Conditional Limit Theorems under Markov Conditioning

IMRE CSISZÁR, THOMAS M. COVER, FELLOW, IEEE,
AND BYOUNG-SEON CHOI, MEMBER, IEEE

Abstract—Let $X_1, X_2, \cdots$ be independent identically distributed random variables taking values in a finite set X and consider the conditional joint distribution of the first m elements of the sample $X_1, \cdots, X_n$ on the condition that $X_1 = x_1$ and the sliding block sample average of a function $h(\cdot,\cdot)$ defined on X^2 exceeds a threshold $\alpha > Eh(X_1, X_2)$. For m fixed and $n \to \infty$, this conditional joint distribution is shown to converge to the m-step joint distribution of a Markov chain started in x_1 which is closest to $X_1, X_2, \cdots$ in Kullback–Leibler information divergence among all Markov chains whose two-dimensional stationary distribution $P(\cdot,\cdot)$ satisfies $\sum P(x,y)h(x,y) \geq \alpha$, provided some distribution P on X^2 having equal marginals does satisfy this constraint with strict inequality. Similar conditional limit theorems are obtained when $X_1, X_2, \cdots$ is an arbitrary finite-order Markov chain and more general conditioning is allowed.

I. Introduction

SANOV'S [13] large deviation theorem for the empirical distribution $\hat{P}_n$ of an independent identically distributed (i.i.d.) sample $X_1, \cdots, X_n$ says that

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\hat{P}_n \in \Pi\} = -\inf_{P\in\Pi} D(P\|Q). \quad (1)$$

Here Π is a given set of probability distributions on the common range of the X_i's satisfying some regularity conditions, Q is the distribution of the X_i's, and $D(P\|Q)$ designates Kullback–Leibler information divergence (also called relative entropy or information for discrimination). General sufficient conditions for the limit relation (1) have been given by Groeneboom, Oosterhoff, and Ruymgaart [9].

A result closely related to (1) is the convergence of the conditional joint distribution of $X_1, \cdots, X_m$ under the condition $\hat{P}_n \in \Pi$ (for m fixed and $n \to \infty$) to the mth Cartesian power of the I-projection of Q on Π, i.e., of the distribution minimizing $D(P\|Q)$ subject to $P \in \Pi$ (cf. Csiszár [4] and previous literature cited there; the theorem in [4] covers also the case when a minimizing $P \in \Pi$ does not exist).

Manuscript received December 13, 1985; revised February 12, 1987. This paper was presented at the IEEE Symposium on Information Theory, Brighton, England, June 24–28, 1985.

I. Csiszár is with the Department of Electrical Engineering, University of Maryland, College Park, MD 20742-3011, USA, on leave from the Mathematical Institute of the Hungarian Academy of Sciences and the L. Eötvös University, Budapest, Hungary.

T. M. Cover is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA.

B. S. Choi is with the Department of Applied Statistics, Yonsei University, Seoul, 120, Korea.

IEEE Log Number 8717367.

An important special case is

$$\Pi = \{P: E_P h_j \geq \alpha_j, j = 1, \cdots, k\} \quad (2)$$

where $h_1, \cdots, h_k$ are given functions defined on the range of the X_i's and $\alpha_1, \cdots, \alpha_k$ are given constants. Then the event $A_n = \{\hat{P}_n \in \Pi\}$ is

$$A_n = \left\{\frac{1}{n}\sum_{i=1}^{n} h_j(X_i) \geq \alpha_j, j = 1, \cdots, k\right\}. \quad (3)$$

For Π as in (2), the I-projection of Q on Π belongs, under weak regularity conditions, to the exponential family through Q determined by the h_j's; i.e. $P(x) = cQ(x)\exp(\sum \lambda_j h_j(x))$. In this case, the conditional limit theorem mentioned above was established by Van Campenhout and Cover [15]. As they pointed out, this result can be construed as a justification of the maximum entropy (or minimum discrimination information) principle (cf. also Csiszár [5]).

This paper is motivated by the question of what happens if the event (3) is replaced by

$$A_n = \left\{\frac{1}{n}\sum_{i=1}^{n} h_j(X_i, X_{i+1}) \geq \alpha_j, j = 1, \cdots, k\right\}, \quad (4)$$

where $h_1, \cdots, h_k$ are given functions of two variables. This event is not determined by the empirical distribution of the sample $X_1, \cdots, X_{n+1}$; rather it depends on its second-order empirical distribution $\hat{P}_n^{(2)}$ (cf. Definition 1 in Section II). Thus we are led to consider events of form $A_n = \{\hat{P}_n^{(2)} \in \Pi\}$ where Π is now a set of two-dimensional distributions. We expect the limiting conditional distribution of $X_1, \cdots, X_m$, given A_n, to be first-order Markov. This suggests relaxing the assumption that $X_1, X_2, \cdots$ is i.i.d. to include the possibility that $X_1, X_2, \cdots$ is a Markov chain. For convenience, we restrict the state space to be finite. This enables us to use a simple but powerful counting approach (Whittle [17] and Billingsley [1]).

For the event that the second-order empirical distribution $\hat{P}_n^{(2)}$ of a finite state Markov chain with transition probability matrix W belongs to a given set Π of two-dimensional distributions, the analog of (1) is

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\hat{P}_n^{(2)} \in \Pi\} = -\min_{P\in\Pi_0} D(P\|W), \quad (5)$$

where Π_0 is the set of those distributions in the closure of Π whose two marginals are equal, and $D(P\|W)$ is defined by (12) in Section II.

0018-9448/87/1100-0788$01.00

Under suitable regularity conditions, (5) can be easily established by the mentioned counting approach (cf. Boza [2] and Natarajan [12]). Alternatively, it could be derived from the large deviation theorem of Donsker and Varadhan [8] for general Markov processes, though this would mean using much deeper tools than the problem requires.

We will weaken the regularity conditions available for (5) in a manner essential for our purposes (Lemma 2). Our main result is, however, that whenever (5) holds, the conditional joint distributions of the random variables X_i under the condition $\hat{P}_n^{(2)} \in \Pi$ approach a Markov chain determined by the $P^* \in \Pi_0$ attaining the minimum in (5), in a sense made precise in Theorems 2 and 3, provided that this P^* is unique. Simple sufficient conditions for the latter are given in Lemma 1. A corollary of our main results for conditioning on events of form (4) will be formulated as Theorem 4.

Intuitively, Theorems 2–4 provide a justification of the "maximum entropy principle" for the case of constraints on two-dimensional distributions (typically forcing dependence) in the same sense as discussed in [15] and [5] for constraints on one-dimensional distributions only. In particular, when $X_1, X_2, \cdots$ are i.i.d. and have uniform distribution, the conditional distributions converge to those of a Markov chain having maximum entropy rate among all processes with stationary two-dimensional distributions belonging to Π_0.

Our results easily extend to higher order empirical distributions and higher order Markov chains (cf. Section IV).

II. Preliminaries and Statement of Results

Let $\boldsymbol{X}$ be a finite set and let $\Lambda^{(k)}$ designate the set of all probability distributions on $\boldsymbol{X}^k$, the kth Cartesian power of $\boldsymbol{X}$. Throughout this paper, distributions on finite sets are identified with their probability mass functions. The support of any $P \in \Lambda^{(k)}$, $k = 1, 2, \cdots$, will be denoted by $S(P)$ and, for any subset Π of $\Lambda^{(k)}$, the union of the supports of all $P \in \Pi$ will be denoted by $S(\Pi)$. The cardinality of a finite set A will be denoted by $|A|$.

Definition 1: The kth-order type of a sequence $\boldsymbol{x} = (x_1, \cdots, x_{n+k-1})$ of elements of $\boldsymbol{X}$ is the distribution $P_{\boldsymbol{x}}^{(k)} \in \Lambda^{(k)}$, defined by the relative frequencies

$$P_{\boldsymbol{x}}^{(k)}(\boldsymbol{y}) = \frac{1}{n}\left|\{i \in \{1, \cdots, n\} : (x_i, \cdots, x_{i+k-1}) = \boldsymbol{y}\}\right|, \quad \boldsymbol{y} \in \boldsymbol{X}^k.$$

For a given sequence $X_1, X_2, \cdots$ of random variables with values in $\boldsymbol{X}$, the kth-order type of the sample $(X_1, \cdots, X_{n+k-1})$ is called the kth-order empirical distribution $\hat{P}_n^{(k)}$.

The first-order type (empirical distribution) is commonly called the type (empirical distribution).

In this paper, limit theorems known for first-order empirical distributions of i.i.d. sequences of random variables, summarized in Theorem 1 below, will be generalized to second and higher order empirical distributions of Markov chains. Basic for these results is Kullback–Leibler information divergence, which is a nonsymmetric measure of distance between distributions in the sense that for any two distributions P and Q on $\boldsymbol{X}^k$, say,

$$D(P\|Q) = \sum_{\boldsymbol{x} \in \boldsymbol{X}^k} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} \tag{6}$$

is nonnegative and equals 0 if and only if $P = Q$. We use logarithms to the base e, with the standard notational conventions $\log 0 = -\infty$, $\log \frac{a}{0} = \infty$ if $a > 0$, $0 \log 0 = 0 \log \frac{0}{0} = 0$.

Topological concepts for distributions will refer to the topology of pointwise convergence. The closure of any set $\Pi \subset \Lambda^{(k)}$ of distributions on $\boldsymbol{X}^k$ will be denoted by $\operatorname{cl} \Pi$.

For any fixed Q, the divergence $D(P\|Q)$ is a continuous function of P restricted to $\{P: S(P) \subset S(Q)\}$. Thus the minimum of $D(P\|Q)$ subject to $P \in \operatorname{cl} \Pi$ is attained, and if $S(\Pi) \subset S(Q)$, this minimum is the same as the infimum of $D(P\|Q)$ subject to $P \in \Pi$.

Theorem 1: Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with common distribution Q such that $S(Q) = \boldsymbol{X}$, and let $\hat{P}_n$ denote the first-order empirical distribution.

a) A necessary and sufficient condition for a set $\Pi \subset \Lambda^{(1)}$ of distributions on $\boldsymbol{X}$ to satisfy

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr\left\{\hat{P}_n \in \Pi\right\} = -\min_{P \in \operatorname{cl} \Pi} D(P\|Q) \tag{7}$$

is the existence, for every sufficiently large n, of distributions $P_n \in \Pi$ equal to the (first-order) type of some $\boldsymbol{x} \in \boldsymbol{X}^n$, such that $D(P_n\|Q)$ converges to the minimum in (7) as $n \to \infty$. A sufficient condition is that the infimum of $D(P\|Q)$ for $P \in \Pi$ be the same as for P in the interior of Π; this is satisfied if the closure of the interior of Π equals $\operatorname{cl} \Pi$.

b) If (7) holds and the I-projection P^* of Q on $\operatorname{cl} \Pi$ exists, i.e., if the minimum in (7) is attained for a unique P^*, then $\hat{P}_n$ converges to P^* in conditional probability given that $\hat{P}_n \in \Pi$, and the conditional joint distribution of $X_1, \cdots, X_m$, given that $\hat{P}_n \in \Pi$, converges to the mth Cartesian power of P^* as $n \to \infty$, for any fixed m.

Part a) of Theorem 1 dates back to Sanov [13]; the given form is effectively due to Hoeffding [10]. Part b) does not appear in the literature under precisely the above conditions but is well-known to those working in this field. The convergence of $\hat{P}_n$ to P^* in conditional probability given that $\hat{P}_n \in \Pi$ has been termed a "conditional law of large numbers" by Vasicek [16] because it means that for every function h on $\boldsymbol{X}$, the sample average $n^{-1}\sum_{i=1}^n h(X_i)$ converges to $E_{P^*}h = \sum_{x \in \boldsymbol{X}} P^*(x)h(x)$ in conditional probability given that $\hat{P}_n \in \Pi$. Following a referee's suggestion, we will give a proof of Theorem 1, preceding the proof of our new results, to exhibit the main ideas in this simple case free of technical difficulties.

In the rest of this paper, unless stated otherwise, $X_1, X_2, \cdots$ will be a Markov chain with state space $\boldsymbol{X}$, stationary transition probabilities $W(\cdot|\cdot)$, and initial distri-

bution $Q^{(1)}$:

$$\Pr\{X_1 = x_1, \cdots, X_{n+1} = x_{n+1}\} = Q^1(x_1) \prod_{i=1}^{n} W(x_{i+1}|x_i). \tag{8}$$

Clearly, the probability (8) depends on $\boldsymbol{x} = (x_1, \cdots, x_{n+1})$ only through its first element and second-order type. For convenience, we assume that the initial probabilities $Q^1(x)$, $x \in \boldsymbol{X}$, are all positive. The transition probability matrix W may have zero entries; i.e.,

$$S(W) = \{(x, y): W(y|x) > 0\} \tag{9}$$

may be a proper subset of $\boldsymbol{X}^2$. At this point, we do not even require the irreducibility of the Markov chain $X_1, X_2, \cdots$; this, however, will be implicit in the hypotheses of some of our results.

We shall be interested in the asymptotic behavior of the probability of the event

$$A_n = \left\{\hat{P}_n^{(2)} \in \Pi\right\} \tag{10}$$

where $\Pi \subset \Lambda^{(2)}$ is some given set of distributions on $\boldsymbol{X}^2$, and of the conditional joint distribution of the X_i's given A_n. Notice that (4) is a particular case of (10), with

$$\Pi = \left\{P: \sum_{x,y} P(x, y) h_j(x, y) \geq \alpha_j,\ j = 1, \cdots, k\right\}. \tag{11}$$

For any $P \in \Lambda^{(2)}$, we denote by $\overline{P}$ and $\overline{\overline{P}}$ the two marginals of P. Let $P(y|x) = P(x, y)/\overline{P}(x)$, for $\overline{P}(x) > 0$. We designate by $\Lambda_0^{(2)}$ the set of all distributions $P \in \Lambda^{(2)}$ such that $\overline{P} = \overline{\overline{P}}$.

A key role will be played by the Kullback–Leibler information divergence of a distribution $P \in \Lambda^{(2)}$ from that defined by the probabilities $\overline{P}(x)W(y|x)$. For brevity, this divergence will be denoted by $D(P\|W)$; i.e.,

$$D(P\|W) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{\overline{P}(x) W(y|x)} = \sum_{x,y} P(x, y) \log \frac{P(y|x)}{W(y|x)}. \tag{12}$$

Definition 2: If for a subset Π_0 of $\Lambda_0^{(2)}$ there exists a unique $P^* \in \Pi_0$ with $D(P^*\|W) = \min_{P \in \Pi_0} D(P\|W) < \infty$, this P^* is called the Markov I-projection on Π_0 of the transition probability matrix W.

Clearly, $S(P^*) \subset S(W)$. As motivation, we notice that every $P \in \Lambda_0^{(2)}$ determines a stationary Markov chain with two-dimensional distribution P. The m-dimensional distribution of this Markov chain is given by

$$P^m(x_1, \cdots, x_m) = \begin{cases} \overline{P}(x_1) \prod_{i=1}^{m-1} (x_{i+1}|x_i), \\ \qquad \text{if } x_i \in S(\overline{P}),\ i = 1, \cdots, m \\ 0, \qquad \text{else.} \end{cases} \tag{13}$$

If the joint distribution of $X_1, \cdots, X_m$ is denoted by Q^m, then (6), (8), and (13) imply that

$$D(P^m\|Q^m) = \sum_{x_1, \cdots, x_m} P^m(x_1, \cdots, x_m) \log \frac{P^m(x_1, \cdots, x_m)}{Q^m(x_1, \cdots, x_m)} = \sum_{x_1} \overline{P}(x_1) \log \frac{\overline{P}(x_1)}{Q^1(x_1)} + (m-1) \sum_{x,y} P(x, y) \log \frac{P(y|x)}{W(y|x)}.$$

Thus the divergence rate from $X_1, X_2, \cdots$ of the Markov chain defined by (13) is

$$\lim_{m \to \infty} \frac{1}{m} D(P^m\|Q^m) = D(P\|W). \tag{14}$$

It is easy to see that among all stationary processes with the same two-dimensional distributions, Markov chains have the smallest divergence rate from the given Markov chain $X_1, X_2, \cdots$. Hence the minimum divergence rate from $X_1, X_2, \cdots$ of stationary processes with two-dimensional distribution in Π_0 is attained for the Markov chain determined by the Markov I-projection of W on Π_0.

We will say that a subset E of $\boldsymbol{X}^2$ is *irreducible* if the directed graph with vertex set $\boldsymbol{X}$ and edge set E is strongly connected. If, in addition, the greatest common divisor of the lengths of all circuits in this graph is equal to 1, we say that E is *aperiodic*. A distribution $P \in \Lambda^{(2)}$ will be called irreducible (and aperiodic) if $\overline{P} = \overline{\overline{P}}$ and $S(P)$ is an irreducible (and aperiodic) subset of $\boldsymbol{X}^2$. Clearly, this means that $S(\overline{P}) = \boldsymbol{X}$ and the Markov chain defined by (13) is irreducible (and aperiodic).

As $D(P\|W)$ is a continuous function of P restricted to $\{P: S(P) \subset S(W)\}$, the infimum of $D(P\|W)$, subject to $P \in \Pi_0$, equals its minimum, subject to $P \in \operatorname{cl} \Pi_0$, for any $\Pi_0 \subset \Lambda_0^{(2)}$ with $S(\Pi_0) \subset S(W)$. The last minimum may be attained for several $P^* \in \operatorname{cl} \Pi_0$, but the uniqueness (and irreducibility) of the minimizing P^* can often be asserted if Π_0 is convex.

Lemma 1: Let Π_0 be a closed convex subset of $\Lambda_0^{(2)}$ such that $S(\Pi_0) \subset S(W)$. If, in addition, $S(\Pi_0)$ is irreducible, then the Markov I-projection P^* of W on Π_0 exists (i.e., $\min D(P\|W)$ subject to $P \in \Pi_0$ is attained for a unique P^*), $S(P^*) = S(\Pi_0)$, and

$$\sum_{x,y} P(x, y) \log \frac{P^*(y|x)}{W(y|x)} \geq D(P^*\|W), \quad \text{for each } P \in \Pi_0. \tag{15}$$

If $S(\Pi_0)$ is not irreducible, the weaker uniqueness assertion holds that if P_1^* and P_2^* both attain $\min D(P\|W)$, subject to $P \in \Pi_0$, then $P_2^*(\cdot|x) = P_2^*(\cdot|x)$ for all $x \in S(\overline{P}_1^*) \cap S(\overline{P}_2^*)$.

A related result appears in [2, Theorem 5.5]. Still, for the reader's convenience, we will given a complete proof in the Appendix. Notice that (15) is equivalent to

$$D(P\|W) \geq D(P\|P^*(\cdot|\cdot)) + D(P^*\|W), \quad \text{for every } P \in \Pi_0.$$

This inequality is an analog of a well-known property of ordinary I-projections (cf. [3, Theorem 2.2]).

The extension of Theorem 1 to the Markov case is rather straightforward, except for the second assertion in part b). Lemma 2 below covers the easy part; the hard part will be the subject of Theorems 2–4. All these results will be proved in Section III.

Since $\Pr\{\hat{P}_n^{(2)} = P\} = 0$ for every $P \in \Lambda^{(2)}$ with $S(P) \not\subset S(W)$, in the statement of our results we assume, without any loss of generality, that $S(\Pi) \subset S(W)$.

To formulate Lemma 2, let

$$U(P, \epsilon) = \left\{ P' : \max_{x,y} |P'(x,y) - P(x,y)| < \epsilon \right\} \quad (16)$$

denote the ϵ-neighborhood of a $P \in \Lambda^{(2)}$. Further, for any $\Pi \subset \Lambda^{(2)}$, let Π^i be the set of those irreducible $P \in \Pi_0^{(2)}$ to which there exists $\epsilon = \epsilon(P) > 0$ such that every $P' \in U(P, \epsilon)$ with $S(P') = S(P)$ also belongs to Π. This Π^i may be visualized as an "irreducible interior" of Π, even though a $P \in \Pi^i$ need not be in the topological interior of Π (as elements of $U(P, \epsilon)$ with support larger than $S(P)$ are not required to belong to Π); actually, the topological interior of Π is empty whenever $S(\Pi) \neq \boldsymbol{X}^2$.

Lemma 2: Let $\Pi \subset \Lambda^{(2)}$ with $S(\Pi) \subset S(W)$ be arbitrary. Let $\Pi_0 = \Lambda_0^{(2)} \cap \operatorname{cl} \Pi$, and write

$$D = \min_{P \in \Pi_0} D(P \| W). \quad (17)$$

a) $\limsup_{n \to \infty} (1/n) \log \Pr\{\hat{P}_n^{(2)} \in \Pi | X_1 = u\} \leq -D$ for every $u \in \boldsymbol{X}$.

b) A necessary and sufficient condition of

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr\left\{ \hat{P}_n^{(2)} \in \Pi | X_1 = u \right\} = -D \quad (18)$$

is the existence, for every sufficiently large n, of $P_n \in \Pi$ equal to the second-order type of some $\boldsymbol{x} \in \boldsymbol{X}^{n+1}$ with $x_1 = u$, such that $D(P_n \| W) \to D$ as $n \to \infty$. A sufficient condition is that the infimum of $D(P \| W)$ for $P \in \Pi^i$ (defined in the paragraph preceding Lemma 2) be equal to D; this condition is fulfilled, e.g., if $\operatorname{cl} \Pi^i = \Pi_0$.

c) If W has Markov I-projection P^* on Π_0, the following are equivalent, for any given $u \in \boldsymbol{X}$:

1) for every $\epsilon > 0$, $\Pr\{\hat{P}_n^{(2)} \in U(P^*, \epsilon) | \hat{P}_n^{(2)} \in \Pi, X_1 = u\}$ is (defined and) positive if n is sufficiently large;
2) $\hat{P}_n^{(2)}$ converges to P^* in conditional probability given that $\hat{P}_n^{(2)} \in \Pi$ and $X_1 = u$; i.e., for every $\epsilon > 0$,
$$\lim_{n \to \infty} \Pr\left\{ \hat{P}_n^{(2)} \in U(P^*, \epsilon) | \hat{P}_n^{(2)} \in \Pi, X_1 = u \right\} = 1;$$
3) the limit relation (18) holds.

Similar equivalences hold when the conditions $X_1 = u$ are everywhere deleted; for 3), this means replacing (18) by

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr\left\{ \hat{P}_n^{(2)} \in \Pi \right\} = -D. \quad (19)$$

Remark: In Lemma 2 (part c), 2) is a "conditional law of large numbers"; it means that for every function $h(\cdot, \cdot)$ defined on $\boldsymbol{X}^2$,

$$\Pr\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i, X_{i+1}) - E_{P^*} h \right| < \epsilon | \hat{P}_n^{(2)} \in \Pi, X_1 = u \right\} \to 1$$

where $E_{P^*} h$ denotes the expectation of h with respect to the distribution P^*.

Lemma 2 is related to previous results of Boza [2] and Natarajan [12] but their results were not immediately suitable for our purposes. Natarajan [12] proved an analog of (19) using a circular version of second-order types (called Markov types, following Davisson, Longo, and Sgarro [7]). He postulated strict positivity of W and his assumptions rule out those cases when no distribution in Π_0 has support equal to $\boldsymbol{X}^2$. Removing these restrictions is relevant for generalizations to higher order empirical distributions, cf. Section IV. (The device of sliding blocks to reduce order necessarily leads to excluded transitions $W(x|y) = 0$.) While our sufficient condition in Lemma 2 (part b) appears somewhat artificial, it is often easy to use, as in the proof of Theorem 4 below.

Theorem 2: Let Π be any subset of $\Lambda^{(2)}$ with $S(\Pi) \subset S(W)$ such that W has Markov I-projection P^* on $\Pi_0 = \Lambda_0^{(2)} \cap \operatorname{cl} \Pi$. Then for every $m \geq 2$ and $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$ with $x_1 \in S(\bar{P}^*)$ we have, writing

$$P^{*m}(x_2, \cdots, x_m | x_1) = \begin{cases} \prod_{i=1}^{m-1} P^*(x_{i+1} | x_i), & \text{if } x_i \in S(\bar{P}^*),\ i = 1, \cdots, m \\ 0, & \text{otherwise,} \end{cases}$$

1) if (18) holds for $u = x_1$, then

$$\lim_{n \to \infty} \Pr\left\{ X_2 = x_2, \cdots, X_m = x_m | \hat{P}_n^{(2)} \in \Pi, X_1 = x_1 \right\} = P^{*m}(x_2, \cdots, x_m | x_1); \quad (20)$$

2) if (19) holds, then

$$\lim_{n \to \infty} \left(\Pr\left\{ X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n^{(2)} \in \Pi \right\} - \Pr\left\{ X_1 = x_1 | \hat{P}_n^{(2)} \in \Pi \right\} P^{*m}(x_2, \cdots, x_m | x_1) \right) = 0. \quad (21)$$

Remark: The hypothesis of assertion 2) is weaker than that of assertion 1). In fact, while obviously (18) $\Rightarrow$ (19) (for any fixed $u \in \boldsymbol{X}$), the opposite implication holds if and only if

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr\left\{ X_1 = u | \hat{P}_n^{(2)} \in \Pi \right\} = 0.$$

As no assertion could be made for $x_1 \notin S(\bar{P}^*)$, Theorem 2 is valuable mainly in the case when $S(\bar{P}^*) = \boldsymbol{X}$, e.g, when P^* is irreducible. If P^* is also aperiodic then the following theorem holds.

Theorem 3: Let Π and P^* be as in Theorem 2 and suppose, in addition, that P^* is irreducible and aperiodic. Then for every m and $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$, and every sequence of positive integers l_n with $l_n \to \infty$, $n - l_n \to \infty$, we

have

$$\lim_{n\to\infty} \Pr\left\{ X_{l_n+1}=x_1,\cdots, X_{l_n+m}=x_m | \hat{P}_n^{(2)} \in \Pi \right\} = \bar{P}^*(x_1) \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i) \quad (22)$$

providing (19) holds.

We notice that since $S(P^*) \subset S(W)$, the hypothesis of Theorem 3 implicitly includes the irreducibility and aperiodicity of the Markov chain $X_1, X_2, \cdots$. A similar remark applies to Theorem 4 below.

Theorem 4: Let E be a given irreducible subset of $\boldsymbol{X}^2$ such that $E \subset S(W)$. Let $h_1,\cdots, h_k$ be given functions on $\boldsymbol{X}^2$ and $\alpha_1,\cdots,\alpha_k$ be constants, and put

$$A_n = \left\{ \frac{1}{n}\sum_{i=1}^{n} h_j(X_i, X_{i+1}) \ge \alpha_j,\ j=1,\cdots,k; \cdot(X_i, X_{i+1}) \in E, i=1,\cdots,n \right\}. \quad (23)$$

Suppose that there exists some $P \in \Lambda_0^{(2)}$ with

$$\sum_{x,y} P(x,y)h_j(x,y) > \alpha_j, \qquad j=1,\cdots,k;\ S(P) \subset E. \quad (24)$$

Then the Markov I-projection P^* of W on $\Pi_0 = \Lambda_0^{(2)} \cap \Pi$ exists, where

$$\Pi = \left\{ P:\ \sum_{x,y} P(x,y)h_j(x,y) \ge \alpha_j, \quad j=1,\cdots,k;\ S(P) \subset E \right\}, \quad (25)$$

P^* has support equal to E, and (18)–(21) hold with $\{\hat{P}_n^{(2)} \in \Pi\} = A_n$. If, in addition, E is aperiodic, then also (22) holds with $\{\hat{P}_n^{(2)} \in \Pi\} = A_n$, whenever $l_n \to \infty$, $n - l_n \to \infty$.

Remarks: 1) Events of form (23) can always be represented in form (4) as well, simply by introducing a new function $h_{k+1} = 1_E$ (i.e., $h_{k+1}=1$ on E and 0 outside E) and a corresponding constant $\alpha_{k+1}=1$. The representation (23) was chosen to get a simple sufficient condition, viz. (24), for the limit relations (18)–(22).

2) Theorem 4 applies, in particular, also with $k=0$, i.e., for $A_n = \{(X_i, X_{i+1}) \in E,\ i=1,\cdots,n\}$. Then Π is simply the set of all distributions $P \in \Lambda^{(2)}$ with $S(P) \subset E$, and condition (24) becomes vacuous. In this case, P^* has a simple explicit form, cf. (30).

3) While in (23) the k-dimensional vector of the empirical means $(1/n)\sum_{i=1}^n h_j(X_i, X_{i+1})$ is required to be in the set $\{(t_1,\cdots,t_k):\ t_j \ge \alpha_j,\ j=1,\cdots,k\}$, it could as well be required to be in some other convex set F in k-space. For events A_n so defined, Theorem 4 remains valid, by the same proof, if hypothesis (24) is appropriately modified, namely, so that for some $P \in \Lambda^{(2)}$ with $S(P) \subset E$, the vector with components $\sum P(x,y)h_j(x,y)$ is in the interior of F. In this generalization of Theorem 4, P^* is the Markov I-projection of W on $\Pi_0 = \Lambda^{(2)} \cap \mathrm{cl}\,\Pi$ where Π consists of those $P \in \Lambda^{(2)}$ with $S(P) \subset E$ for which the vector with components $\sum P(x,y)h_j(x,y)$ is in F.

If $X_1, X_2, \cdots$ are i.i.d., one might expect (22) to hold even with $l_n = 0$. In view of (21), this would be equivalent to

$$\lim_{n\to\infty} \Pr\left\{ X_1 = x_1 | \hat{P}_n^{(2)} \in \Pi \right\} = \bar{P}^*(x_1). \quad (26)$$

Unfortunately, (26) is false even in very "nice" cases, cf. Example 4 in Section IV. Actually, the (existence and) evaluation of this limit remains an open problem.

Notice that if $X_1, X_2, \cdots$ are i.i.d., (26) would immediately follow from 2) of Lemma 2 (part c) if the conditioning event were defined in terms of circular "Markov types," e.g., if in (4) X_{n+1} were replaced by X_1. It is rather surprising that such an apparently minor change in the condition can substantially affect the conditional distribution.

Finally, we mention that the Markov I-projection P^* in Theorem 4 can be represented as follows. Let $\lambda(\zeta)$ denote the largest eigenvalue of the $|\boldsymbol{X}|\times|\boldsymbol{X}|$ matrix Q_ζ whose (x,y) entry is

$$Q_\zeta(x,y) = \begin{cases} W(y|x)\exp\sum_{j=1}^{k} \zeta_j h_j(x,y), & \text{if } (x,y) \in E \\ 0, & \text{else,} \end{cases} \quad (27)$$

where $\zeta = (\zeta_1,\cdots,\zeta_k)$, $\zeta_j \ge 0$, $j=1,\cdots,k$, and let u_ζ and v_ζ be the corresponding left and right eigenvectors, normalized to have inner product 1. Then

$$\min_{P\in\Pi_0} D(P\|W) = \max_{\zeta} \left(\sum_{j=1}^{k} \zeta_j \alpha_j - \log\lambda(\zeta) \right) \quad (28)$$

where the maximum is taken subject to $\zeta_j \ge 0$, $j=1,\cdots,k$. The Markov I-projection P^* of W on Π_0 is given by

$$P^*(x,y) = \frac{u_\zeta(x) Q_\zeta(x,y) v_\zeta(y)}{\lambda(\zeta)}, \quad (29)$$

for ζ attaining the maximum in (28). In the simple case mentioned in Remark 2) to Theorem 4, (29) reduces to

$$P^*(x,y) = \begin{cases} \lambda^{-1} u(x) W(y|x) v(y), & \text{if } (x,y) \in E \\ 0, & \text{if } (x,y) \notin E \end{cases} \quad (30)$$

where λ is the largest eigenvalue of the matrix obtained from W by replacing the entries $(x,y) \notin E$ by zeros, and u and v are the corresponding left and right eigenvectors.

The proof of (28) and (29) will be omitted. They can be derived from known properties of ordinary I-projections (cf., e.g. [4, Theorem 2]), keeping in mind that P^* equals the I-projection on Π_0 of the (two-dimensional) distribution consisting of the probabilities $\bar{P}^*(x)W(y|x)$. A result of Justesen and Hoholdt [11] is equivalent to the special case $W(y|x) = \text{constant}$ of (29). In this case P^* gives what they call the "maxentropic Markov chain." A result related to theirs was obtained earlier by Spitzer [14].

III. Proofs

First we give a proof of Theorem 1. This proof should be routine for information theorists familiar with the method of types (cf. Csiszár and Körner [6]). Readers less practiced in working with types might find it helpful to get a first overview of our basic approach in this simple case.

Proof of Theorem 1: Let $T_n(P)$ denote the set of those sequences $\boldsymbol{x} \in \boldsymbol{X}^n$ whose (first-order) type equals a given $P \in \Lambda^{(1)}$, and let $\boldsymbol{P}_n = \{P\colon T_n(P) \neq \phi\}$. Then for $P \in \boldsymbol{P}_n$,

$$(n+1)^{|X|}\exp\{-nD(P\|Q)\} \leq Q^n(T_n(P)) \leq \exp\{-nD(P\|Q)\}, \quad (31)$$

(cf. [6, p. 32]). Hence, using the obvious bound $|\boldsymbol{P}_n| \leq (n+1)^{|X|}$, the probability

$$\Pr\{\hat{P}_n \in \Pi\} = \sum_{P \in \Pi \cap \boldsymbol{P}_n} Q^n(T_n(P))$$

can be bounded from above and from below as

$$\Pr\{\hat{P}_n \in \Pi\} \leq (n+1)^{|X|} \max_{P \in \Pi \cap \boldsymbol{P}_n} Q^n(T_n(P)) \leq (n+1)^{|X|}\exp\Big\{-n \min_{P \in \Pi \cap \boldsymbol{P}_n} D(P\|Q)\Big\}, \quad (32)$$

$$\Pr\{\hat{P}_n \in \Pi\} \geq \max_{P \in \Pi \cap \boldsymbol{P}_n} Q^n(T_n(P)) \geq (n+1)^{|X|}\exp\Big\{-n \min_{P \in \Pi \cap \boldsymbol{P}_n} D(P\|Q)\Big\}. \quad (33)$$

The first assertion of part a) immediately follows from these bounds. Since an arbitrarily small neighborhood of any $P \in \Lambda^{(1)}$ contains some $P \in \boldsymbol{P}_n$ if n is sufficiently large, it follows by continuity that for any $\epsilon > 0$,

$$\min_{P \in \Pi \cap \boldsymbol{P}_n} D(P\|Q) \leq \inf_{P \in \operatorname{int}\Pi} D(P\|Q) + \epsilon$$

if n is large enough, where $\operatorname{int}\Pi$ denotes the interior of Π. This and (32), (33) prove that the equality of the last infimum to $\inf_{P \in \Pi} D(P\|Q) = \min_{P \in \operatorname{cl}\Pi} D(P\|Q) = D$ is a sufficient condition for (7). Clearly, this sufficient condition is satisfied if $\operatorname{cl}(\operatorname{int}\Pi) = \operatorname{cl}\Pi$.

Turning now to part b), notice that if D is attained for a unique $P^* \in \operatorname{cl}\Pi$, then the minimum D_ϵ of $D(P\|Q)$ for P ranging over the compact set

$$\Pi^\epsilon\colon \{P\colon P \in \operatorname{cl}\Pi, \max|P(x) - P^*(x)| \geq \epsilon\}$$

is larger than D, for every $\epsilon > 0$. The bound (32) applied to Π^ϵ instead of Π gives that

$$\Pr\{\hat{P}_n \in \Pi^\epsilon\} \leq (n+1)^{|X|}\exp\{-nD_\epsilon\}.$$

Hence if (7) holds, i.e.,

$$\frac{1}{n}\log\Pr\{\hat{P}_n \in \Pi\} \to -D,$$

then

$$\Pr\Big\{\max_{x \in \boldsymbol{X}} |\hat{P}_n(x) - P^*(x)| \geq \epsilon | \hat{P}_n \in \Pi\Big\} \leq \frac{\Pr\{\hat{P}_n \in \Pi^\epsilon\}}{\Pr\{\hat{P}_n \in \Pi\}} \to 0$$

for every $\epsilon > 0$. This means that $\hat{P}_n \to P^*$ in conditional probability given that $\hat{P}_n \in \Pi$, as claimed.

Finally, fix any $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$, denote by $k(x)$ the number of indices $1 \leq i \leq m$ with $x_i = x$, and notice that for any $P \in \boldsymbol{P}_n$ with $nP(x) = f(x)$, say, we have

$$(x_1, \cdots, x_m, x_{m+1}, \cdots, x_n) \in T_n(P), \quad \text{iff } (x_{m+1}, \cdots, x_m) \in T_{n-m}(P'),$$

where $(n-m)P'(x) = f(x) - k(x)$. Since $\Pr\{(X_1, \cdots, X_n) = \boldsymbol{x}\}$ is constant for $\boldsymbol{x} \in T_n(P)$, it follows that

$$\begin{aligned}
&\Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n = P\} \\
&\quad = |T_{n-m}(P')| \Big/ |T_n(P)| \\
&\quad = \frac{(n-m)!}{\prod_{x \in X}(f(x)-k(x))!} \Big/ \frac{n!}{\prod_{x \in X} f(x)!} \\
&\quad = \frac{1}{n(n-1)\cdots(n-m+1)} \prod_{x:\,k(x)>0} \\
&\qquad \cdot [f(x)\cdots(f(x)-k(x)+1)] \\
&\quad = \frac{1}{\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)} \prod_{x:\,k(x)>0} \\
&\qquad \cdot \left[P(x)\cdots\left(P(x)-\frac{k(x)-1}{n}\right)\right].
\end{aligned}$$

This shows that $\Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n = P\}$ converges to

$$\prod_{x \in \boldsymbol{X}} P(x)^{k(x)} = \prod_{i=1}^m P(x_i)$$

as $n \to \infty$, uniformly in $P \in \boldsymbol{P}_n$. This, in turn, implies that to any $\eta > 0$, there exists $\epsilon > 0$ and n_0 such that

$$\left|\Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n = P\} - \prod_{i=1}^m P^*(x_i)\right| < \eta \quad (34)$$

if $P \in \boldsymbol{P}_n$ belongs to the ϵ-neighborhood of P^* and $n \geq n_0$. As

$$\begin{aligned}
&\Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n \in \Pi\} \\
&\quad = \sum_{P \in \Pi \cap \boldsymbol{P}_n} \Pr\{\hat{P}_n = P | \hat{P}_n \in \Pi\} \\
&\qquad \cdot \Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n = P\},
\end{aligned}$$

and here the contribution of the terms with P outside an arbitrarily small neighborhood of P^* is negligible if n is large enough (because (7) was shown to imply that $\hat{P}_n \to P^*$

in conditional probability), (34) gives rise to

$$\lim_{n\to\infty} \Pr\{X_1 = x_1, \cdots, X_m = x_m | \hat{P}_n \in \Pi\} = \prod_{i=1}^{m} P^*(x_i).$$

This completes the proof of Theorem 1.

The proofs of Lemma 2 and Theorems 2–4 will be basically similar to that of Theorem 1, relying upon properties of second-order types which we now summarize.

For $P \in \Lambda^{(2)}$ and $u \in \boldsymbol{X}$, let $T_n(P, u)$ designate the set of those sequences $\boldsymbol{x} \in \boldsymbol{X}^{n+1}$ with $x_1 = u$ whose second-order type equals P. Write $\boldsymbol{P}_n(u) = \{P\colon P \in \Lambda^{(2)}, T_n(P, u) \neq \phi\}$.

Obvious necessary conditions for $P \in \boldsymbol{P}_n(u)$ are that the numbers $f(x, y) = nP(x, y)$ be nonnegative integers satisfying

$$\sum_{x,y} f(x, y) = n, \tag{35}$$

and for some $v \in \boldsymbol{X}$,

$$\sum_y f(x, y) - \delta(u, x) = \sum_y f(y, x) - \delta(v, x) \geq 0, \qquad x \in \boldsymbol{X}. \tag{36}$$

Here $\delta(x, y) = 1$ if $x = y$, and 0 otherwise. Clearly, v is uniquely determined by P and u; it is the last element of any $\boldsymbol{x} \in T_n(P, u)$.

Notice that (36) implies for $P \in \boldsymbol{P}_n(u)$ that $\overline{P}(x) \neq \overline{\overline{P}}(x)$ happens only if $u \neq v$ and x equals u or v, in which case $\overline{P}(u) - \overline{\overline{P}}(u) = \overline{\overline{P}}(v) - \overline{P}(v) = 1/n$.

The following proposition counts sequences of a given second-order type and is due to Whittle [17]; for a simple proof see [1].

Proposition W: If the numbers $f(x, y) = nP(x, y)$ are nonnegative integers satisfying (35) and (36), then

$$|T_n(P, u)| = F_{vu}^*(P) \prod_{x \in \boldsymbol{X}} \frac{f(x)!}{\prod_{y \in \boldsymbol{X}} f(x, y)!} \tag{37}$$

where $f(x) = \sum_y f(x, y)$ and $F_{vu}^*(P)$ is the (v, u)-cofactor of the $|\boldsymbol{X}| \times |\boldsymbol{X}|$ matrix $F^*(P)$ whose entries are

$$F^*(x, y) = \begin{cases} \delta(x, y) - P(x, y)/\overline{P}(x), & \text{if } x \in S(\overline{P}) \\ \delta(x, y), & \text{else.} \end{cases} \tag{38}$$

The conditions (35) and (36) are necessary but not sufficient for $P \in \boldsymbol{P}_n(u)$, because $F_{vu}^*(P)$ in (37) may be zero. Necessary and sufficient is that in addition to (35) and (36), for a suitable ordering $x_0, \cdots, x_l$ of the elements of $S(\overline{P}) \cup \{v\}$ with $x_0 = u$, $x_l = v$, we have $(x_{i-1}, x_i) \in S(P)$, $i = 1, \cdots, l$. This follows from the proof of Proposition W in [1] but will not be used in this paper.

The following consequence of Proposition W is an analog of (31); it suffices to prove Lemma 2.

Lemma 3: For $P \in \boldsymbol{P}_n(u)$, we have

$$\begin{aligned} &(n+1)^{-|\boldsymbol{X}|^2 - |\boldsymbol{X}|} \exp\{-nD(P\|W)\} \\ &\quad \leq \Pr\{\hat{P}_n^{(2)} = P | X_1 = u\} \\ &\quad \leq \exp\{-nD(P\|W)\}. \end{aligned} \tag{39}$$

This is effectively Theorem 3.1(a) of Boza [2]. Clearly, (39) trivially holds if $S(P) \not\subset S(W)$ (with $\exp(-\infty) = 0$). Otherwise, it is an immediate consequence of (8), (37), and the standard rough bounds on multinomial coefficients (cf. [6, p. 30]) since the factor $F_{vu}^*(P)$ in (37) can be neither less than $(n+1)^{-|\boldsymbol{X}|}$ nor greater than 1.

Proof of Lemma 2: As each $P \in \Pi \cap \boldsymbol{P}_n(u)$ is also in

$$\Pi_n = \Pi \cap \left\{P\colon \max_x \left|\overline{P}(x) - \overline{\overline{P}}(x)\right| \leq \frac{1}{n}\right\},$$

we have

$$\Pr\{\hat{P}_n^{(2)} \in \Pi | X_1 = u\} = \sum_{P \in \Pi_n \cap \boldsymbol{P}_n(u)} \Pr\{\hat{P}_n^{(2)} = P | X_1 = u\}. \tag{40}$$

If $\Pi_m = \phi$ for some m, then $\Pr\{\hat{P}_n^{(2)} \in \Pi | X_1 = u\} = 0$ for $n \geq m$, and Lemma 2 assertion a) holds trivially. Otherwise

$$D_n = \inf_{P \in \Pi_n} D(P\|W) < \infty, \qquad n = 1, 2, \cdots, \tag{41}$$

and from (40) and (39) we obtain

$$\Pr\{\hat{P}_n^{(2)} \in \Pi | X_1 = u\} \leq |\boldsymbol{P}_n(u)| \exp(-nD_n).$$

Since $|\boldsymbol{P}_n(u)|$ grows only polynomially with n, this will prove assertion a) if we show that

$$\lim_{n\to\infty} D_n = D = \min_{P \in \Pi_0} D(P\|W). \tag{42}$$

By continuity, (41) implies that D_n equals the minimum of $D(P\|W)$ subject to $P \in \operatorname{cl} \Pi_n$; let $P_n \in \operatorname{cl} \Pi_n$ attain this minimum. Picking a convergent subsequence, say $P_{n_k} \to P_0$, we have $P_0 \in \Pi_0 = \Lambda_0^{(2)} \cap \operatorname{cl} \Pi$, thus

$$\lim_{k\to\infty} D_{n_k} = \lim_{k\to\infty} D(P_{n_k}\|W) = D(P_0\|W) \geq D.$$

As the sequence D_n is nondecreasing and cannot exceed D (because $\operatorname{cl} \Pi_m \supset \operatorname{cl} \Pi_n \supset \Pi_0$ for $m < n$), this proves (42) and thereby assertion a).

The first part of b) immediately follows from a) and Lemma 3. Next, notice that to any irreducible $P \in \Lambda_0^{(2)}$, $\epsilon > 0$, and $u \in \boldsymbol{X}$ there exists

$$P' \in \boldsymbol{P}_n(u) \cap U(P, \epsilon), \qquad \text{with } S(P') = S(P) \tag{43}$$

for every sufficiently large n. This follows, e.g., from the law of large numbers applied to the irreducible Markov chain determined by P, cf. (13).

Given any $P \in \Pi^i$ and $\delta > 0$, pick $\epsilon > 0$ so small that (43) implies both $P' \in \Pi$ (possible by the definition of Π^i) and $D(P'\|W) < D(P\|W) + \delta$ (possible by continuity). Then by Lemma 3,

$$\begin{aligned} &\Pr\{\hat{P}_n^{(2)} \in \Pi | X_1 = u\} \\ &\quad \geq \Pr\{\hat{P}_n^{(2)} = P' | X_1 = u\} \\ &\quad \geq (n+1)^{-|\boldsymbol{X}|^2 - |\boldsymbol{X}|} \exp\{-nD(P'\|W)\} \\ &\quad \geq \exp\{-n(D(P\|W) + 2\delta)\} \end{aligned}$$

if n is large enough. This proves that

$$\inf_{P \in \Pi^i} D(P\|W) = D$$

is a sufficient condition for (18); this condition is clearly satisfied if $\operatorname{cl} \Pi^i = \Pi_0$.

Turning to part c), notice that 1) means that for every sufficiently large n, there exists $P_n \in \boldsymbol{P}_n(u) \cap \Pi$ such that $P_n \to P^*$ as $n \to \infty$. Thus the necessary and sufficient condition of part b) is satisfied, and 1) $\Rightarrow$ 3).

Further, if P^* is the Markov I-projection of W on Π_0, i.e., P^* is the unique $P \in \Pi_0$ attaining the minimum in (17), then

$$\min_{P \in \Pi_0 - U(P^*, \epsilon)} D(P\|W) > D, \qquad \text{for every } \epsilon > 0,$$

and thus, by assertion a),

$$\limsup_{n \to \infty} \frac{1}{n} \log \Pr\left\{ \hat{P}_n^{(n)} \in \Pi - U(P^*, \epsilon) | X_1 = u \right\} < -D. \tag{44}$$

Hence the implication 3) $\Rightarrow$ 2) directly follows, since (18) and (44) result in

$$\Pr\left\{ \hat{P}_n^{(2)} \notin U(P^*, \epsilon) | \hat{P}_n^{(2)} \in \Pi, X_1 = u \right\} = \frac{\Pr\left\{ \hat{P}_n^{(2)} \in \Pi - U(P^*, \epsilon) | X_1 = u \right\}}{\Pr\left\{ \hat{P}_n^{(2)} \in \Pi | X_1 = u \right\}} \to 0.$$

The remaining implication 2) $\Rightarrow$ 1) is trivial. The mutual equivalence of the analogs of 1), 2), and 3) obtained by deleting the conditions $X_1 = u$ can be proved similarly. Thus the proof of Lemma 2 is complete.

While the bounds (39) were sufficient for Lemma 2, we will need the exact formula (37) to prove Theorems 2 and 3. Also, two further lemmas will be needed (Lemma 5 for Theorem 3 only).

Lemma 4: The factor $F_{vu}^*(P)$ in Proposition W equals $F_{vv}^*(P)$, which can be expressed as the sum of certain products of conditional probabilities $P(y|x)$. More exactly,

$$F_{vu}^*(P) = F_{vv}^*(P) = \sum_{\phi \in \Phi} \prod_{x \in S - \{v\}} P(\phi(x)|x), \tag{45}$$

where $S = S(\bar{P}) \cup \{v\}$ and Φ is a set of mappings ϕ: $(S - \{v\}) \to S$ uniquely determined by S and v.

Proof: Without any loss of generality, we assume that $X = \{0, 1, \cdots, N\}$, $v = 0$, and that $S(\bar{P})$ is either $\{0, 1, \cdots, s\}$ or $\{1, \cdots, s\}$ for some s with $u \le s \le N$. By the paragraph after (36), we have for each $x \neq v = 0$ either $\bar{\bar{P}}(x) = \bar{P}(x)$ or $\bar{\bar{P}}(x) = \bar{P}(x) - 1/n$, hence $S(\bar{\bar{P}}) \subset S = \{0, 1, \cdots, s\}$. Thus if $s \neq N$, the matrix $F^* = F^*(P)$ defined by (38) can be decomposed as

$$F^* = \left(\begin{array}{c|c} G & 0 \\ \hline 0 & I \end{array} \right)$$

where G is an $(s+1) \times (s+1)$ matrix and I is the $(N-s) \times (N-s)$ unit matrix. It follows that the $(0, u)$-cofactor of F^* is equal to that of G. As the $s \times (s+1)$ submatrix of G obtained by deleting the 0th row has entries

$$G(x, y) = F^*(x, y) = \delta(x, y) - P(y|x),$$

its row sums are 0. Hence the $(0, j)$-cofactors of G are the same for all $0 \le j \le s$; in particular,

$$F_{0u}^* = G_{0u} = G_{00} = F_{00}^*.$$

Now denote by A the $s \times s$ matrix obtained by deleting the 0th row and column of G. Then $G_{00} = \det A$ and

$$A = \begin{pmatrix} a_{10} + a_{12} + a_{1s} & -a_{12} & \cdots & -a_{1s} \\ -a_{21} & a_{20} + a_{21} + a_{23} + \cdots + a_{2s} & \cdots & -a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{s1} & -a_{s2} & \cdots & a_{s0} + a_{s1} + \cdots + a_{s,s-1} \end{pmatrix} \tag{46}$$

with $a_{xy} = P(y|x)$, $x = 0, 1, \cdots, s$, $y = 1, \cdots, s$, $y \neq x$.

Lemma 4 will be proved if we show that for every $s \ge 1$, there exists a set Φ_s of mappings ϕ: $\{1, \cdots, s\} \to \{0, \cdots, s\}$ with $\phi(i) \neq i$, $i = 1, \cdots, s$, such that for every $s \times s$ matrix of form (46),

$$\det A = \sum_{\phi \in \Phi_s} \prod_{i=1}^{s} a_{i\phi(i)}. \tag{47}$$

For any A of form (46), $\det A$ is a polynomial of the a_{ij}'s consisting of terms $a_{1j_1} \cdots a_{sj_s}$ (with $j_i \neq i$, $i = 1, \cdots, s$), possibly with coefficients depending on $(j_1, \cdots, j_s)$. To establish (47), we prove by induction that all the nonzero coefficients in this polynomial are equal to 1. This is obvious for $s = 1$ when there is a single term a_{10}.

For $s \ge 2$ we use the easily checked identity that for any $s \times s$ matrix B and diagonal matrix $C = \operatorname{diag}(c_1, \cdots, c_s)$,

$$\det(B + C) = \sum_I \left(\prod_{i \in I} c_i \right) \det B(I).$$

Here the sum extends for all subsets I of $\{1, \cdots, s\}$ and $B(I)$ denotes the matrix obtained by deleting the rows and columns of indices $i \in I$ of B (for $I = \{1, \cdots, s\}$ we understand $\det B(I) = 1$). We apply this identity to $C = A_0 = \operatorname{diag}(a_{10}, \cdots, a_{s0})$ and $B = A - A_0$. Then the determinant of $B(\phi) = B = A - A_0$ is 0 because the row sums of $A - A_0$ are all 0. Thus we get that

$$\det A = \sum_I \left(\prod_{i \in I} a_{i0} \right) \det(A - A_0)(I)$$

where the sum extends to the nonvoid subsets I of $\{1, \cdots, s\}$. Here all the matrices $(A - A_0)(I)$ are of form (46), with $\sum_{j \in I^c} a_{ij}$ playing the role of a_{i0} (where $I^c = \{1, \cdots, s\} - I$). As the dimension of the matrices $(A - A_0)(I)$ is less than s, it follows that if the induction hypothesis is true for $1, \cdots, s-1$, then it is true also for s.

This completes the proof of (47) and thereby of Lemma 4.

Lemma 5: For every $0 < i < n$ and $P \in \bigcup_{u \in X} \boldsymbol{P}_n(u)$ with $S(P) \subset S(W)$, we have

$$\Pr\left\{ \hat{P}_i^{(2)} \in U(P, \epsilon) | \hat{P}_n^{(2)} = P \right\} > 1 - (n+1)^{2|X|^2} \exp\left\{ -i \min_{P' \notin U(P, \epsilon)} D(P'\|P(\cdot|\cdot)) \right\} \tag{48}$$

and

$$\Pr\left\{\hat{P}_{n,i}^{(2)} \in U(P,\epsilon) | \hat{P}_n^{(2)} = P\right\} > 1-(n+1)^{2|X|^2} \exp\left\{-(n-i) \min_{P' \notin U(P,\epsilon)} D(P'\|P(\cdot|\cdot))\right\} \tag{49}$$

where $\hat{P}_i^{(2)}$ and $\hat{P}_{n,i}^{(2)}$ are the second-order types of $(X_1,\cdots,X_{i+1})$ and $(X_{i+1},\cdots,X_{n+1})$, respectively, and the minima are understood subject to the additional constraint that P' is a possible value of $\hat{P}_i^{(2)}$ resp. $\hat{P}_{n,i}^{(2)}$ when $\hat{P}_n^{(2)} = P$.

Remark: The point of Lemma 5 is that if i and $n-i$ are sufficiently large, the conditional probabilities in (48) and (49) are arbitrarily close to 1.

Proof: Fix P as above and consider any pair (P_1, P_2) such that for some $x = (x_1,\cdots,x_{n+1}) \in X^{n+1}$ of second-order type P, $(x_1,\cdots,x_{i+1})$ has second-order type P_1 and $(x_{i+1},\cdots,x_{n+1})$ has second-order type P_2; of course, then $iP_1 + (n-i)P_2 = nP$. Let $u \in X$ be such that $P \in P_n$. Then

$$\begin{aligned}\Pr\left\{\hat{P}_i^{(2)} = P_1 | \hat{P}_n^{(2)} = P, X_1 = u\right\} &= \Pr\left\{\hat{P}_{n,i}^{(2)} = P_2 | \hat{P}_n^{(2)} = P, X_1 = u\right\} \\ &= \Pr\left\{\hat{P}_i^{(2)} = P_1, \hat{P}_{n,i}^{(2)} = P_2 | X_1 = u\right\} \\ &\quad / \Pr\left\{\hat{P}_n^{(2)} = P | X_1 = u\right\}. \end{aligned} \tag{50}$$

Here, by the Markov property and Lemma 3, we have

$$\begin{aligned}\Pr\left\{\hat{P}_i^{(2)} = P_1, \hat{P}_{n,i}^{(2)} = P_2 | X_1 = u\right\} &\le \Pr\left\{\hat{P}_i^{(2)} = P_1 | X_1 = u\right\} \max_{x \in X} \Pr\left\{\hat{P}_{n,i}^{(2)} = P_2 | X_{i+1} = x\right\} \\ &\le \exp\left\{-iD(P_1\|W) - (n-i)D(P_2\|W)\right\}.\end{aligned}$$

Now we use an identity that can be easily verified from (12), namely, that if $P = \alpha P_1 + (1-\alpha)P_2$ with some $0 < \alpha < 1$, then

$$\begin{aligned}\alpha D(P_1\|W) + (1-\alpha)D(P_2\|W) &= D(P\|W) + \alpha D(P_1\|P(\cdot|\cdot)) \\ &\quad + (1-\alpha)D(P_2\|P(\cdot|\cdot)).\end{aligned} \tag{51}$$

Rewriting our last upper bound on the numerator in (50) according to (51) (with $\alpha = i/n$) and using the lower bound on the denominator provided by Lemma 3, it follows that the two equal conditional probabilities in (50) are upperbounded by

$$(n+1)^{|X|^2+|X|} \exp\left\{-iD(P_1\|P(\cdot|\cdot)) - (n-i)D(P_2\|P(\cdot|\cdot))\right\}.$$

Since $|P_n(u)| < (n+1)^{|X|^2-|X|}$, this implies, in particular, that

$$\Pr\left\{\hat{P}_i^{(2)} \notin U(P,\epsilon) | \hat{P}_n^{(2)} = P, X_1 = u\right\} < (n+1)^{2|X|^2} \exp\left\{-\min_{P' \notin U(P,\epsilon)} D(P'\|P(\cdot|\cdot))\right\}$$

and

$$\Pr\left\{\hat{P}_{n,i}^{(2)} \notin U(P,\epsilon) | \hat{P}_n^{(2)} = P, X_1 = u\right\} < (n+1)^{2|X|^2} \exp\left\{-(n-i) \min_{P' \notin U(P,\epsilon)} D(P'\|P(\cdot|\cdot))\right\}.$$

As these bounds hold for every $u \in X$ with $\Pr\{X_1 = u | \hat{P}_n^{(2)} = P\} > 0$, they remain valid also if the condition $X_1 = u$ is deleted. This proves (48) and (49).

Proof of Theorem 2: First we prove that to any $P^* \in \Lambda_0^{(2)}$ with $S(P^*) \subset S(W)$ and to every m and $\eta > 0$, there exist $\epsilon > 0$ and n_0 such that if $n \ge n_0$ then for every $(x_1,\cdots,x_m) \in X^m$ with $x_1 \in S(\bar{P}^*)$ and $P \in U(P^*,\epsilon) \cap P_n(x_1)$ with $S(P) \subset S(W)$,

$$\left|\Pr\left\{X_2 = x_2,\cdots,X_m = x_m | \hat{P}_n^{(2)} = P, X_1 = x_1\right\} - P^{*m}(x_2,\cdots,x_m|x_1)\right| < \eta. \tag{52}$$

This will imply assertion 1) of Theorem 2 exactly as (34) implied the last assertion of Theorem 1.

Fix an m-tuple $(x_1,\cdots,x_m) \in X^m$ and write

$$k(x,y) = \left|\{i: (x_i, x_{i+1}) = (x,y), 1 \le i \le m-1\}\right|. \tag{53}$$

Then a sequence $(x_1,\cdots,x_{n+1}) \in X^{n+1}$, whose initial m-tuple equals the given one, has the second-order type P with $nP(x,y) = f(x,y)$ if and only if the second-order type P' of $(x_m,\cdots,x_{n+1})$ is given by

$$\begin{aligned}P'(x,y) &= \frac{f(x,y) - k(x,y)}{n-m+1} \\ &= \frac{n}{n-m+1}P(x,y) - \frac{k(x,y)}{n-m+1}.\end{aligned} \tag{54}$$

Since $\Pr\{X_1 = x_1,\cdots,X_{n+1} = x_{n+1}\}$ is constant for $(x_1,\cdots,x_{n+1}) \in T_n(P,x_1)$, Proposition W yields

$$\begin{aligned}&\Pr\left\{X_2 = x_2,\cdots,X_m = x_m | \hat{P}_n^{(2)} = P, X_1 = x_1\right\} \\ &= \frac{|T_{n-m+1}(P', x_m)|}{|T_n(P,x_1)|} \\ &= \frac{F^*_{vx_m}(P')}{F^*_{vx_1}(P)} \prod_{x:k(x)>0} \frac{\prod_{y:k(x,y)>0}[f(x,y)\cdots(f(x,y)-k(x,y)+1)]}{f(x)\cdots(f(x)-k(x)+1)} \\ &= \frac{F^*_{vx_m}(P')}{F^*_{vx_1}(P)} \prod_{x:k(x)>0} \frac{\prod_{y:k(x,y)>0}\left[P(x,y)\cdots\left(P(x,y)-\frac{k(x,y)-1}{n}\right)\right]}{\bar{P}(x)\cdots\left(\bar{P}(x)-\frac{k(x)-1}{n}\right)},\end{aligned} \tag{55}$$

where $f(x) = \sum_y f(x,y)$, $k(x) = \sum_y k(x,y)$, and v is the common last element x_{n+1} of the sequences in $T_n(P,x_1)$.

We claim that for $n \to \infty$ the last expression in (55) converges to

$$\prod_{x: k(x)>0} \left(\prod_{y: k(x,y)>0} P(x,y)^{k(x,y)} \overline{P}^{-k(x)} \right) = \prod_{i=1}^{n} P(x_{i+1}|x_i) \quad (56)$$

uniformly for $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$ and $P \in \boldsymbol{P}_n(x_1)$ such that

$$P(x_i, x_{i+1}) \geq \delta, \qquad i = 1, \cdots, m, \quad (57)$$

where δ is an arbitrary but fixed positive number.

As (57) means that $P(x, y) \geq \delta$ if $k(x, y) > 0$, this claim will be established if we show that $F^*_{vx_m}(P')/F^*_{vx_1}(P) \to 1$ uniformly, subject to (57). This is nontrivial, because even though the numerator and denominator will be arbitrarily close to each other if n is large, both may be arbitrarily close to 0. Actually, this is the point where we need Lemma 4.

Now, as $P(x, y) \geq \delta$ if $k(x, y) > 0$, (54) gives

$$\frac{n}{n-m+1} - \frac{m}{(n-m+1)\delta} \leq \frac{P'(x,y)}{P(x,y)} \leq \frac{n}{n-m+1}, \qquad \text{if } P(x,y) > 0.$$

Thus $S(P') = S(P)$ if $n > m\delta^{-1}$; further, $P'(x, y)/P(x, y)$ and hence also $\overline{P}'(x)/\overline{P}(x)$ and $P'(y|x)/P(y|x)$ converge uniformly to 1 if $(x, y) \in S(P)$. Hence for $n > m\delta^{-1}$, the same Φ appears in the expansions of $F^*_{vx_m}(P')$ and $F^*_{vx_1}(P)$ by Lemma 4, and the ratios of those corresponding terms which do not both vanish converge to 1 uniformly as $n \to \infty$. Since all terms are nonnegative, this implies the desired uniform convergence $F^*_{vx_m}(P')/F^*_{vx_1}(P) \to 1$.

If $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$ satisfies

$$(x_i, x_{i+1}) \in S(P^*), \qquad i = 1, \cdots, m-1, \quad (58)$$

then (57) holds for all $P \in U(P^*, \epsilon)$ if ϵ is sufficiently small (with any δ such that $\delta + \epsilon$ is less than the smallest positive $P^*(x, y)$). Thus by the result just proved, $\Pr\{X_2 = x_2, \cdots, X_m = x_m | \hat{P}_n = P, X_1 = x_1\}$ will be arbitrarily close to (56), and hence also to $P^{*m}(x_2, \cdots, x_m|x_1)$, uniformly for all $P \in P_n(x_1) \cap U(P^*, \epsilon)$, if ϵ is sufficiently small and n is sufficiently large.

This already proves (52) for $(x_2, \cdots, x_m) \in \boldsymbol{X}^{m-1}$ with the property (58). As $\Sigma P^{*m}(x_2, \cdots, x_m|x_1)$ for all these $(x_2, \cdots, x_m)$ (with $x_1 \in S(\overline{P}^*)$ fixed) is 1, this in turn implies that for all other $(x_2, \cdots, x_m) \in \boldsymbol{X}^{m-1}$, the conditional probabilities $\Pr\{X_2 = x_2, \cdots, X_m = x_m | \hat{P}^{(2)}_n = P, X_1 = x_1\}$ must be close to 0 uniformly for $P \in \boldsymbol{P}_n(x_1) \cap U(P^*, \epsilon)$. This completes the proof of (52).

Now suppose that P^* is the Markov I-projection of W on Π_0 and (18) holds for $u = x_1$. Then by the equivalence 2) $\Leftrightarrow$ 3) in Lemma 2 part c), we have

$$\Pr\{\hat{P}^{(2)}_n \in U(P^*, \epsilon) | \hat{P}^{(2)}_n \in \Pi, X_1 = x_1\} > 1 - \eta \quad (59)$$

whenever $n \geq n_1$, say. Since (52) holds for every $P \in U(P^*, \epsilon) \cap \boldsymbol{P}_n(x_1)$ and $n \geq n_0$, it follows that

$$\left|\Pr\{X_2 = x_2, \cdots, X_m = x_m | \hat{P}^{(2)}_n \in \Pi, X_1 = x_1\} - P^{*m}(x_2, \cdots, x_m|x_1)\right| < 2\eta \quad (60)$$

if $n \geq \max(n_0, n_1)$. This proves (20).

If instead of (18), only (19) is postulated, we claim that (60) still holds at least for those sufficiently large n that satisfy

$$\Pr\{X_1 = x_1 | \hat{P}^{(2)}_n \in \Pi\} \geq \eta. \quad (61)$$

In fact, Lemma 2 c) (with the condition $X_1 = u$ deleted) guarantees that

$$\Pr\{\hat{P}^{(2)}_n \in U(P^*, \epsilon) | \hat{P}^{(2)}_n \in \Pi\} > 1 - \eta^2$$

if $n \geq n_2$, say. As this inequality implies (59) if (61) holds, we get, as claimed, that (60) holds for $n \geq \max(n_0, n_2)$ satisfying (61). But then the left side of (60) multiplied by $\Pr\{X_1 = x_1 | \hat{P}^{(2)}_n \in \Pi\}$ will be less than 2η for every $n \geq \max(n_0, n_2)$. This proves (21).

Remark: After having submitted this paper, we learned from Persi Diaconis that Zaman [18] (cf. also Zaman [19]) had obtained results similar to (52) in a different context. The goals and method of his work were quite different from ours, and we could not easily determine whether his results could also have been used to prove Theorem 2.

Proof of Theorem 3: Since P^* is irreducible, $S(\overline{P}^*) = \boldsymbol{X}$, thus

$$P^{*m}(x_2, \cdots, x_m|x_1) = \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i)$$

for all $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$. As P^* is also aperiodic, to any $\eta > 0$ there exists a k such that the k-step transition probabilities of the Markov chain determined by P^* differ by less than η from the stationary probabilities; that is,

$$\left| \sum_{(u_2, \cdots, u_k) \in \boldsymbol{X}^{k-1}} P^{*(k+1)}(u_2, \cdots, u_k, x|u) - \overline{P}^*(x) \right| < \eta$$

for every u and x in $\boldsymbol{X}$. Fixing such a k, apply (52) to $k + m$ instead of m and $|\boldsymbol{X}|^{-k+1}\eta$ instead of η. It follows that for any $(u_1, \cdots, u_k, x_1, \cdots, x_m) \in \boldsymbol{X}^{k+m}$,

$$\Pr\{X_2 = u_2, \cdots, X_k = u_k, X_{k+1} = x_1, \cdots, X_{k+m} = x_m | \hat{P}^{(2)}_n = P, X_1 = u_1\}$$

differs by less than $|\boldsymbol{X}|^{-k+1}\eta$ from

$$P^{*(k+m)}(u_2, \cdots, u_k, x_1, \cdots, x_m|u_1) = P^{*(k+1)}(u_2, \cdots, u_k, x_1|u_1) P^{*m}(x_2, \cdots, x_m|x_1)$$

if $P \in \boldsymbol{P}_n(u_1) \cap U(P^*, \epsilon)$ and $n \geq n_0$ (with suitable ϵ and n_0). Summing for all $(u_2, \cdots, u_k) \in \boldsymbol{X}^{k-1}$, we obtain that

$$\left|\Pr\{X_{k+1} = x_1, \cdots, X_{k+m} = x_m | \hat{P}^{(2)}_n = P, X_1 = u_1\} - \overline{P}^*(x_1) P^{*m}(x_2, \cdots, x_m|x_1)\right| < 2\eta.$$

Of course, this result is not affected when shifting the starting point of time, say by $i = l - k$; i.e., we also have

$$\left| \Pr\left\{ X_{l+1} = x_1, \cdots, X_{l+m} = x_m | \hat{P}^{(2)}_{n,i} = P,\ X_{i+1} = u \right\} - \bar{P}^*(x_1) \prod_{i=1}^{n-1} P^*(x_{i+1}|x_i) \right| < 2\eta \quad (62)$$

whenever $P \in U(P^*, \epsilon) \cap \boldsymbol{P}_{n-i}(u)$ and $n - i \geq n_0$, $i = l - k$. Here $l \geq k$ is arbitrary, it may depend on n, while k (depending on η) is fixed as above.

Now, the hypothesis (19) implies by Lemma 2 c) that

$$\lim_{n\to\infty} \Pr\left\{ \hat{P}^{(2)}_n \in U(P^*, \epsilon') | \hat{P}^{(2)}_n \in \Pi \right\} = 1, \quad \text{for every } \epsilon' > 0. \quad (63)$$

Further, there exist $\epsilon' > 0$ and $\delta > 0$ such that in (49) of Lemma 5,

$$\min_{P' \notin U(P,\epsilon)} D(P' \| P(\cdot|\cdot)) > \delta, \qquad \text{for all } P \in U(P^*, \epsilon'), \quad (64)$$

if $n - i$ is sufficiently large. In fact, otherwise for certain $P_{n_k} \to P^*$ and $P'_{n_k} \notin U(P^*, \varepsilon/2)$ we would have $D(P'_{n_k} \| P_{n_k}(\cdot|\cdot)) \to 0$, where P'_{n_k} should be a possible value of $\hat{P}^{(2)}_{n_k, i_k}$ with $n_k - i_k \to \infty$. Picking a convergent subsequence of P'_{n_k}, the last condition implies that its limit P^{**} must be in $\Lambda^{(2)}_0$, while by the previous ones $P^{**} \neq P^*$ and $D(P^{**} \| P^*(\cdot|\cdot)) = 0$. This contradicts the irreducibility of P^*.

On account of (63), (64), and Lemma 5, for any sequence of integers i_n with $1 \leq i_n \leq \gamma n$ (for some fixed $\gamma < 1$),

$$\lim_{n\to\infty} \Pr\left\{ \hat{P}^{(2)}_{n,i_n} \in U(P^*, \epsilon) | \hat{P}^{(2)}_n \in \Pi \right\} = 1, \quad \text{for every } \epsilon > 0. \quad (65)$$

Since $\Pr\{ \hat{P}^{(2)}_{n,i} \in U(P^*, \epsilon) | \hat{P}^{(2)}_n \in \Pi > 1 - \eta$ and (62) imply, using the Markov property, that

$$\left| \Pr\left\{ X_{l+1} = x_1, \cdots, X_{l+m} = x_m | \hat{P} \in \Pi \right\} - \bar{P}^*(x_1) \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i) \right| < 3\eta,$$

this proves (22) for the case when $l_n \to \infty$, $l_n < \gamma n$.

If instead of $l_n < \gamma n$ only $n - l_n \to \infty$ is assumed, our proof of (65) breaks down when $n - i_n$ goes to infinity too slowly. In the case $l_n > \gamma n$, however, a similar argument can be used looking at the sample "backwards." More specifically, we then set $i = l + m + k$ (rather than $i = l - k$), we use instead of (65) the fact

$$\Pr\left\{ \hat{P}^{(2)}_i \in U(P^*, \epsilon) | \hat{P}^{(2)}_n \in \Pi \right\} \to 1, \qquad \text{for every } \epsilon > 0,$$

(also a consequence of Lemma 5), and we use instead of (52) its analog for the terminal m-tuple of the sample (giving the role of n and m to i and $m + k$).

Proof of Theorem 4: Since E is an irreducible subset of $\boldsymbol{X}^2$, there exists $P_1 \in \Lambda^{(2)}_0$ with $S(P_1) = E$. By assumption, some $P_0 \in \Lambda^{(2)}_0$ satisfies (24) and consequently so does $P_\beta = (1 - \beta)P_0 + \beta P_1$ if $\beta > 0$ is sufficiently small. Hence the set of those $P \in \Lambda^{(2)}_0$ which satisfy (24) with $S(P) = E$ is nonvoid; denote it by Π'_0. Clearly, Π'_0 is a subset of the "irreducible interior" Π' appearing in Lemma 2 b). As every $P \in \Pi_0$ belongs to the closure of Π'_0 (take $P = \lim_{\beta\to 1}[(1 - \beta)P_0 + \beta P]$ with any $P_0 \in \Pi'_0$), Lemma 2 b) applies and gives (18) and (19). Further, $\Pi_0 = \Lambda^{(2)}_0 \cap \Pi$ satisfies the hypotheses of Lemma 1, thus the Markov I-projection P^* of W on Π_0 exists and $S(P^*) = E$. Now the remaining assertions of Theorem 4 follow from Theorems 2 and 3.

IV. Comments and Counterexamples

The large deviation result (19) cannot hold for arbitrary $\Pi \subset \Lambda^{(2)}$. It may well happen, e.g., that Π does not contain any $P \in \boldsymbol{P}_n(u)$, even though $\min_{P \in \Pi_0} D(P\|W)$ is finite. This is also possible when Π is required to be convex. A necessary and sufficient condition for (19) appears in Lemma 2 b); that condition, however, may not be easy to verify. One merit of the sufficient condition given in Lemma 2 b) is that it easily applied to the important situation of Theorem 4.

The first example shows that for the convergence of $\hat{P}^{(2)}_n$ in conditional probability to the Markov I-projection, the latter need not be irreducible.

Example 1: Let $X_1, X_2, \cdots$ be i.i.d. random variables uniformly distributed on $\boldsymbol{X} = \{0,1\}$; thus $W(y|x) = 1/2$ for all $(x, y) \in \boldsymbol{X}^2$. Let

$$\Pi = \{ P:\ P(1,0)P(0,1) = 0, \quad -P(1,0) \leq \bar{P}(1) - \bar{P}(0) \leq P(0,1) \}.$$

Then Π_0 consists of a single distribution P^* with $P^*(0,0) = P^*(1,1) = 1/2$ and this P^* is the Markov I-projection of W on Π_0. The second-order type of a sequence $\boldsymbol{x} = (x_1, \cdots, x_{n+1}) \in \boldsymbol{X}^{n+1}$ belongs to Π iff its first $\lceil n/2 \rceil$ digits are 0's and the others are 1's, or the first $\lceil n/2 \rceil$ digits are 1's and the rest are 0's (where $\lceil\ \rceil$ denotes "smallest integer not less than"). In this example, the mutally equivalent conditions in Lemma 2 c) are clearly fulfilled (both for $u = 0$ and $u = 1$) and the assertions of Theorem 2 are immediately obvious.

Notice that the Markov I-projection P^* enters the assertions of Theorem 2 through the conditional probabilities $P^*(\cdot|\cdot)$ only. A minor modification of the proof shows that instead of the existence of Markov I-projection, i.e., of a unique P^* minimizing $D(P\|W)$ subject to $P \in \Pi_0$, it suffices to adopt the weaker hypothesis that for any two minimizing P_1^* and P_2^*, both $P_1^*(\cdot|\cdot) = P_2^*(\cdot|\cdot)$ and $S(\bar{P}_1^*) = S(\bar{P}_2^*)$. By Lemma 1, the first one of these conditions is always satisfied if Π_0 is convex. It appears likely that in the convex case, the last condition can be dispensed with, so that then (18) with $u = x_1$ always implies (20) whenever $x_1 \in S(\bar{P}^*)$ for some $P^* \in \Pi_0$ minimiz-

ing $D(P\|W)$. In general, however, the uniqueness of $P^*(\cdot|\cdot)$ is not a sufficient substitute for that of P^* in Theorem 2, as the second part of the following example shows.

Example 2: Let $X_1, X_2 \cdots$ be as in Example 1. Then for $\Pi = \{P\colon P(1,0) = 0\}$, Π_0 consists of all distributions with $P(0,0) + P(1,1) = 1$, and $D(P\|W) = \log 2$ is constant for $P \in \Pi_0$. Thus P^* is not unique but $P^*(\cdot|\cdot)$ is, and it equals the unit matrix. Clearly, Theorem 2 is valid with this $P^*(\cdot|\cdot)$. On the other hand, let $\Pi = \{P\colon P(0,0) = 1$ or $P(0,0) = 0,\ P(0,1) = 2P(1,0)\}$. Then Π_0 consists of two elements, concentrated on $(0,0)$ and $(1,1)$; both achieve $\min D(P\|W)$ subject to $P \in \Pi_0$, and $P^*(\cdot|\cdot)$ is again the unit matrix. Now an $(x_1, \cdots, x_{n+1}) \in \{0,1\}^{n+1}$ with $x_1 = 0$ belongs to Π if and only if either $x_i = 0$, $i = 1, \cdots, n+1$, or $x_i = 1$ for $i = 2, \cdots, n+1$ except for exactly one $i \le n$. As all these sequences have probability 2^{-n-1}, we see that (18) is valid. Further, for $u = 0$,

$$\Pr\left\{X_2 = 0 \middle| \hat{P}_n^{(2)} \in \Pi,\ X_1 = 0\right\} = \frac{1}{n},$$

so that the assertion of Theorem 2 does not hold in this case.

The next example shows that the aperiodicity of P^* is essential for Theorem 3.

Example 3: Let $X_1, X_2, \cdots$ be i.i.d. Bernoulli random variables with $\Pr\{X_i = 0\} = q$, $0 < q < 1/2$. Let Π be the set of those distributions P on $\{0,1\}^2$ for which $P(0,0) = P(1,1) = 0$. Then Π_0 consists of a single distribution P_0 with $P_0(0,1) = P_0(1,0) = 1/2$, and (18) and (19) hold as does (20). The condition $\hat{P}_n^{(2)} \in \Pi$ now means that the sample $X_1, \cdots, X_{n+1}$ is an alternating sequence of zeros and ones. The two possible such sequences of length $n+1$ are equiprobable if n is odd and have probabilities qa_n and $(1-q)a_n$ if n is even, where $a_n = q^{n/2}(1-q)^{n/2}$. It follows that for every $0 \le i \le n+1$,

$$\Pr\left\{X_i = 0 \middle| \hat{P}_n^{(2)} \in \Pi\right\} = \begin{cases} \frac{1}{2}, & \text{if } n \text{ is odd} \\ q, & \text{if } n \text{ is even and } i \text{ is odd} \\ 1-q, & \text{if } n \text{ and } i \text{ are even.} \end{cases}$$

Thus in this example, $\lim_{n\to\infty} \Pr\{X_{l_n} = 0 | \hat{P}_n^{(2)} \in \Pi\}$ does not exist for any choice of l_n.

Example 4: Let $X_1, X_2 \cdots$ be i.i.d. random variables uniformly distributed on $\mathcal{X} = \{0,1\}$. Let Π be the set of those distributions P on $\mathcal{X}^2$ for which $P(0,0) \ge \alpha$, with $1/4 < \alpha < 1$. Then Π is of the form (11), with a single function h (the indicator of the point $(0,0)$), and $\hat{P}_n^{(2)} \in \Pi$ means that the count of $(0,0)$ pairs in the sample $X_1, \cdots, X_{n+1}$ is at least αn. By Theorem 4, the assertions (18)–(22) are valid in this case, where P^* minimizes

$$\sum_{(x,y) \in \{0,1\}^2} P(x,y) \log P(y|x)$$

subject to $P(0,0) \ge \alpha$ and $P(0,1) = P(1,0)$. This example represents about the most regular case conceivable. We claim that (26) is false even in this "nice" case.

Let us denote by N_n and $N_{n,2}$ the count of $(0,0)$ pairs in the sample $X_1, \cdots, X_{n+1}$ and $X_3, \cdots, X_{n+1}$, respectively. In view of (21), our claim will be established if we show that

$$\Pr\{X_1 = 0 | N_n \ge \alpha n\} - \Pr\{X_2 = 0 | N_n \ge \alpha n\}$$

does not tend to 0 as $n \to \infty$. This difference equals

$$\begin{aligned} &\frac{\Pr\{X_1 = 0,\ X_2 = 1,\ N_n \ge \alpha n\} - \Pr\{X_1 = 1,\ X_2 = 0,\ N_n \ge \alpha n\}}{\Pr\{N_n \ge \alpha n\}} \\ &= \frac{1}{4} \frac{\Pr\{N_{n,2} \ge \alpha n\} - \Pr\{N_{n,2} \ge \alpha n \text{ or } X_3 = 0,\ N_{n,2} = \lceil \alpha n \rceil - 1\}}{\Pr\{N_n \ge \alpha n\}} \\ &= -\frac{1}{8} \frac{\Pr\{N_{n,2} = \lceil \alpha n \rceil - 1 | X_3 = 0\}}{\Pr\{N_n \ge \alpha n\}}, \end{aligned}$$

where $\lceil\ \rceil$ denotes "smallest integer not less than." Using Proposition W, a simple calculation shows that this does not tend to zero as $n \to \infty$; thus (26) is, indeed, false.

The results in this paper easily generalize to kth-order empirical distributions with $k > 2$, i.e., to events $\hat{P}_n^{(k)} \in \Pi$ where Π is now a subset of $\Lambda^{(k)}$; at the same time, the hypothesis on $X_1, X_2, \cdots$ may be weakened to Markovity of order higher than 1. If X_1, X_2 is a Markov chain of order $k-1$ with

$$\Pr\{X_{l+k} = x_k | X_{l+1} = x_1, \cdots, X_{l+k-1} = x_{k-1}\} = W(x_k | x_1, \cdots, x_{k-1}), \qquad l = 0, 1, \cdots,$$

then $Y_1, Y_2, \cdots$ defined by $Y_i = (X_i, \cdots, X_{i+k-2})$ is a (first-order) Markov chain with state space $\mathcal{X}^{k-1}$ and transition probability matrix $\tilde{W}$ where

$$\tilde{W}(x_1', \cdots, x_{k-1}' | x_1, \cdots, x_{k-1}) = \begin{cases} W(x_k | x_1, \cdots, x_{k-1}), & \text{if } (x_1', \cdots, x_{k-1}') = (x_2, \cdots, x_k) \\ 0, & \text{if } (x_1', \cdots, x_{k-2}') \ne (x_2, \cdots, x_{k-1}) \end{cases}.$$

For any distribution $P \in \Lambda^{(k)}$, let $\tilde{P}$ denote its image under the mapping $(x_1, \cdots, x_k) \to ((x_1, \cdots, x_{k-1}), (x_2, \cdots, x_k))$. Then $\hat{P}_n^{(k)} \in \Pi$ if and only if the second-order empirical distribution of $Y_1, \cdots, Y_{n+1}$ belongs to $\tilde{\Pi} = \{\tilde{P}\colon P \in \Pi\}$, and for elements of $\tilde{\Pi}$ we have

$$D(\tilde{P}\|\tilde{W}) = \sum_{x_1, \cdots, x_k} P(x_1, \cdots, x_k) \log \frac{P(x_k | x_1, \cdots, x_{k-1})}{W(x_k | x_1, \cdots, x_{k-1})}. \tag{66}$$

The extensions of Theorems 1–3 to kth-order empirical distributions of Markov chains of order $k-1$ are obtained by applying these very theorems to the Markov chain $Y_1, Y_2, \cdots$. In these extensions, the role of $\min_{P \in \Pi_0} D(P\|W)$ will be played by the minimum of (66) for $P \in \operatorname{cl} \Pi$ with $\bar{P} = \bar{\bar{P}}$ where $\bar{P}$ and $\bar{\bar{P}}$ are now defined by

$$\bar{P}(x_1, \cdots, x_{k-1}) = \sum_{x_k} P(x_1, \cdots, x_k),$$

$$\bar{\bar{P}}(x_2, \cdots, x_k) = \sum_{x_1} P(x_1, \cdots, x_k). \tag{67}$$

The role of the Markov I-projection will be played by the (k-dimensional) distribution attaining this minimum, and instead of the Markov chain determined by the former, we will have the Markov chain of order $(k-1)$ determined by the latter. Notice that $\tilde{W}$ has many zeros, and the support of each $\tilde{P} \in \tilde{\Pi}$ is contained in a proper subset of $\boldsymbol{X}^{k-1} \times \boldsymbol{X}^{k-1}$. Hence for the extensions of Theorems 2–4 just mentioned, it is essential that the hypotheses of these theorems do not require a strictly positive transition probability matrix, nor the existence of a $P \in \Pi_0$ with support $S(P) = \boldsymbol{X}^2$.

We formulate explicitly only the extension of Theorem 4. To this end, let a subset E of $\boldsymbol{X}^k$ be called irreducible if to any $(x_1, \cdots, x_{k-1}) \in \boldsymbol{X}^{k-1}$ and $x \in \boldsymbol{X}$ there exist some $l \geq k$ and elements $x_k, \cdots, x_l$ of $\boldsymbol{X}$ with $x_l = x$ such that $(x_i, \cdots, x_{i+k-1}) \in E$, $i = 1, \cdots, l-k+1$. If such $x_k, \cdots, x_l$ exist for every sufficiently large l, we say that E is aperiodic.

Theorem 5: Let E be a given irreducible subset of $\boldsymbol{X}^k$ such that $W(x_k|x_1, \cdots, x_{k-1}) > 0$ for each $(x_1, \cdots, x_k) \in E$. Let $h_1, \cdots, h_r$ be given functions on $\boldsymbol{X}^k$ and $\alpha_1, \cdots, \alpha_r$ be constants, and put

$$A_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} h_j(X_i, \cdots, X_{i+k-1}) \geq \alpha_j, \quad j = 1, \cdots, r; \right.$$
$$\left. (X_i, \cdots, X_{i+k-1}) \in E, \quad i = 1, \cdots, n \right\}.$$

Then there exists a unique $P^* \in \Lambda^{(k)}$ minimizing

$$\sum_{x_1, \cdots, x_k} P(x_1, \cdots, x_k) \log \frac{P(x_k|x_1, \cdots, x_{k-1})}{W(x_k|x_1, \cdots, x_{k-1})}$$

subject to

$$S(P) \subset E, \ \bar{P} = \bar{\bar{P}} \tag{68}$$

and

$$\sum_{x_1, \cdots, x_k} P(x_1, \cdots, x_k) h_j(x_1, \cdots, x_k) \geq \alpha_j,$$
$$j = 1, \cdots, r, \tag{69}$$

whenever there exists some $P \in \Lambda^{(k)}$ satisfying (68) and the strict inequalities in (69). In this case for every $m \geq k$ and $(x_1, \cdots, x_m) \in \boldsymbol{X}^m$,

$$\lim_{n \to \infty} \Pr\{ X_k = x_k, \cdots, X_m = x_m | A_n, X_1 = x_1, \cdots, X_{k-1} = x_{k-1} \} = \prod_{i=0}^{m-k} P^*(x_{i+k}|x_{i+1}, \cdots, x_{i+k-1}).$$

If, in addition, E is aperiodic, then for any sequence of integers l_n with $l_n \to \infty$, $n - l_n \to \infty$,

$$\lim_{n \to \infty} \Pr\{ X_{l_n+1} = x_1, \cdots, X_{l_n+m} = x_m | A_n \} = \bar{P}^*(x_1, \cdots, x_{k-1}) \prod_{i=0}^{m-k} P^*(x_{i+k}|x_{i+1}, \cdots, x_{i+k-1}).$$

V. Conclusions

If $X_1, X_2, \cdots$ is a Markov chain with transition probability matrix W, the probability that $X_1, \cdots, X_{n+1}$ has second-order type $\hat{P}_n^{(2)} = P$ is approximately $\exp\{-nD(P\|W)\}$. Since these probabilities decrease exponentially in n, the exponent of the probability that $\hat{P}_n^{(2)} \in \Pi$ is determined by those second-order types in Π which are close to P^*, where P^* minimizes $D(P\|W)$ over all $P \in \Pi$ having equal marginals. Thus, under certain regularity conditions, $\Pr\{\hat{P}_n^{(2)} \in \Pi\}$ will be approximately $\exp\{-nD(P^*\|W)\}$, and the conditional probability that $\hat{P}_n^{(2)}$ is near P^* given that $\hat{P}_n^{(2)} \in \Pi$ tends to 1 as $n \to \infty$. It is then expected that the conditional joint distribution of the X_i's, given that $\hat{P}_n^{(2)} \in \Pi$, will be close to the distribution of the Markov chain determined by P^*. In fact, using the exact formula for the number of sequences of a given second-order type starting with a given $x_1 \in \boldsymbol{X}$, and using the fact that all such sequences have the same probability, we have proved that

$$\Pr\{ X_2 = x_2, \cdots, X_m = x_m | \hat{P}_n^{(2)} \in \Pi, \ X_1 = x_1 \} \to \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i)$$

as $n \to \infty$. The initial state $X_1 = x_1$ requires special treatment because

$$\Pr\{ X_1 = x_1, \ X_2 = x_2, \cdots, X_m = x_m | \hat{P}_n^{(2)} \in \Pi \}$$

does not converge to the unconditional Markov probability

$$\bar{P}^*(x_1) \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i).$$

This sensitivity to end effects can be eliminated by looking at interior segments, where it is indeed true that

$$\Pr\{ X_{l_n+1} = x_1, \cdots, X_{l_n+m} | \hat{P}_n^{(2)} \in \Pi \} \to \bar{P}^*(x_1) \prod_{i=1}^{m-1} P^*(x_{i+1}|x_i)$$

if both l_n and $n - l_n$ go to infinity as $n \to \infty$. These results are then specialized to

$$\Pi = \left\{ P: \sum_{x,y} P(x,y) h_j(x,y) \geq \alpha_j, \quad j = 1, \cdots, k \right\},$$

in which case the condition $\hat{P}_n^{(2)} \in \Pi$ is identical to

$$\frac{1}{n}\sum_{i=1}^{n} h_j(X_i, X_{i+1}) \geq \alpha_j, \qquad j=1,\cdots,n.$$

Our results support the so-called "maximum entropy" or "minimum discrimination information" principle: If new information requires "updating" of an original probability assignment, the new probability assignment should be the closest possible to the original in the sense of Kullback–Leibler information divergence.

APPENDIX

Proof of Lemma 1: Let $P^* \in \Pi_0$ minimize $D(P\|W)$ subject to $P \in \Pi_0$ (since Π_0 is closed, such a P^* surely exists) and pick an arbitrary $P \in \Pi_0$. Then

$$P_\alpha = \alpha P + (1-\alpha)P^* \in \Pi_0$$

for every $0 \leq \alpha \leq 1$ (by convexity) and $D(P_\alpha\|W)$ is minimized for $\alpha = 0$. A simple calculation yields, for $0 < \alpha < 1$,

$$\frac{\partial}{\partial \alpha} D(P_\alpha\|W) = \sum_{x,y} (P(x,y) - P^*(x,y)) \log \frac{P_\alpha(x,y)}{\overline{P}_\alpha(x) W(y|x)}.$$

As

$$\lim_{\alpha \to 0} \frac{P_\alpha(x,y)}{\overline{P}_\alpha(x)} = \begin{cases} P^*(x,y)/\overline{P}^*(x), & \text{if } x \in S(\overline{P}^*) \\ P(x,y)/\overline{P}(x), & \text{if } x \in S(\overline{P}) - S(\overline{P}^*), \end{cases}$$

it follows that

$$\begin{aligned} 0 &\leq \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} D(P_\alpha\|W) \\ &= \sum_{x \in S(\overline{P}^*)} \sum_{y \in X} (P(x,y) - P^*(x,y)) \log \frac{P^*(x,y)}{\overline{P}^*(x) W(y|x)} \\ &\quad + \sum_{x \in X - S(\overline{P}^*)} \sum_{y \in X} P(x,y) \log \frac{P(x,y)}{\overline{P}(x) W(y|x)}. \end{aligned} \tag{A.1}$$

One consequence of (A.1) is that $P^*(x,y) > 0$ whenever $P(x,y) > 0$ and $\overline{P}^*(x) > 0$. In particular, if P is irreducible, then necessarily $S(P) \subset S(P^*)$. This means that P^* is irreducible if Π_0 contains some irreducible P. As, by convexity, there exists $P \in \Pi_0$ with $S(P) = S(\Pi_0)$, this proves that $S(P^*) = S(\Pi_0)$ if $S(\Pi_0)$ is irreducible. Then $S(\overline{P}^*) = X$ and thus (A.1) gives (15).

To prove the last assertion and the uniqueness of P^* in the case when $S(\Pi_0)$ is irreducible, suppose that P_1^* and P_2^* both attain $\min D(P\|W)$ subject to $P \in \Pi_0$, and set $P^* = \alpha P_1^* + (1-\alpha)P_2^*$ $(0 < \alpha < 1)$. Then by the identity (51), it follows that

$$D(P_1^*\|P^*(\cdot|\cdot)) = D(P_2^*\|P^*(\cdot|\cdot)) = 0,$$

for otherwise $D(P^*\|W)$ would be strictly less than $D(P_1^*\|W) = D(P_2^*\|W)$. Thus $P_1^*(\cdot|x) = P_2^*(\cdot|x) = P^*(\cdot|x)$ for $x \in S(\overline{P}_1) \cap S(\overline{P}_2)$, as claimed. If $S(\Pi_0)$ is irreducible, then a P^* attaining $\min D(P\|W)$ must be irreducible. Hence, the last result means that P^* is unique.

REFERENCES

[1] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, pp. 12–40; Correction: p. 1343, 1961.

[2] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *Ann. Math. Statist.*, vol. 42, pp. 1992–2007, 1971.

[3] I. Csiszár, "I-Divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[4] ____, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, pp. 768–793, 1984.

[5] ____, "A generalized maximum entropy principle and its Bayesian justification," *Bayesian Statistics*, 2, pp. 83–98, Proc. 2nd Valencia Int'l Symp. on Bayesian Statistics, North Holland, 1985.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[7] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 431–438, 1981.

[8] M. D. Donsker, and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time I–III," *Comm. Proc. App. Math.*, vol. 28, pp. 1–47, 279–301, and vol. 29, pp. 389–461, 1975–76.

[9] P. Groeneboom, J. Oosterhoff, and F. H. Ruymgaart, "Large deviation theorems for empirical probability measures," *Ann. Probab.*, vol. 7, pp. 553–586, 1979.

[10] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.* vol. 36, pp. 1916–1921, 1965.

[11] J. Justesen, and T. Hoholdt, "Maxentropic Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 665–667, 1984.

[12] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 360–365, 1985.

[13] I. N. Sanov, "On the probability of large deviations of random variables," *Mat. Sb.*, vol. 42, pp. 11–44, 1957 (in Russian). English translation in *Sel. Transl. Math. Statist. Probab.*, vol. 1, pp. 213–244, 1961.

[14] F. Spitzer, "A variational characterization of finite Markov chains," *Ann. Math. Statist.*, vol. 43, pp. 303–307, 1972.

[15] J. M. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 483–489, 1981.

[16] O. A. Vasicek, "A conditional law of large numbers," *Ann. Probab.*, vol. 8, pp. 142–147, 1980.

[17] P. Whittle, "Some distributions and moment formulae for the Markov chain, *J. Roy. Stat. Soc.*, Ser. B, 17, pp. 235–242, 1955.

[18] A. Zaman, "An approximation theorem for finite Markov exchangeability," Tech. Rep. No. 176, Dept. of Statistics, Stanford University, Stanford, CA, 1981.

[19] ____, "A finite form of DeFinetti's theorem for stationary Markov exchangeability," *Ann. Probab.*, vol. 14, pp. 1418–1427, 1986.