of the measure of x_1^n:

$$\mu(x_1^n) = \prod_{a \in A} \big(\mu(a)\big)^{f(a|x_1^n)} = 2^{-n(\hat{H}_1 + D(\hat{\mu}_1 \| \mu_1))}, \tag{7}$$

where

$$\hat{H}_1 = -\sum_a \hat{\mu}_1(a) \log \hat{\mu}_1(a),$$

is the entropy of the empirical 1-block distribution and

$$D(\hat{\mu}_1\|\mu_1) = \sum_a \hat{\mu}_1(a)\log\frac{\hat{\mu}_1(a)}{\mu_1(a)}$$

is the so-called divergence of $\hat{\mu}_1$ relative to μ_1. Let $\hat{\mu}(\cdot|x_1^n)$ be the product measure of A^n defined by the first-order empirical distribution $\hat{\mu}(\cdot|x_1^n)$ and use this in place of μ in the product formula (7) to obtain

$$\hat{\mu}(x_1^n \mid x_1^n) = 2^{-n\hat{H}_1}. \tag{8}$$

The second key to the lemma is to note that both (7) and (8) depend only on the empirical 1-block distribution and not on anything else about x_1^n. Let us say that x_1^n is equivalent to y_1^n if $\hat{\mu}_1(\cdot|x_1^n) = \hat{\mu}_1(\cdot|y_1^n)$ and denote the equivalence class of x_1^n by $\mathcal{E}(x_1^n)$. Both μ and $\hat{\mu}(\cdot|x_1^n)$ are then constant on $\mathcal{E}(x_1^n)$. Furthermore, since $\hat{\mu}(\cdot|x_1^n)$ is a probability measure and because of (8), we must have

$$\hat{\mu}\Big(\mathcal{E}(x_1^n) \mid x_1^n\Big) = \big|\mathcal{E}(x_1^n)\big|2^{-n\hat{H}_1} \le 1,$$

so that $|\mathcal{E}(x_1^n)| \le 2^{n\hat{H}_1}$. We can combine this with (7) to obtain the following bound:

$$\mu\Big(\mathcal{E}(x_1^n)\Big) \le 2^{-nD(\hat{\mu}_1\|\mu_1)}. \tag{9}$$

The third key to the lemma is the observation that

(10) There are at most $(n+1)^{|A|}$ equivalence classes.

This is because each empirical 1-block probability has the form q/n, where q is an integer in the range $0 \le q \le n$.

Now, to complete the proof, define

$$B(n,\varepsilon) = \big\{x_1^n\colon\ |\hat{\mu}_1 - \mu_1| \ge \varepsilon\big\}$$

and partition this into disjoint sets of the form $B(n,\varepsilon) \cap \mathcal{E}(x_1^n)$. Then combine (9) and (10) to obtain the bound

$$\mu\big(B(n,\varepsilon)\big) \le (n+1)^{|A|}2^{-nD_*},$$

where

$$D_* = \min\big\{D(\hat{\mu}_1\|\mu_1)\colon\ \big|\hat{\mu}_1 - \mu_1\big| \ge \varepsilon\big\}.$$

It is not hard to see that the worst case occurs when the alphabet is binary and a calculus argument [see Csiszár and Körner (1981), Exercise 3.17] then yields the bound

$$D_* \ge \frac{1}{2\log 2}\big|\hat{\mu}_1 - \mu_1\big|^2,$$

proving the lemma. □

The preceding lemma is enough to obtain the admissibility theorem we want for the case of unbiased coin tossing; the following extended form will be needed in order to obtain admissibility results for other i.i.d. processes and for more general processes.

LEMMA 2. *There is a positive constant C such that for any ε in the range $0 < \varepsilon < 0.25$, for any finite set A, for any $n > 0$, for any i.i.d. process μ with finite alphabet A and for any $B \subset A$ such that $\mu(B) > 1 - \varepsilon$ and $|B| \geq 2$, the following holds:*

$$\mu\big(\{x_1^n\colon\ |\hat{\mu}_1 - \mu_1| > 5\varepsilon\}\big) \leq (n+1)^{|B|}2^{-nC\varepsilon^2}.$$

PROOF. Define

$$y_n = y_n(x) = \begin{cases} 0, & \text{if } x_n \in B, \\ 1, & \text{otherwise,} \end{cases}$$

so that $\{y_n\}$ is a binary i.i.d. process with $\text{prob}(y_n = 1) < \varepsilon$. Let

$$C_n = \left\{x_1^n\colon \sum_1^n y_i > 2\varepsilon n\right\}$$

and apply Lemma 1 to obtain

$$\mu(C_n) \leq (n+1)^2 2^{-cn4\varepsilon^2}. \tag{11}$$

For each $m \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_m \leq n$, let $A(i_1, \ldots, i_m)$ denote the set of all x_1^n such that

$$x_i \notin B, \qquad i \in \{i_1, i_2, \ldots, i_m\}$$

and

$$x_i \in B, \qquad i \notin \{i_1, i_2, \ldots, i_m\}.$$

The sets $A(i_1, \ldots, i_m)$ are disjoint for different $\{i_1, i_2, \ldots, i_m\}$ and have union A^n. Furthermore, the sets $\{A(i_1, \ldots, i_m)\colon m > 2\varepsilon n\}$ have union C_n so that (11) yields the bound

$$\sum_{\{i_1,\ldots,i_m\}\colon m > 2\varepsilon n} \mu\big(A(i_1, \ldots, i_m)\big) \leq (n+1)^2 2^{-cn4\varepsilon^2}. \tag{12}$$

Fix a set $\{i_1, \ldots, i_m\}$ with $m \leq 2\varepsilon n$, put $s = n - m$ and for $x_1^n \in A(i_1, \ldots, i_m)$ define $\overline{x} = \overline{x}(x_1^n)$ to be the sequence of length s obtained by deleting $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ from x_1^n. Let $\mu_{1,B}$ be the conditional distribution on B defined by μ_1 and let μ_B be the corresponding product measure on B^s defined by $\mu_{1,B}$. We then have

$$\begin{aligned} &\mu\Big(\big\{x_1^n \in A(i_1, \ldots, i_m)\colon\ |\hat{\mu}_1 - \mu_1| > 5\varepsilon\big\}\Big) \\ &\qquad \leq \mu\Big(A(i_1, \ldots, i_m)\Big)\mu_B\Big(\big\{\overline{x}_1^s \in B^s\colon\ |\hat{\mu}_1(\cdot|\overline{x}_1^s) - \mu_1| > 3\varepsilon\big\}\Big). \end{aligned} \tag{13}$$

Since

$$\begin{aligned}|\mu_1 - \mu_{1,B}| &= \sum_{b\in B}\left[\frac{\mu(b)}{\mu(B)} - \mu(b)\right] + \sum_{b\notin B}\mu(b)\\ &= \left(\frac{1}{\mu(B)} - 1\right)\mu(B) + 1 - \mu(B) = 2\big(1 - \mu(B)\big) \le 2\varepsilon,\end{aligned}$$

we can apply Lemma 1 and continue (13) to obtain

$$\begin{aligned}&\mu\big(\{x_1^n \in A(i_1,\dots,i_m)\colon |\widehat{\mu}_1 - \mu_1| > 5\varepsilon\}\big)\\ &\quad\le \mu\big(A(i_1,\dots,i_m)\big)\mu_B\big(\{\overline{x}_1^s \in B^s\colon |\widehat{\mu}_1(\cdot|\overline{x}_1^s) - \mu_{1,B}| > \varepsilon\}\big)\\ &\quad\le \mu\big(A(i_1,\dots,i_m)\big)(s+1)^{|B|}2^{-sc\varepsilon^2}\\ &\quad\le \mu\big(A(i_1,\dots,i_m)\big)(n+1)^{|B|}2^{-n(1-2\varepsilon)c\varepsilon^2}.\end{aligned}$$

The sum of the $\mu(A(i_1,\dots,i_m))$ for which $m \le 2\varepsilon n$ cannot exceed 1. This, combined with (12) and our assumption that $\varepsilon < 0.25$, then establishes the lemma. □

To apply the preceding lemma to obtain our positive admissibility results, we shall need to convert from the distribution of overlapping k-blocks to that of nonoverlapping k-blocks. To assist in this task, we develop some notation and terminology. Given $k \le n/2$, define integers $t = t(n,k)$ and $u \in [0,k)$ such that $n = tk + k + u$. For each $r \in [1,k]$ define

$$f^r(a_1^k \mid x_1^n) = \big|\{j \in [1,t]\colon x_{r+(j-1)k}^{r+(j-1)k+k-1} = a_1^k\}\big|,$$

that is, we partition x_1^n into blocks of length k, starting at x_r, and count the number of blocks that agree with a_1^k, ignoring the final k-block if $r \le u$ (so that we always divide by the same amount to obtain relative frequencies). The *r-shifted nonoverlapping empirical k-block distribution* $\widehat{\mu}_k^r = \widehat{\mu}_k^r(\cdot|x_1^n)$ is the distribution on A^k defined by

$$\widehat{\mu}_k^r(a_1^k) = \frac{f^r(a_1^k \mid x_1^n)}{t}.$$

The overlapping-block measure $\widehat{\mu}_k$ is (almost) an average of the measures $\widehat{\mu}_k^r$, where "almost" is needed to account for end effects; hence, the following result holds.

LEMMA 3. *Given $\delta > 0$, there is a $\gamma > 0$ such that if $k/n < \gamma$ and $|\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \ge \gamma$, then there is an $r \in [1,k]$ such that $|\widehat{\mu}_k^r(\cdot|x_1^n) - \mu_k| \ge \delta/2$.*

We are now ready to prove our first positive theorem. We give this proof in detail and then show how the ideas can be modified to yield our other theorems.

THEOREM 2. *If μ is i.i.d. and $k(n) \le (\log n)/(H + \varepsilon)$, then $k(n)$ is admissible for μ.*

PROOF. Let δ be a given positive number. We will show that

$$\sum_n \mu\big(\{x_1^n\colon |\widehat{\mu}_{k(n)} - \mu_{k(n)}| > \delta\}\big) < \infty, \tag{14}$$

which immediately implies the theorem. The idea of the proof is to use Lemma 3 to replace overlapping blocks by nonoverlapping blocks and then apply Lemma 2 with A replaced by A^k and B replaced by a suitably chosen set $T = T(k, \varepsilon')$ of typical sequences. Since we can control the number of typical sequences, we can guarantee that the polynomial factor in the key bound (6), which now essentially has the form $(1+n/k)^{2^{k(H-\varepsilon')}}$, is dominated by the exponential factor. The rigorous proof is given in the following paragraphs.

From Lemma 3 we can assume n is large enough so that, for $k = k(n)$ and $n = tk + k + u$, $0 \le u < k$, the following holds:

$$\{x_1^n\colon |\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \ge \delta\} \subseteq \bigcup_{r=1}^{k} \{x_1^n\colon |\widehat{\mu}_k^r(\cdot|x_1^n) - \mu_k| \ge \delta/2\}. \tag{15}$$

Fix such an n and for $1 \le r \le k$ define

$$\widetilde{x}_j(r) = x_{r+(j-1)k}^{r+(j-1)k+k-1}, \qquad 1 \le j \le t.$$

Since μ is a product measure we have

$$\sum_{x_1^{r-1}} \mu(x_1^n) \le \mu\left(\bigcap_{j=1}^{t} [\widetilde{x}_j(r)]\right) = \prod_{j=1}^{t} \mu(\widetilde{x}_j(r)). \tag{16}$$

Thus, if we let $\widetilde{A} = A^k$ and define the measure μ^* on $\widetilde{A}^t$ by the formula

$$\mu^*(y_1^t) = \prod_{j=1}^{t} \mu(y_j), \qquad y_j \in \widetilde{A},$$

then the following holds:

$$\mu\Big(\{x_1^n\colon |\widehat{\mu}_k^r(\cdot|x_1^n) - \mu_k| \ge \delta/2\}\Big) \le \mu^*\big(\{y_1^t\colon |\widehat{\mu}_1^* - \mu_1^*| \ge \delta/2\}\big). \tag{17}$$

Moreover, we can assume that n is large enough so that, for $k = k(n)$, the set $T(k, \varepsilon/2)$ of typical sequences, defined by (3), satisfies $\mu(T(k, \varepsilon/2)) > 1 - \delta/10$. Since $T(k, \varepsilon/2) \subseteq A^k = \widetilde{A}$ we can replace μ by μ^* and obtain

$$\mu^*\big(T(k, \varepsilon/2)\big) > 1 - \delta/10. \tag{18}$$

Now we can apply Lemma 2 with A replaced by $\widetilde{A}$, n by t, B by $T(k, \varepsilon/2)$ and ε by $\delta/10$, together with the typical sequence bound (4),

$$|T(k, \varepsilon/2)| \le 2^{k(H+\varepsilon/2)},$$

to obtain the bound

$$\mu^*\big(\{y_1^t\colon\ |\widehat{\mu}_1^* - \mu_1^*| \geq \delta/2\}\big) \leq (t+1)^{2^{k(H+\epsilon/2)}} 2^{-tC\delta^2/100}.$$

We can combine this with (17) and (15) to obtain the bound

$$\mu\big(\{x_1^n\colon\ |\widehat{\mu}_{k(n)} - \mu_{k(n)}| > \delta\}\big) \leq k(n)(t+1)^{2^{k(n)(H+\epsilon/2)}} 2^{-tC\delta^2/100}. \tag{19}$$

Let us put

$$\alpha = C\delta^2/100 \quad \text{and} \quad \gamma = \frac{H+\varepsilon/2}{H+\varepsilon},$$

so that $\gamma < 1$. We can then use the bound $k(n) \leq (\log n)/(H+\varepsilon)$ and rewrite the right-hand side of (19) in the form

$$\frac{\log n}{H+\varepsilon}(t+1)^{n^\gamma} 2^{-\alpha t}. \tag{20}$$

This bound is summable in n, since $t \sim n/k(n)$ and $\gamma < 1$, which establishes (14) and thereby completes the proof of Theorem 2. □

REMARK 1. A somewhat sharper form of Lemma 2 can be obtained by replacing the variational bound $|\widehat{\mu}_1 - \mu_1|$ by the divergence $D(\widehat{\mu}_1\|\mu_1)$, which yields the bound

$$\mu\big(\{x_1^n\colon D(\widehat{\mu}_1\|\mu_1) \geq \varepsilon\}\big) \leq (n+1)^{|A|} 2^{-nc\varepsilon}.$$

This, in turn, leads to slightly stronger forms of our later results.

REMARK 2. It was pointed out to us by Tamás Móri that a stronger version of Theorem 2 for unbiased coin tossing was obtained by Flajolet, Kirschenhofer and Tichy (1988), who used the distance $\sup_{a_1^k} 2^k|\widehat{\mu}_k(a_1^k) - 2^{-k}|$ in place of the variational distance. For this metric a necessary and sufficient condition for admissibility is that $k(n) = \log n - \log\log n - \gamma(n)$, where $\gamma(n) \to \infty$.

REMARK 3. Results about the asymptotic distribution of $k(n)$-blocks for i.i.d. processes, for the case $k(n) \sim (c\log n)/H$ with $c \geq 1$, were obtained by Erdős and Rényi (1970).

To extend our results to the ψ-mixing case, we need a slight extension of Lemma 3, formulated so as to allow gaps between the blocks. Fix a positive integer g. For $k \leq (n-g)/2$ define $t = t(n,k,g)$ such that

$$n = t(k+g) + (k+g) + u, \qquad 0 \leq u < k+g. \tag{21}$$

For each $r \in [1, k+g]$ define

$$f_g^r(a_1^k \mid x_1^n) = \big|\{j \in [1,t]\colon x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+k-1} = a_1^k\}\big|.$$

The *r-shifted nonoverlapping empirical k-block distribution with gap g* is the distribution $\widehat{\mu}_{k,g}^r = \widehat{\mu}_{k,g}^r(\cdot|x_1^n)$ on A^k defined by

$$\widehat{\mu}_{k,g}^r(a_1^k) = \frac{f_g^r(a_1^k \mid x_1^n)}{t}.$$

Lemma 3 easily extends to the following result.

LEMMA 4. *Given $\delta > 0$ and g, there is a $\gamma > 0$ and a $K > 0$ such that if $k/n < \gamma$, $k > K$ and $|\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta$, then there is an $r \in [1, k+g]$ such that $|\widehat{\mu}_{k,g}^r(\cdot|x_1^n) - \mu_k| \geq \delta/2$.*

THEOREM 3. *If μ is ψ-mixing and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

PROOF. Our proof of Theorem 2 will be modified to allow gaps of a fixed length g, independent of k, between blocks. This will contribute an exponential factor to the key bound, (19), a factor that is dominated by the other exponential factor, provided g is large enough. The details are given in the following paragraphs.

Let $\delta > 0$ and use the ψ-mixing property to choose g so large that $\psi(g) < 2^{C\delta^2/200}$, where C is the constant of Lemma 2. From Lemma 4 we can assume that n is so large that

$$\{x_1^n\colon |\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta\} \subseteq \bigcup_{r=1}^{k+g} \{x_1^n\colon |\widehat{\mu}_{k,g}^r(\cdot|x_1^n) - \mu_k| \geq \delta/2\}. \tag{22}$$

For $n = t(k+g) + k + g + u$, where $0 \leq u < k+g$, we modify our prior notation, setting

$$\widetilde{x}_j(r) = x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+k-1}, \qquad j \in [1,t],\ r \in [1,k+g].$$

The ψ-mixing property gives the inequality

$$\mu\left(\bigcap_{j=1}^{t} [\widetilde{x}_j(r)]\right) \leq [\psi(g)]^t \prod_{j=1}^{t} \mu(\widetilde{x}_j(r)). \tag{23}$$

As in the proof of Theorem 2, we set $\widetilde{A} = A^k$ and let μ^* be the product measure on $\widetilde{A}^t$ defined by μ_k, so that (23), combined with $\psi(g) < 2^{C\delta^2/200}$, leads to the bound

$$\mu(\{x_1^n\colon |\widehat{\mu}_{k,g}^r(\cdot|x_1^n) - \mu_k| \geq \delta/2\}) \leq 2^{tC\delta^2/200}\mu^*(\{y_1^t\colon |\widehat{\mu}_1^* - \mu_1^*| \geq \delta/2\}). \tag{24}$$

As before, we can assume that $\mu(T(k,\varepsilon/2)) > 1 - \delta/10$ and apply Lemma 2, combined with (22) and (24), to obtain the bound

$$\begin{aligned}&\mu(\{x_1^n\colon |\widehat{\mu}_{k(n)} - \mu_{k(n)}| > \delta\}) \\ &\quad \leq 2^{tC\delta^2/200}[k(n)+g](t+1)^{2^{k(n)(H+\varepsilon/2)}}2^{-tC\delta^2/100}.\end{aligned} \tag{25}$$

Finally, we set

$$\alpha = \frac{C\delta^2}{200} \quad \text{and} \quad \gamma = \frac{H + \varepsilon/2}{H + \varepsilon}$$

and obtain the bound

$$\mu\big(\{x_1^n\colon\ |\widehat{\mu}_{k(n)} - \mu_{k(n)}| > \delta\}\big) \le \frac{\log n}{H+\varepsilon}(t+1)^{n^\gamma} 2^{-\alpha t}, \tag{26}$$

which, as before, is summable in n. This completes the proof of Theorem 3. □

The preceding theorem applies to aperiodic Markov chains and to functions of such chains, since these are ψ-mixing. The following corollary extends our admissibility results to periodic Markov chains and functions thereof.

COROLLARY 1. *If μ is a function of an irreducible Markov chain and $k(n) \le (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

PROOF. We need only consider the periodic case. We give the proof only for the Markov case; the extension to functions of Markov chains is straightforward. Let μ be an irreducible Markov chain with period $d > 1$ and partition A into $C_1, C_2, \ldots, C_d$ such that

$$\text{Prob}(X_{n+1} \in C_{s+1} \mid X_n \in C_s) = 1, \qquad 1 \le s \le d,$$

where addition is mod d. Define the function $c\colon A \mapsto [1,d]$ by putting $c(a) = s$ if $a \in C_s$. Also let $\mu^{(s)}$ denote the measure μ conditioned on $X_1 \in C_s$.

Let g be a gap length, which we can assume is divisible by d and small relative to $k = k(n)$. We can also assume that k is divisible by d, for we can always increase or decrease k by no more than d to achieve this, which has no effect on the asymptotics. For $r \in [1, k+g]$ the nonoverlapping block measure $\widehat{\mu}_{k,g}^r(\cdot|x_1^n)$ satisfies the following:

$$\widehat{\mu}_{k,g}^r(a_1^k) = 0 \quad \text{unless } c(a_1) = c(x_r).$$

Since μ_k is an average of the $\mu_k^{(s)}$, it follows that, if k/g is large enough, then the following will hold:

$$\{x_1^n\colon\ |\widehat{\mu}_k - \mu_k| > \delta\} \subseteq \bigcup_{r=1}^{k+g} \{x_1^n\colon\ |\widehat{\mu}_{k,g}^r - \mu_k^{(c(x_r))}| > \delta/2\}.$$

The measure $\mu^{(s)}$ is, however, an aperiodic Markov measure with state space $C(s) \times C(s+1) \times \cdots \times C(s+d-1)$, so our previous theory applies to each set $\{x_1^n\colon\ |\widehat{\mu}_{k,g}^r - \mu_k^{(c(x_r))}| > \delta/2\}$ separately; thus we can conclude that

$$\mu\big(\{x_1^n\colon\ |\widehat{\mu}_k - \mu_k| \ge \delta\}\big)$$

is summable in n, which proves the corollary. □

We next show how the argument of Theorem 3 can be refined to yield a rate result, at least for the Markov case. For this result, we use the notation

$$B_{k,n}(\delta) = \{x_1^n\colon\ |\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta\}.$$

COROLLARY 2. *If μ is a function of an irreducible Markov chain and $k(n) \leq (\log n)/(H+\varepsilon)$, then $\mu(B_{k(n),n}(1/k(n)^2))$ is summable in n, for each $\varepsilon > 0$.*

PROOF. We consider only the aperiodic Markov case; the extension to the more general case can be achieved by using the technique of the preceding corollary. Thus we assume that μ is an aperiodic Markov chain. In this case the ψ-function can be taken to have the following stronger form:

$$\psi(g) = 2^{L\lambda^g}, \qquad \lambda < 1, \tag{27}$$

where L and λ are constants that depend on μ. The corollary is obtained by allowing the gap length g to be a function of $k = k(n)$, namely, $g = \sqrt{k}$. We also define $\delta_k = 1/k^2$ and apply the argument of the theorem. Formula (25) is then replaced by the following bound, in which $k = k(n)$:

$$\mu(\{x_1^n\colon\ |\widehat{\mu}_k - \mu_k| > 1/k^2\}) \leq 2^{tL\lambda^{\sqrt{k}}}[k+\sqrt{k}](t+1)^{2^{k(H+\varepsilon/2)}}2^{-tC/100k^4}. \tag{28}$$

We can then proceed as before, using the additional information (27) to show that this is summable in n, which establishes the corollary. □

The ψ-mixing admissibility result, Theorem 3, will now be extended to the weak Bernoulli case. A key step in the ψ-mixing proof was the observation, Lemma 4, that if the nonoverlapping k-block distribution is not close to the true distribution, then the r-shifted distribution cannot be close to the true distribution for at least one $r \leq k+g$. In fact, the r-shifted distribution cannot be close to the true distribution for a *positive fraction* of the indices $r \leq k+g$. This sharper form is easy to prove; we state it as follows in the form we will use for the weak Bernoulli case.

LEMMA 5. *Given $\delta > 0$, there is a positive $\gamma < 1/2$ such that for all g there is a $K = K(g,\gamma)$ such that if $k \geq K$, if $k/n < \gamma$ and if $|\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta$, then $|\widehat{\mu}_{k+g}^r(\cdot|x_1^n) - \mu_k| \geq \delta/2$ for at least $2\gamma(k+g)$ indices $r \in [1, k+g]$.*

The second key step in the ψ-mixing proof was the product bound, (23), which was a simple consequence of the ψ-mixing condition. It is not necessary that such a bound hold for all x_1^n and all the nonoverlapping $(k+g)$-blocks, but only that it eventually almost surely holds for a large fraction of the nonoverlapping $(k+g)$-blocks. To help make this idea precise, some notation and terminology will be introduced.

Fix a positive number γ and positive integers k, g and n such that $n = t(k+g) + k + g + r$, where $t > 0$ and $0 \leq r < k+g$. For each $j \in [1,t]$ and $r \in [1, k+g]$, let

$$\widetilde{x}_j(r) = x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+k-1}.$$

A set $J = J_r \subseteq [1, t]$ will be called a *splitting set* for x_1^n, associated with the shift r, if

$$\mu\left(\bigcap_{j\in J}[\widetilde{x}_j(r)]\right) \leq (1+\gamma)^t \prod_{j\in J}\mu(\widetilde{x}_j(r)). \tag{29}$$

(To keep the notation from getting out of hand, we have suppressed the dependence of these definitions on k, g and γ.)

Our key lemma for the weak Bernoulli case asserts that, eventually almost surely, most shifts r of x_1^n will have a large splitting set J_r.

LEMMA 6. *If μ is weak Bernoulli and $0 < \gamma < 1/2$, then there is a gap $g = g(\gamma)$, there are integers $k(\gamma)$ and $t(\gamma)$ and there are sets $G_n = G_n(\gamma) \subset A^n$, $n \geq 1$, such that the following hold*:

(a) $x_1^n \in G_n$, *eventually almost surely.*

(b) *If $k \geq k(\gamma)$, if $t \geq t(\gamma)$, if $t(k+g) \leq n < (t+1)(k+g)$ and if $x_1^n \in G_n$, then there is a set $R = R(x_1^n) \subset [1, k+g]$ of cardinality at least $(1-\gamma)(k+g)$ such that associated with each $r \in R(x_1^n)$ there is a splitting set $J_r = J_r(x_1^n)$ of cardinality at least $(1-\gamma)t$.*

PROOF. First we use the weak Bernoulli property, (2), to choose $g = g(\gamma)$ so large that, for any n,

$$\sum_{x_{-g-n}^{-g}, x_1^n} \mu\left([x_{-g-n}^{-g}] \cap [x_1^n]\right)\left|1 - \frac{\mu(x_{-g-n}^{-g})\mu(x_1^n)}{\mu([x_{-g-n}^{-g}] \cap [x_1^n])}\right| < \frac{\gamma^4}{4}.$$

For g fixed the functions

$$f_n(x) = \frac{\mu(x_{-g-n}^{-g})\mu(x_1^n)}{\mu([x_{-g-n}^{-g}] \cap [x_1^n])}$$

form a martingale with respect to the σ-algebras $\mathcal{F}_n$ generated by the cylinders $[x_{-g-n}^n]$, and therefore f_n converges almost surely to some f.

Fatou's lemma implies that $\int |1 - f(x)|\, d\mu \leq \gamma^4/4$, so there is an M such that if

$$C_M = \{x\colon\ |1 - f_n(x)| \leq \gamma^2/2, \forall\ n \geq M\},$$

then $\mu(C_M) > 1 - \gamma^2/2$. The ergodic theorem implies that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=0}^{N-1} \mathcal{I}_{C_M}(T^i x) > 1 - \frac{\gamma^2}{2} \quad \text{a.s.},$$

where $\mathcal{I}_{C_M}$ denotes the indicator function of the set C_M. In particular, if

$$\widetilde{G}_n = \left\{x\colon\ \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{I}_{C_M}(T^i x) > 1 - \frac{\gamma^2}{2}\right\},$$

then $x \in \widetilde{G}_n$, eventually almost surely. Thus, if we let G_n be the projection of $\widetilde{G}_n$ onto A^n, then property (a) holds.

Let us put $k(\gamma) = M$ and let $t(\gamma)$ be any number larger than $2/\gamma^2$. Fix $k \geq k(\gamma)$, $t \geq t(\gamma)$ and $t(k+g) \leq n < (t+1)(k+g)$ and fix $x \in \widetilde{G}_n$. The definition of G_n and the assumption that $t \geq 2/\gamma^2$ imply that

$$\frac{1}{t(k+g)} \sum_{i=1}^{t(k+g)} \mathcal{I}_{C_M}\left(T^{i-1}x\right) = \frac{1}{k+g} \sum_{r=1}^{k+g} \frac{1}{t} \sum_{j=1}^{t} \mathcal{I}_{C_M}\left(T^{r+(j-1)(k+g)-1}x\right) > 1-\gamma^2,$$

so there is a subset $R = R(x) \subset [1, k+g]$ of cardinality at least $(1-\gamma)(k+g)$ such that

$$\frac{1}{t} \sum_{j=1}^{t} \mathcal{I}_{C_M}\left(T^{r+(j-1)(k+g)-1}x\right) > 1-\gamma.$$

But if this holds for a given $r \in [1, k+g]$, the definition of C_M implies that there exist $m = m_r \geq (1-\gamma)t$ indices $1 \leq j_1 < j_2 < \cdots < j_m \leq t$ such that,

$$\left|1 - \frac{\mu\left(x_{r+(j-1)(k+g)-g-N}^{r+(j-1)(k+g)-g}\right)\mu\left(x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+N-1}\right)}{\mu\left(\left[x_{r+(j-1)(k+g)-g-N}^{r+(j-1)(k+g)-g}\right] \cap \left[x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+N-1}\right]\right)}\right| < \gamma^2, \tag{30}$$

for $j \in J_r = \{j_1, \ldots, j_m\}$ and for all $N \geq M$.

Let $j = j_i < j_{i+1} = l$ be two successive members of J_r and recall that we assumed that $k \geq k(\gamma) = M$. The inequality (30), together with the assumption that $\gamma < 1/2$, implies that

$$\begin{aligned}\mu\left(x_{r+(j-1)(k+g)-g}^{n}\right) &\leq (1+\gamma)\mu\left(x_{r+(j-1)(k+g)}^{r+(j-1)(k+g)+k-1}\right)\mu\left(x_{r+(l-1)(k+g)-g}^{n}\right) \\ &= (1+\gamma)\mu\left(\widetilde{x}_j(r)\right)\mu\left(x_{r+(l-1)(k+g)-g}^{n}\right).\end{aligned}$$

Inductive application of this bound starting with $j = j_1$ produces the desired inequality, (29), thereby completing the proof of Lemma 6. □

We are now ready to prove our basic weak Bernoulli result.

THEOREM 4. *If μ is weak Bernoulli and $k(n) \leq (\log n)/(H+\varepsilon)$, then $k(n)$ is admissible for μ.*

PROOF. Fix $\delta > 0$ and choose a positive $\gamma < 1/2$ so that Lemma 4 holds. Then choose integers $g = g(\gamma)$, $k(\gamma)$ and $t(\gamma)$ and sets $G_n = G_n(\gamma) \subset A^n$, $n \geq 1$, such that conditions (a) and (b) of Lemma 6 hold. Fix $t \geq t(\gamma)$ and $t(k+g) \leq n < (t+1)(k+g)$, where $k(\gamma) \leq k = k(n) \leq (\log n)/(H+\varepsilon)$.

For each $r \in [1, k+g]$ and $J \subseteq [1,t]$, let $D_n(r,J)$ be the set of all x_1^n such that J is a splitting set for x_1^n associated with r. If $x_1^n \in G_n(\gamma)$, then Lemma 6 implies that there are at least $(1-\gamma)(k+g)$ indices $r \in [1, k+g]$ which have splitting sets of cardinality at least $(1-\gamma)t$, while, if $|\widehat{\mu}_k(\cdot|x_1^n) - \mu_k| \geq \delta$, then Lemma 4 implies that $|\widehat{\mu}^r_{k,g}(\cdot|x_1^n) - \mu_k| \geq \delta/2$ for at least $2\gamma(k+g)$ indices $r \in [1, k+g]$. Thus

$$\{x_1^n\colon\ |\widehat{\mu}_k - \mu_k| \geq \delta\} \cap G_n(\gamma) \subseteq \bigcup_{r=1}^{k+g} \bigcup_{\substack{J \subseteq [1,t] \\ |J| \geq (1-\gamma)t}} \{x_1^n\colon\ |\widehat{\mu}^r_{k,g} - \mu_k| \geq \delta/2\} \cap D_n(r,J).$$

The argument used to establish (25), with inequality (23) replaced by (29), produces the upper bound

$$2^{-2t\gamma \log \gamma}(1+\gamma)^t \big[k(n)+g\big](t+1)^{2^{k(n)(H+\epsilon/2)}} 2^{-t(1-\gamma)C\delta^2/100}, \tag{31}$$

for $\mu(\{x_1^n\colon\ |\widehat{\mu}_{k(n)} - \mu_{k(n)}| > \delta\} \cap G_n(\gamma))$, for sufficiently large t, where the extra factor, $2^{-2t\gamma \log \gamma}$, bounds the number of subsets $J \subseteq [1,t]$ for which $|J| \geq (1-\gamma)t$. If γ is small enough, then, as before, (31) will be summable in n. Since $x_1^n \in G_n$, eventually almost surely, this establishes Theorem 4. □

4. The Ornstein–Weiss problem. We now show how our results are connected to the Ornstein–Weiss $\bar{d}$-estimation problem [Ornstein and Weiss (1990)]. The $\bar{d}$-distance is defined as follows. First, the distance between two n-sequences is defined by

$$d_n(a_1^n, b_1^n) = \frac{1}{n}\sum_{i=1}^n d(a_i, b_i),$$

where $d(a,b)$ is 0 or 1, depending on whether $a = b$ or $a \neq b$. Next, let $J_n(\mu,\nu)$ be the set of all measures λ on $A^n \times A^n$ that have μ and ν as marginals and define the $\bar{d}_n$-distance by

$$\bar{d}_n(\mu,\nu) = \min_{\lambda \in J_n(\mu,\nu)} E_\lambda\Big(d_n\big(x_1^n, y_1^n\big)\Big),$$

where E_λ denotes expected value with respect to λ. If μ and ν are stationary processes with alphabet A, then

$$\bar{d}(\mu,\nu) = \lim_n \bar{d}_n(\mu_n, \nu_n),$$

a limit which can be shown to exist. The processes that are $\bar{d}$-limits of the mixing multistep Markov chains are called the "almost Markov" processes. (Other names, which arise from other characterizations of these processes, are also used, such as finitely determined, very weak Bernoulli and almost block independent.)

Let us say that a sequence $\{k(n)\}$ is $\bar{d}$-admissible for the ergodic process μ if

$$\lim_n \bar{d}_n(\widehat{\mu}_{k(n)}, \mu_{k(n)}) = 0 \quad \text{a.s.}$$

Ornstein and Weiss have shown that, if μ is a finitely determined process and $k(n) \leq (\log n)/H$, then $\{k(n)\}$ is $\bar{d}$-admissible for μ [see also Ornstein and Shields (1993) for an extension of these results]. Since we always have

$$\bar{d}_n(\mu, \nu) \leq \sum_{a_1^n} |\mu(a_1^n) - \nu(a_1^n)|,$$

our results include theirs for the case when μ is aperiodic Markov and $k(n) \leq (\log n)/(H + \varepsilon)$. It is trivial to show, however, that the class of processes for which a given sequence $\{k(n)\}$ satisfying $k(n) \leq (\log n)/(H + \varepsilon)$ is $\bar{d}$-admissible is closed in the $\bar{d}$-metric; hence our results include theirs for the general finitely determined case, at least when $k(n) \leq (\log n)/(H + \varepsilon)$. With a little effort it can be shown that a finitely determined process can always be approximated in $\bar{d}$ by multistep mixing Markov chains of smaller entropy; hence Ornstein–Weiss results can be derived from ours, even in the case $k(n) \leq (\log n)/H$.

In summary, our proofs are not as elegant as the Ornstein–Weiss proofs, but we have been able to show, at least in part, how their results are connected to approximation in variational distance, and hence that they are closely connected to classical statistical questions.

REFERENCES

BILLINGSLEY, P. (1965). *Ergodic Theory and Information.* Wiley, New York.

CSISZÁR, I. and KÖRNER, J. (1981). *Information Theory. Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest.

ERDŐS, P. and RÉNYI, A. (1970). On a new law of large numbers. *J. Analyse Math.* **23** 103–111.

FLAJOLET, P., KIRSCHENHOFER, P. and TICHY, R. F. (1988). Deviations from uniformity in random strings. *Probab. Theory Related Fields* **80** 139–150.

FRIEDMAN, N. and ORNSTEIN, D. (1970). On isomorphism of weak Bernoulli transformations. *Adv. in Math.* **5** 365–394.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–400.

ORNSTEIN, D. (1974). *Ergodic Theory, Randomness, and Dynamical Systems.* Yale Univ. Press.

ORNSTEIN, D. and SHIELDS, P. (1993). The $\bar{d}$-recognition of processes. *Adv. in Math.* To appear.

ORNSTEIN, D. and WEISS, B. (1990). How sampling reveals a process. *Ann. Probab.* **18** 905–930.

PINSKER, M. (1960). *Information and Information Stability of Random Variables and Processes. Problemy Peredači Informacii* **7**. Akad Nauk SSSR, Moscow (in Russian). [English translation (1964) by Holden-Day, San Francisco.]

SANOV, I. (1957). On the probability of large deviations of random variables. *Mat. Sb.* **42** 11–44. [English translation in *Select. Transl. Math. Statist. Probab.* **1** (1961) 213–244.]

SHIELDS, P. (1993). Waiting times: positive and negative results on the Wyner–Ziv problem. *J. Theoret. Probab.* **6** 499–519.

WYNER, A. and ZIV, J. (1989). Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory* **IT-35** 1250–1258.

MATHEMATICS INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES
P.O.B. 127
1364 BUDAPEST
HUNGARY

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606