The Annals of Statistics
2000, Vol. 28, No. 6, 1601–1619

THE CONSISTENCY OF THE BIC MARKOV ORDER ESTIMATOR

BY IMRE CSISZÁR[1] AND PAUL C. SHIELDS[2]

Hungarian Academy of Sciences and University of Toledo

The Bayesian Information Criterion (BIC) estimates the order of a Markov chain (with finite alphabet A) from observation of a sample path $x_1, x_2, \ldots, x_n$, as that value $k = \hat{k}$ that minimizes the sum of the negative logarithm of the kth order maximum likelihood and the penalty term $\frac{|A|^k(|A|-1)}{2} \log n$. We show that $\hat{k}$ equals the correct order of the chain, eventually almost surely as $n \to \infty$, thereby strengthening earlier consistency results that assumed an apriori bound on the order. A key tool is a strong ratio-typicality result for Markov sample paths. We also show that the Bayesian estimator or minimum description length estimator, of which the BIC estimator is regarded as an approximation, fails to be consistent for the uniformly distributed i.i.d. process.

1. Introduction. A Markov chain is a discrete stochastic process $\{X_n\colon n \geq 1\}$ with values in a set A, called the alphabet, of cardinality $|A| < \infty$, for which there is a $k \geq 1$ such that

$$\text{(1.1)} \qquad \text{Prob}(X_1^n = x_1^n) = \text{Prob}\left(X_1^k = x_1^k\right) \prod_{i=k+1}^{n} Q\left(x_i | x_{i-k}^{i-1}\right), \qquad n \geq k,\ x_1^n \in A^n,$$

for suitable transition probabilities $Q(\cdot|\cdot)$. Here and in the sequel, x_m^n denotes the sequence $x_m, x_{m+1}, \ldots, x_n$. The class of processes such that (1.1) holds for a given $k \geq 1$ will be denoted by $\mathscr{M}_k$, and $\mathscr{M}_0$ will denote the class of i.i.d. processes. The *order* of a process in $\mathscr{M} = \cup_{k=0}^{\infty} \mathscr{M}_k$ is the smallest integer k_0 such that for some $\ell \geq 1$, $\{X_n\colon n \geq \ell\}$ is in $\mathscr{M}_{k_0}$.

One popular approach to model selection is the so-called Bayesian Information Criterion (BIC) of Schwarz [16]. Applied to estimating the order of a Markov chain, this gives the estimator

$$\text{(1.2)} \qquad \hat{k}_{\text{BIC}} = \hat{k}_{\text{BIC}}(x_1^n) = \arg\min_k \left(-\log P_{\text{ML}(k)}(x_1^n) + \frac{|A|^k(|A|-1)}{2} \log n \right),$$

where $P_{\text{ML}(k)}(x_1^n)$ is the kth order maximum likelihood, that is, the largest probability given to x_1^n by processes in $\mathscr{M}_k$. Our principal result is that this estimator is consistent.

It is a kind of folklore that "BIC is consistent," that is, recovers the right model class, eventually almost surely, provided one of the candidate model

Received February 1999; revised October 2000.

[1]Supported in part by a joint NSF-Hungarian Academy Grant 92 and by the Hungarian National Foundation for Scientific Research Grant T26041.

[2]Supported in part by A joint NSF-Hungarian Academy Grant INT-9515485.

AMS 2000 *subject classifications*. Primary 62F12, 62M05; secondary 62F13, 60J10.

Key words and phrases. Bayesian Information Criterion, order estimation, ratio-typicality, Markov chains.

classes is right. Consistency proofs are available for various situations such as i.i.d. processes with distributions from exponential families (Haughton [10]), autoregressive processes (Hannan and Quinn [9]) and Markov chains (Finesso [6]). All these proofs include the assumption that the number of candidate model classes is finite; for Markov chains, this means that there is a known upper bound k^* on the order of the process and the minimization in (1.2) is for $k \le k^*$. Consistency results of this kind may satisfy the practitioner, to whom consistency is of limited interest anyhow, for the true law of the process seldom belongs exactly to one of the candidate model classes (and the practitioner may not assume that a true law exists at all). From a theoretical point of view, however, restriction to a finite number of candidate model classes appears artificial, and it is a relevant question whether consistency also holds without such restriction. We answer this question in the positive, for the Markov chain case.

A process in $\mathscr{M} = \cup_{k=0}^{\infty} \mathscr{M}_k$ is irreducible if it is either i.i.d. or belongs to $\mathscr{M}_k$ for some $k \ge 1$ and has the additional property that the k-blocks that occur with positive probability communicate.

THE BIC CONSISTENCY THEOREM. *For any irreducible process in $\mathscr{M}$, $\hat{k}_{\mathrm{BIC}}(X_1^n)$ is eventually almost surely equal to the order of the process.*

The restriction to the irreducible case is justifiable for two reasons. First, if transient blocks were permitted, the order of the process could depend on those, and then it could not be estimated consistently. The definition of order could, however, be modified so as to keep the theorem valid for this case also. Second, the theorem is not true if the process is a mixture of two processes of different orders, because the sample paths drawn from the lower order component have positive probability and on these the BIC estimator is eventually almost surely equal to the lower order.

Consistent Markov order estimators of the penalized likelihood type, not assuming a prior bound on the order, have been known before, see Kieffer [11]. There, the penalty term was larger than the BIC penalty term $\frac{1}{2}|A|^k(|A| - 1)\log n$ in (1.2) and BIC consistency was raised as an open question. We note that the previously cited consistency proofs (assuming a finite number of model classes) were not restricted to the BIC, rather, consistency was proved even with penalty terms growing as slow as $\log\log n$ (but sample-size independent penalty terms, such as in AIC, do not suffice for consistency). Our proof uses the particular form of the BIC penalty term rather strongly, and it remains open whether smaller penalty terms suffice for consistency in the absence of a prior bound on the order.

In the sequel, processes will be identified with their distributions. Thus, if a probability measure P on A^∞ is the distribution of a process in $\mathscr{M}$, we will write $P \in \mathscr{M}$ and say that P is a process in $\mathscr{M}$. Note that to each each irreducible $P \in \mathscr{M}$ there exists a stationary (i.e., shift-invariant) $Q \in \mathscr{M}$ whose order and transition matrix are the same as those of P. This "stationary modification" of the process P is obtained by replacing $\mathrm{Prob}(X_1^k = x_1^k)$

in (1.1) by the limit as $n \to \infty$ of the arithmetic mean of the probabilities $\text{Prob}(X_i^{i+k-1} = x_1^k)$, $i = 1, \ldots, n$. Any irreducible process in $\mathscr{M}$ is absolutely continuous with respect its stationary modification, hence it suffices to prove the consistency theorem for stationary irreducible processes in $\mathscr{M}$.

An essential tool in proving the consistency of the BIC order estimator is a strong ratio-typicality result we establish for stationary, irreducible Markov chains, a result that appears to be of independent interest. It says, loosely, that as long as k does not grow too rapidly with sample path length, the ratio of the empirical relative frequency of each k-block to its probability is uniformly close to 1. The empirical distribution of k-blocks in x_1^n is defined by the formula

$$\hat{P}(a_1^k|x_1^n) = \frac{1}{n-k+1} N(a_1^k|x_1^n), \qquad a_1^k \in A^k, \tag{1.3}$$

where $N(a_1^k|x_1^n) = |\{i \in [1, n-k+1]\colon\ x_i^{i+k-1} = a_1^k\}|$ is the number of occurrences of a_1^k in x_1^n.

The sequence x_1^n is called (k, ε)-*typical* for a process Q if $\hat{P}(a_1^k|x_1^n) = 0$, whenever $Q(a_1^k) = 0$, and

$$\left| \frac{\hat{P}(a_1^k|x_1^n)}{Q(a_1^k)} - 1 \right| < \varepsilon \quad \text{whenever } Q(a_1^k) > 0. \tag{1.4}$$

THE TYPICALITY THEOREM. *For any stationary irreducible process $Q \in \mathscr{M}$ and any $0 < \beta < 1/2$ there exists $\alpha > 0$ such that eventually almost surely as $n \to \infty$, the sequence x_1^n is $(k, n^{-\beta})$-typical for every $k \le \alpha \log n$.*

Earlier typicality results in which block length grows with sample path length include the following.

1. Marton and Shields obtain large deviations bounds for the variational distance between the empirical k-block distribution and the theoretical k-block distribution that are valid for all $k \le (\log n)(H + \varepsilon)$, where H is the process entropy [13]. Their results do yield our typicality theorem for $k = o(\log \log n)$, but this is not sufficient for our proof of the BIC consistency theorem.
2. Flajolet, Kirschenhofer and Tichy have obtained similar results for the case when Q is the unbiased coin-tossing process [8]. They consider longer blocks but do not give an error rate.

The relevance of the typicality theorem for our consistency proof is that the probabilities

$$Q\left(\left\{x_1^n\colon\ \hat{k}_{\text{BIC}} = k,\, x_1^n \text{is } \left(k, n^{-\beta}\right)\text{-typical}\right\}\right)$$

can be bounded by direct counting arguments to show that their sum over all $k \in (k^*, \alpha \log n)$ is itself summable in n if k^* is sufficiently large. A second

idea, which comes from a more careful look at the Bayesian framework that underlies the BIC, will be used to show that

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \notin \left(\frac{\log\log n}{\log |A|}, n \right] \quad \text{eventually a.s.} \tag{1.5}$$

These results combine with the known consistency results, subject to $k \le k^*$ in (1.2), to establish the BIC consistency theorem.

The Bayesian framework starts with a prior distribution $\{p_k\}$ on the possible orders, together with a prior distribution on $\mathscr{M}_k$, for each k; the latter defines a mixture distribution, that is, a weighted average of the processes in $\mathscr{M}_k$. The Bayesian order estimator is that k for which the product of p_k and the mixture probability of x_1^n is largest. We consider here the mixture of those kth order Markov chains whose starting distribution is uniform on A^k, taking as prior the $|A|^k$-fold product of the $(\frac{1}{2}, \ldots, \frac{1}{2})$ Dirichlet distributions put on the transition probability matrices $Q(\cdot|\cdot)$. This mixture distribution plays a distinguished role in the theory of universal coding; see Krichevsky and Trofimov [12]. We denote it by KT_k; its explicit form is given later. The expression minimized in the definition (1.2) of $\hat{k}_{\mathrm{BIC}}(x_1^n)$ is an approximation to $-\log \mathrm{KT}_k(x_1^n)$ when k is fixed and $n \to \infty$. We will show, however, that for large n and k, it substantially overestimates $-\log \mathrm{KT}_k(x_1^n)$, a fact that easily leads to the result (1.5).

We will also consider the Bayesian order estimator

$$\hat{k}_{\mathrm{KT}}(x_1^n) = \arg\min_k \Big(-\log p_k - \log \mathrm{KT}_k(x_1^n) \Big), \tag{1.6}$$

which is also a minimum description length (MDL) estimator; see Rissanen [15] or Barron, Rissanen and Yu [2]. This is because $-\log \mathrm{KT}_k(x_1^n)$ is the description length (code length) for a universal code tailored to the model class $\mathscr{M}_k$.

In contrast to the BIC consistency theorem, we will prove the following.

THE INCONSISTENCY THEOREM. *If $\{X_n\}$ is i.i.d. with X_n uniformly distributed on A, then $\hat{k}_{\mathrm{KT}}(x_1^n) \to +\infty$, almost surely, provided the p_k of* (1.6) *are taken, as usual, to be slowly decreasing, say $p_k = ck^{-2}$.*

Bayesian inconsistency phenomena similar to ours are well known; see, for example, Diaconis and Freedman [5]. Our inconsistency result is interesting for three reasons:

(i) It concerns a natural problem with choice of priors commonly used in the literature.

(ii) The contrast to the BIC consistency theorem suggests a deficiency in the usual interpretation of the BIC order estimator as an approximation to the Bayesian.

(iii) An MDL-inspired result of Barron [1] (see also [2]) says the following: if the processes of $\mathscr{M}_k$ are parametrized by an $|A|^k(|A| - 1)$ dimensional parameter, then for Lebesgue almost all choices of the parameter, the estimator

$\hat{k}_{\mathrm{KT}}(x_1^n)$ is eventually almost surely equal to the correct order. Our inconsistency theorem shows that "for almost every choice of the parameter," is not vacuous.

The question remains open whether the Bayesian order estimator $\hat{k}_{\mathrm{KT}}$ may be inconsistent also for other processes than the uniform i.i.d. process. If the answer is no, as we conjecture, the inconsistency could be remedied by putting the uniform i.i.d. process into a class of its own (of order -1, say) and, accordingly, assigning to it a positive prior probability.

In addition to the Bayesian estimator $\hat{k}_{\mathrm{KT}}$ there are other MDL order estimators that correspond to different coding schemes tailored to model class $\mathscr{M}_k$. Another form of optimal code for the class $\mathscr{M}_k$, due to Shtarkov [18], has codeword lengths $-\log \mathrm{NML}_k(x_1^n)$, where NML_k is the normalized maximum likelihood

$$\mathrm{NML}_k(x_1^n) = P_{\mathrm{ML}(k)}(x_1^n) \Big/ \sum_{y_1^n \in A^n} P_{\mathrm{ML}(k)}(y_1^n).$$

The corresponding order estimator is obtained replacing KT_k by NML_k in (1.6). We will show that the inconsistency theorem also holds for this estimator. The estimator is related to $\hat{k}_{\mathrm{BIC}}$ is the same way as $\hat{k}_{\mathrm{KT}}$ is, that is, $-\log \mathrm{NML}_k(x_1^n)$ is also asymptotically equal to the expression minimized in (1.2) when k is fixed and $n \to \infty$.

As a reviewer suggested, we mention some recent results about model selection for a more refined model in which probabilities are assumed to depend on a variable number of steps in the past, with a bound on the depth of lookback. Such a process is Markov (with order equal to the depth of lookback), but now the user wishes to determine not only the order but also the explicit "context tree" that determines the probabilities. Weinberger, Rissanen and Feder [19], showed that the context tree can be consistently estimated by a search that assumes a bounded depth but the bound is allowed to grow slowly as the sample increases. Bühlman and Wyner [3], using a modified algorithm that does not require a prior depth bound, obtained consistency in probability even if the "true model" is allowed to depend on sample size (subject to regularity assumptions). Willems et al. [21] showed that a modified version of the "weighting algorithm" they develop for data compression, [20], is also suitable for consistent estimation of the context tree; here, however, a prior bound on depth is assumed. While none of these consistency results employ a BIC estimator, they may render our BIC consistency theorem more plausible.

Finally, we mention the paper of Papangelou [14] whose flavor is somewhat similar to our work, although no direct relationship of results is apparent. We should also mention that our methods, particularly the ratio-typicality idea, might be useful in establishing consistency in other settings; for brevity and to keep this paper focused we do not consider these here.

In outline, here is the structure of the remainder of the paper.

1. The proof of the BIC consistency theorem, assuming the typicality theorem, is given in Section 2. As alluded to earlier, separate arguments to rule

out moderate and gross overestimation are given in Sections 2.1 and 2.2, respectively.
2. The typicality theorem is proved using a martingale argument in Section 3.
3. The inconsistency theorem is established in Section 4. The key is that if k is large enough to make it unlikely that no k-block appears twice, then $\mathrm{KT}_k \sim |A|^{-n}$ while $\mathrm{KT}_0 \sim |A|^{-n} n^{\frac{|A|-1}{2}}$.

2. Proof of the BIC consistency theorem. We make use of the fact that the maximum likelihood $P_{\mathrm{ML}(k)}(x_1^n)$ can be expressed in terms of the empirical distribution of $(k+1)$-blocks. Indeed, using the notation $a_1^k \in x_1^n$ to mean that $N(a_1^k|x_1^n) > 0$, where, as defined earlier, $N(a_1^k|x_1^n)$ is the number of occurrences of a_1^k in x_1^n, the Markov condition (1.1) can be expressed as

$$\text{(2.1)} \qquad \mathrm{Prob}(X_1^n = x_1^n) = \mathrm{Prob}\left(X_1^k = x_1^k\right) \prod_{a_1^{k+1} \in x_1^n} Q(a_{k+1}|a_1^k)^{N(a_1^{k+1}|x_1^n)}.$$

For fixed $x_1^n \in A^n$, this probability is maximized if $\mathrm{Prob}(X_1^k = x_1^k) = 1$ and $Q(a_{k+1}|a_1^k) = N(a_1^{k+1}|x_1^n)/N(a_1^k|x_1^{n-1})$ whenever $a_1^k \in x_1^{n-1}$. It follows that

$$\text{(2.2)} \qquad P_{\mathrm{ML}(k)}(x_1^n) = \exp\left(-\sum_{a_1^{k+1} \in x_1^n} N(a_1^{k+1}|x_1^n) \log \frac{N(a_1^k|x_1^{n-1})}{N(a_1^{k+1}|x_1^n)}\right)$$

where $N(a_1^k|x_1^{n-1})$ is replaced by n in the case when $k = 0$.

As noted in the Introduction, it suffices to prove consistency of the BIC order estimator for stationary, irreducible processes in $\mathscr{M}$. Throughout this section Q will denote such a process of order k_0 and almost sure statements are with respect to Q. Further, we abbreviate $|A|^{k+1} - |A|^k$ by $\Delta_k A$ and put

$$B_{n,k} = \left\{x_1^n\colon\ \hat{k}_{\mathrm{BIC}}(x_1^n) = k\right\}.$$

Since $Q \in \mathscr{M}_{k_0}$ implies $Q(x_1^n) \le P_{\mathrm{ML}(k_0)}(x_1^n)$, it follows from the definition (1.2) of $\hat{k}_{\mathrm{BIC}}(x_1^n)$ that

$$\text{(2.3)} \qquad \log Q(x_1^n) \le \log P_{\mathrm{ML}(k)}(x_1^n) - \frac{\Delta_k A}{2} \log n + \frac{\Delta_{k_0} A}{2} \log n, \qquad x_1^n \in B_{n,k}.$$

By the consistency result with a prior bound on the order, see [6], we have, for any fixed $k^* > k_0$,

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \notin [0, k_0) \cup (k_0, k^*] \quad \text{eventually almost surely.}$$

For completeness a proof is given in the Appendix. The consistency theorem will be established if we can show that for suitable $k^* > k_0$,

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \le k^* \quad \text{eventually a.s.}$$

We sketch our initial attempts to prove the preceding via bounding the probabilities $Q(B_{n,k})$. Partition $B_{n,k}$ according to the empirical distribution of $(k+1)$-blocks in x_1^n. As the maximum likelihood is constant on a class of the

partition the class size is bounded by the reciprocal of that maximum likelihood. It follows, using (2.3), that each class has Q-probability $\leq n^{\frac{-\Delta_k A}{2}+\frac{\Delta_{k_0} A}{2}}$. This provides a useful bound for $Q(B_{n,k})$ only if there is available a bound sufficiently less than $n^{\frac{\Delta_k A}{2}-\frac{\Delta_{k_0} A}{2}}$ on the number of classes.

Counting ideas like the preceding are frequently used in information theory, where it is known as the method of types; for previous applications to Markov order estimation, see [7]. Unfortunately, the argument fails for our problem, since the number of classes is too large. It can be salvaged, however, by invoking the typicality theorem. The latter permits restricting attention to the part of $B_{n,k}$ consisting of $(k, n^{-\beta})$-typical sequences, and for that part a satisfactory bound on the number of classes can be obtained.

We show in Section 2.1 how to use the counting idea in conjunction with the typicality theorem to obtain the following result.

PROPOSITION 1. *There exist positive numbers $k^* \geq k_0$ and α, both depending on Q, such that $\hat{k}_{\mathrm{BIC}}(x_1^n) \notin (k^*, \alpha \log n)$, eventually almost surely.*

For large k the penalty term $\Delta_k A \log n$ becomes dominant. For example, if

$$k \geq \frac{\log n}{\log |A|}$$

then $\Delta_k A \geq n \log n$, so that (2.3) yields

$$\log Q(x_1^n) \leq -\frac{n}{2} \log^2 n + O(\log n), \qquad x_1^n \in B_{n,k}.$$

This cannot hold for large n since the Markov property implies there is positive constant c such that

$$\min_{x_1^n:\ Q(x_1^n)>0} Q(x_1^n) \geq c^n, \qquad n \geq 1.$$

A more sophisticated argument will be used in Section 2.2 to establish the following stronger result.

PROPOSITION 2.

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \notin \left(\frac{\log \log n}{\log |A|}, n \right],$$

eventually almost surely.

Combined with Proposition 1 and the preceding arguments, this yields the BIC consistency theorem.

2.1. *Proof of Proposition* 1. Two lemmas will be established, one bounding the number of typical sequences with a given empirical $(k+1)$-block distribution, the other bounding the number of possible such distributions. Used in conjunction with the typicality theorem, these lead to the proof of Proposition 1.

To facilitate the statements and proofs of our two lemmas we introduce the following terminology. A distribution P on $(k+1)$-blocks will be called *a* $(k+1)$*-type* (*with path length* n) if it is the empirical $(k+1)$-block distribution of some n-sequence. The $(k+1)$-type class of P, denoted by $\mathscr{T}_P^n$, is the set of all x_1^n for which $\hat{P}(a_1^{k+1}|x_1^n) = P(a_1^{k+1})$, for all a_1^{k+1}.

The typicality concept (1.4) extends to type classes. For each j, let $\mathscr{S}_j$ denote the support of the jth marginal of Q, that is, the set of all a_1^j for which $Q(a_1^j) > 0$. A $(k+1)$-type class $\mathscr{T}_P^n$ is ε-typical (for the given Q) if some $x_1^n \in \mathscr{T}_P^n$ (and hence every $x_1^n \in \mathscr{T}_P^n$) is $(k+1, \varepsilon)$-typical, that is, if

$$\left| \frac{P(a_1^{k+1})}{Q(a_1^{k+1})} - 1 \right| < \varepsilon, \quad a_1^{k+1} \in \mathscr{S}_{k+1}; \qquad P(a_1^{k+1}) = 0, \quad a_1^{k+1} \notin \mathscr{S}_{k+1}.$$

LEMMA 1. *There is a positive constant* $C = C(Q)$ *such that*

$$|\mathscr{T}_P^n| \le C^{|A|^k} (n-k)^{-\frac{|\mathscr{S}_{k+1}| - |\mathscr{S}_k|}{2}} \exp(-\log P_{\mathrm{ML}(k)}(x_1^n)), \qquad x_1^n \in \mathscr{T}_P,$$

for every ε*-typical* $(k+1)$*-type class* $\mathscr{T}_P^n$, *for all* $\varepsilon < 1/2$, *for all* n, *and for all* $1 \le k < n$.

PROOF. The easy case is $k = 0$ for then

$$|\mathscr{T}_P^m| = \binom{m + |\mathscr{S}(P)| - 1}{|\mathscr{S}(P)| - 1},$$

where $\mathscr{S}(P) = \{a \in A : P(a) > 0\}$, to which Stirling's formula can be applied. Indeed, using the refined version of Stirling's formula (see [4], Exercise 2, Section 1.2), a straightforward calculation yields the 1-*type bound*

$$|\mathscr{T}_P^m| \le \frac{(2\pi m)^{-\frac{|\mathscr{S}(P)|-1}{2}}}{\prod_{a \in \mathscr{S}(P)} \sqrt{P(a)}} \exp\Big(-\log P_{\mathrm{ML}(0)}(x_1^m)\Big), \qquad x_1^m \in \mathscr{T}_P^m. \tag{2.4}$$

For the case when $k > 0$, the idea is to partition $\mathscr{T}_P^n$ according to the first k-terms, then upper bound the cardinality of each set of the partition by the product of the cardinalities of the 1-types given by the conditional counts. This plan is carried out in the following paragraphs.

Fix $0 < \varepsilon < 1/2$ and $1 \le k < n$, and let P be an ε-typical $(k+1)$-type of path length n. Typicality implies that $P(a_1^k) > 0$ iff $a_1^k \in \mathscr{S}_k$ and $P(a_1^{k+1}) > 0$ iff $a_1^{k+1} \in \mathscr{S}_{k+1}$. Furthermore, denoting the support of the conditional distribution $P(\cdot|a_1^k)$ by $\mathscr{S}(\cdot|a_1^k)$, we have

$$\sum_{a_1^k \in \mathscr{S}_k} \frac{|\mathscr{S}(\cdot|a_1^k)| - 1}{2} = \frac{|\mathscr{S}_{k+1}| - |\mathscr{S}_k|}{2}. \tag{2.5}$$

Given $x_1^n \in \mathscr{T}_P^n$, let $x^*(a_1^k)$ denote the sequence of length $(n-k)P(a_1^k)$ consisting of the symbols x_{t_j} that follow the occurrences of a_1^k in x_1^{n-1}, that is, such that $x_{t_j-k}^{t_j-1} = a_1^k$, $k < t_j \le n$. Then $x^*(a_1^k)$ belongs to the 1-type class $\mathscr{T}_{P(\cdot|a_1^k)}^{(n-k)P(a_1^k)}$ and it follows that assigning to x_1^n the $|\mathscr{S}_k|$-tuple $\{x^*(a_1^k): a_1^k \in \mathscr{S}_k\}$ defines a mapping of $\mathscr{T}_P^n$ into the Cartesian product of the sets $\mathscr{T}_{P(\cdot|a_1^k)}^{(n-k)P(a_1^k)}$, $a_1^k \in \mathscr{S}_k$. Furthermore, this mapping is one-to-one when restricted to the set $\mathscr{T}_{P,x_1^k}^n$ consisting of all $x_1^n \in \mathscr{T}_P^n$ that start with a fixed x_1^k. Hence it follows from the 1-type bound (2.4) and the support relations (2.5) that

$$|\mathscr{T}_{P,x_1^k}^n| \le \prod_{a_1^k \in \mathscr{S}_k} \left|\mathscr{T}_{P(\cdot|a_1^k)}^{(n-k)P(a_1^k)}\right|$$

$$\le \Pi(P)(n-k)^{-\frac{|\mathscr{S}_{k+1}|-|\mathscr{S}_k|}{2}} \exp\left\{-\log P_{\mathrm{ML}(k)}(x_1^n)\right\}, \qquad x_1^n \in \mathscr{T}_P^n,$$

where

$$\Pi(P) = \prod_{a_1^k \in \mathscr{S}_k} \frac{(2\pi P(a_1^k))^{-\frac{|\mathscr{S}(\cdot|a_1^k)|-1}{2}}}{\prod_{a_{k+1} \in \mathscr{S}(\cdot|a_1^k)} \sqrt{P(a_{k+1}|a_1^k)}}, \tag{2.6}$$

since

$$\sum_{a_1^k \in \mathscr{S}_k} \log P_{\mathrm{ML}(0)} x^*(a_1^k)) = \log P_{\mathrm{ML}(k)}(x_1^n), \qquad x_1^n \in \mathscr{T}_P^n,$$

as can be seen by using formula (2.2) and some algebra.

Typicality and the assumption that $0 < \varepsilon < 1/2$ imply that

$$\Pi(P) \le c^{|A^k|}\Pi(Q), \tag{2.7}$$

where c is a constant, depending only on Q, and $\Pi(Q)$ is defined by replacing P by Q in (2.6). (One can show that $c = \sqrt{6}$ suffices.) The Markovian property, however, implies the existence of $c_* > 0$ such that $Q(a_1^j) \ge c_*^j$, for any j and any $a_1^j \in \mathscr{S}_j$. This fact, together with (2.7), easily yields the desired result. This completes the proof of Lemma 1. □

LEMMA 2. *The number of ε-typical $(k+1)$-type classes $\mathscr{T}_P^n$ is less than* $|A|^{2k}(1+2\varepsilon(n-k))^{|\mathscr{S}_{k+1}|-|\mathscr{S}_k|}$.

PROOF. It suffices to show that, fixing x_1^k and x_{n-k+1}^n, the number of ε-typical $(k+1)$-type classes that could contain x_1^n is less than $(1+2\varepsilon(n-k))^{|\mathscr{S}_{k+1}|-|\mathscr{S}_k|}$. Note that for a $(k+1,\varepsilon)$-typical x_1^n, among the numbers $N(a_1^{k+1}|x_1^n)$ exactly those with $a_1^{k+1} \in \mathscr{S}_{k+1}$ are positive, and each have less than $1+2\varepsilon(n-k)$ possible values, because typicality implies

$$\left|N\left(a_1^{k+1}|x_1^n\right) - (n-k)Q(a_1^{k+1})\right| < \varepsilon(n-k)Q(a_1^{k+1}) \le \varepsilon(n-k).$$

Hence the proof will be complete if we show that, fixing x_1^k and x_{n-k+1}^n, the numbers $N(a_1^{k+1}|x_1^n)$, $a_1^{k+1} \in \mathcal{S}_{k+1}$, satisfy $|\mathcal{S}_k|$ independent linear constraints, so that $|\mathcal{S}_{k+1}| - |\mathcal{S}_k|$ of these numbers uniquely determine the others.

Clearly, the following equations always hold:

$$\sum_{a_1^{k+1} \in \mathcal{S}_{k+1}} N\left(a_1^{k+1}|x_1^n\right) = n - k,$$

and for each $b_1^k \in \mathcal{S}_k$,

$$\sum_{a_1^{k+1} \in \mathcal{S}_{k+1}} N\left(a_1^{k+1}|x_1^n\right)\left(\mathbb{1}\left(a_1^k = b_1^k\right) - \mathbb{1}\left(a_2^{k+1} = b_1^k\right)\right)$$
$$= \mathbb{1}\left(a_1^k = x_1^k\right) - \mathbb{1}\left(a_1^k = x_{n-k+1}^n\right),$$

where here and in the remainder of the paper, $\mathbb{1}(\cdot)$ denotes the indicator function.

It is easy to see that dropping any one of the last $|\mathcal{S}_k|$ equations, the remaining $|\mathcal{S}_k|$ constraints are independent. Formally, $\mathbb{1}$ (with each component equal to 1) and $\left\{\mathbb{1}(a_1^k = b_1^k) - \mathbb{1}(a_2^{k+1} = b_1^k), a_1^{k+1} \in \mathcal{S}_{k+1}\right\}$ with b_1^k running over all k-blocks in $\mathcal{S}_k$ but one, are linearly independent vectors in $\mathbb{R}^{|\mathcal{S}_{k+1}|}$. □

PROOF OF PROPOSITION 1 COMPLETED (Assuming the typicality theorem). Let $0 < \beta < 1/2$ and $\alpha > 0$ be as in the typicality theorem for Q and let $C = C(Q)$ be the number given by Lemma 1. The bound (2.3) and Lemma 1 imply that

$$Q(\mathcal{T}_P^n \cap B_{n.k}) \le C^{|A|^k}(n-k)^{-\frac{\Delta_k \mathcal{S}}{2}} n^{\frac{-\Delta_k A}{2} + \frac{\Delta_{k_0} A}{2}}, \tag{2.8}$$

for any ε-typical $(k+1)$-type class $\mathcal{T}_P^n$ with $\varepsilon < 1/2$, where we used the notation $\Delta_k \mathcal{S} = |\mathcal{S}_{k+1}| - |\mathcal{S}_k|$ in addition to our earlier notation $\Delta_k A = |A|^{k+1} - |A|^k$.

Let $G_{n.k}$ be the union of the $n^{-\beta}$-typical $(k+1)$-type classes $\mathcal{T}_P^n$. The bound (2.8) and Lemma 2 combine to show that $Q(G_{n.k} \cap B_{n.k})$ is upper bounded by

$$C^{|A|^k}(1 + 2n^{-\beta}(n-k))^{\Delta_k \mathcal{S}}(n-k)^{-\frac{\Delta_k \mathcal{S}}{2}} n^{\frac{-\Delta_k A}{2} + \frac{\Delta_{k_0} A}{2}},$$

which is, in turn, upper bounded by

$$(\text{constant})^{|A|^k} n^{-\beta \Delta_k \mathcal{S}}(n-k)^{\frac{\Delta_k \mathcal{S}}{2}} n^{-\frac{\Delta_k A}{2}} n^{\frac{\Delta_{k_0} A}{2}}. \tag{2.9}$$

Here

$$n^{-\beta \Delta_k \mathcal{S}}(n-k)^{\frac{\Delta_k \mathcal{S}}{2}} n^{-\frac{\Delta_k A}{2}} \le n^{-(\Delta_k A - \Delta_k S(1-2\beta))/2} \le n^{-\beta \Delta_k A},$$

since $0 \le \Delta_k S \le \Delta_k A$. It follows that for k sufficiently large, the bound in (2.9) becomes smaller than any negative power of n, because $\Delta_k A = |A|^k(|A| - 1)$ goes to infinity as $k \to \infty$. Hence, there is a $k^* \ge k_0$ and an n^* such that

$$Q(G_{n.k} \cap B_{n.k}) \le n^{-2}, \qquad n \ge n^*,\ k^* \le k \le \alpha \log n,$$

from which it easily follows that

$$\sum_n Q\left(\bigcup_{k\in[k^*,\alpha\log n]}(G_{n.k}\cap B_{n.k})\right)<\infty.$$

The typicality theorem implies

$$x_1^n\in\bigcap_{k\in[k^*,\alpha\log n]}G_{n.k}\quad\text{eventually a.s.},$$

and an application of Borel-Cantelli yields Proposition 1. □

2.2. *Proof of Proposition* 2. Again we use the notation $a_1^k\in x_1^n$ to mean that $N(a_1^k|x_1^n)>0$. The proof of Proposition 2 relies on properties of the Krichevsky-Trofimov distributions KT_k. For $k=0$, KT_0 is the $(\frac{1}{2},\dots,\frac{1}{2})$-Dirichlet mixture of the i. i. d. distributions with alphabet A. In explicit terms, see [12],

$$\begin{aligned}\mathrm{KT}_0(x_1^n)&=\frac{\Gamma(\frac{|A|}{2})}{\Gamma(n+\frac{|A|}{2})}\prod_{a\in x_1^n}\frac{\Gamma(N(a|x_1^n)+\frac{1}{2})}{\Gamma(\frac{1}{2})}\\&=\frac{\prod_{a\in x_1^n}(N(a|x_1^n)-\frac{1}{2})(N(a|x_1^n)-\frac{3}{2})\cdots(\frac{1}{2})}{(n-1+\frac{|A|}{2})(n-2+\frac{|A|}{2})\cdots(\frac{|A|}{2})}.\end{aligned}\tag{2.10}$$

For $k\geq 1$, KT_k is the mixture of the kth order Markov chains whose starting distribution is uniform on A^k, the mixture taken with the $|A|^k$-fold product of $(\frac{1}{2},\dots,\frac{1}{2})$-Dirichlet distributions put on the transition matrices $Q(\cdot|\cdot)$. The explicit form is

$$\begin{aligned}&\mathrm{KT}_k(x_1^n)\\&=\frac{1}{|A|^k}\prod_{a_1^k\in x_1^{n-1}}\left[\frac{\prod_{a_{k+1}:a_1^{k+1}\in x_1^n}(N(a_1^{k+1}|x_1^n)-\frac{1}{2})(N(a_1^{k+1}|x_1^n)-\frac{3}{2})\cdots(\frac{1}{2})}{(N(a_1^k|x_1^{n-1})-1+\frac{|A|}{2})(N(a_1^k|x_1^{n-1})-2+\frac{|A|}{2})\cdots(\frac{|A|}{2})}\right].\end{aligned}\tag{2.11}$$

For our purposes (2.10) and (2.11) are taken as the definition of $\mathrm{KT}_k(x_1^n)$.

The next lemma summarizes the properties we shall need. A more precise result than inequality (2.12) appears in [20], for the case $|A|=2$.

LEMMA 3. *There is a constant C depending only on the alphabet size $|A|$ such that for every $n\geq 1$ and $x_1^n\in A^n$,*

$$\left|\log KT_0(x_1^n)-\sum_{a\in A}N(a|x_1^n)\log\frac{N(a|x_1^n)}{n}+\frac{|A|-1}{2}\log n\right|\leq C\tag{2.12}$$

and, for every $k\geq 1$,

$$\begin{aligned}&\left|\log KT_k(x_1^n)-\log P_{\mathrm{ML}(k)}(x_1^n)+\frac{|A|-1}{2}\sum_{a_1^k\in x_1^{n-1}}\log N(a_1^k|x_1^{n-1})\right|\\&\qquad\leq C|A|^k,\end{aligned}\tag{2.13}$$

and

$$\log KT_k(x_1^n) \geq \log P_{\mathrm{ML}(k)}(x_1^n) - \frac{\Delta_k A}{2}\left(1 + \log^+ \frac{n}{|A|^k}\right) - C|A|^k, \tag{2.14}$$

where $\log^+ t = \max(\log t, 0)$.

PROOF. The well-known bound (2.12) (see [12]) easily follows from Stirling's formula for Γ-functions and (2.13) is an immediate consequence of (2.12). Indeed, the square-bracketed factors in the definition (2.11) of KT_k are of the form of the definition (2.10) of KT_0, with $N(a_1^k|x_1^{n-1})$ and $N(a_1^{k+1}|x_1^n)$ in the role of n and $N(a|x_1^n)$, respectively. Applying (2.12) to each of these factors and recalling formula (2.2), that is,

$$\sum_{a_1^k \in x_1^{n-1}} \sum_{\substack{a_{k+1}: \\ a_1^{k+1} \in x_1^n}} N(a_1^{k+1}|x_1^n) \log \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})} = \log P_{\mathrm{ML}(k)}(x_1^n),$$

we obtain (2.13), with a somewhat larger C to take care of the logarithm of the $1/|A|^k$ factor that appears in the definition (2.11) of KT_k.

The final bound follows from (2.13), because, by the concavity of

$$f(t) = \begin{cases} \log t, & \text{if } t > 1, \\ t - 1, & \text{if } 0 \leq t \leq 1, \end{cases}$$

we have

$$\begin{aligned} \sum_{a_1^k \in x_1^{n-1}} \log N(a_1^k|x_1^{n-1}) &\leq \sum_{a_1^k \in A^k} f(N(a_1^k|x_1^{n-1})) + |A|^k \\ &\leq |A|^k f\left(\frac{n}{|A|^k}\right) + |A|^k \leq |A|^k \log^+\left(\frac{n}{|A|^k}\right) + |A|^k. \end{aligned}$$

This completes the proof of Lemma 3. □

To establish Proposition 2 we first recall (2.3):

$$\log Q(x_1^n) \leq \log P_{\mathrm{ML}(k)}(x_1^n) - \frac{\Delta_k A}{2}\log n + \frac{\Delta_{k_0} A}{2}\log n, \qquad x_1^n \in B_{n,k}.$$

Adding $-\log \mathrm{KT}_k(x_1^n)$ to both sides and applying the bound (2.14) of Lemma 3, we obtain with $k_n = \frac{\log\log n}{|A|}$ that

$$-\log \mathrm{KT}_k(x_1^n) + \log Q(x_1^n) \leq -3\log n, \ k \geq k_n, \ n \geq n_0, \qquad x_1^n \in B_{n,k}.$$

This implies that

$$Q(\cup_{k=k_n}^n B_{n,k}) \leq n^{-3} \sum_{k=k_n}^n \mathrm{KT}_k(B_{n,k}) \leq n^{-2}, \qquad n \geq n_0,$$

which, in turn, implies

$$\sum_{n=n_0}^{\infty} Q\left(\left\{x_1^n : \hat{k}_{\mathrm{BIC}}(x_1^n) \in [k_n, n]\right\}\right) \leq \sum_{n=n_0}^{\infty} n^{-2} < \infty.$$

An application of the Borel-Cantelli theorem yields Proposition 2. □

3. Proof of the typicality theorem. In this section, Q again denotes the distribution of a stationary irreducible Markov chain $\{X_n\colon\ n \geq 1\}$ of order k_0, and, for each j, $\mathscr{S}_j$ denotes the support of the jth order marginal of Q, that is, the set of all a_1^j for which $Q(a_1^j) > 0$. For brevity, in this section we denote the k-block frequencies $N(a_1^k|x_1^n)$ by $N_n(a_1^k)$ and, accordingly, we write $\hat{P}_n(a_1^k) = N_n(a_1^k)/(n-k+1)$.

To prove the typicality theorem, it suffices to establish the following two propositions.

PROPOSITION 3. *For any fixed k there is a constant C such that*

$$\left|\frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} - 1\right| < C\sqrt{\frac{\log\log n}{n}}, \qquad a_1^k \in \mathscr{S}_k, \text{ eventually a.s.}$$

PROPOSITION 4. *To any $0 < \beta < 1/2$ there exists $\alpha > 0$, which depends on Q, such that, eventually almost surely,*

$$\left|\frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} - \frac{\hat{P}_{n-(k-k_0)}(a_1^{k_0})}{Q(a_1^{k_0})}\right| < n^{-\beta}, \qquad a_1^k \in \mathscr{S}_k,\ k_0 < k \leq \alpha\log n. \tag{3.1}$$

PROOF OF PROPOSITION 3. For each $k \geq k_0$, the law of the iterated logarithm applied to the recurrence times of the (delayed) renewal process $\{\mathbb{1}(X_i^{i+k-1} = a_1^k)\colon\ i \geq 1\}$ shows that the proposition is true; see [6]. The assertion for smaller k obviously follows. □

PROOF OF PROPOSITION 4. The simplest idea would be to overbound the probability that (3.1) does not hold by some γ_n for which $\sum\gamma_n < \infty$, but this does not appear to be feasible. The same idea, however, works when merging "bad events" between consecutive powers of 2. We proceed by noticing first that whenever (3.1) is violated for some $n \in (2^{i-1}, 2^i]$, there exists $k \in (k_0, \alpha\log 2^i]$ and $a_1^k \in \mathscr{S}_k$ such that

$$\max_{2^{i-1}<n\leq 2^i}\left|\frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} - \frac{\hat{P}_{n-(k-k_0)}(a_1^{k_0})}{Q(a_1^{k_0})}\right| > 2^{-i\beta}, \tag{3.2}$$

Thus, letting $B_i(\alpha,\beta)$ be the event that (3.2) occurs for some $k \in (k_0, \alpha\log 2^i]$ and $a_1^k \in \mathscr{S}_k$, it is enough to show that

$$\sum_{i=1}^{\infty}\text{Prob}(B_i(\alpha,\beta)) < \infty, \tag{3.3}$$

for suitable $\alpha > 0$.

Using the factorization

$$Q(a_1^k) = Q(a_1^{k_0}) \prod_{j=k_0+1}^{k} Q(a_j|a_{j-k_0}^{j-1}),$$

inequality (3.2) can be rewritten as

$$\max_{2^{i-1}<n\le 2^i} \left| N_n(a_1^k) - N_{n-(k-k_0)}(a_1^{k_0}) \prod_{j=k_0+1}^{k} Q\left(a_j|a_{j-k_0}^{j-1}\right) \right| > 2^{-i\beta} 2^{i-2} Q(a_1^k). \tag{3.4}$$

The difference $N_n(a_1^k) - N_{n-(k-k_0)}(a_1^{k_0}) \prod_{j=k_0+1}^{k} Q(a_j|a_{j-k_0}^{j-1})$ can be decomposed as the sum

$$\begin{aligned} &\sum_{\ell=1}^{k-k_0} \left[N_{n-\ell+1}\left(a_1^{k-\ell+1}\right) - N_{n-\ell}\left(a_1^{k-\ell}\right) Q\left(a_{k-\ell+1}|a_{k-\ell+1-k_0}^{k-\ell}\right) \right] \\ &\qquad \times \prod_{j=k-\ell+2}^{k} Q\left(a_j|a_{j-k_0}^{j-1}\right), \end{aligned} \tag{3.5}$$

hence, if its absolute value exceeds a positive constant λ, then at least one term of the sum (3.5) has absolute value $> \lambda/(k-k_0)$. Using this fact, the probability of the event (3.4) is upper bounded by the sum, for $1 \le \ell \le k-k_0$, of the probabilities of the events

$$\begin{aligned} &\max_{2^{i-1}<n\le 2^i} \left| N_{n-\ell+1}\left(a_1^{k-\ell+1}\right) - N_{n-\ell}\left(a_1^{k-\ell}\right) Q\left(a_{k-\ell+1}|a_{k-\ell+1-k_0}^{k-\ell}\right) \right| \\ &\qquad > \frac{2^{-i\beta} 2^{i-2} Q(a_1^{k-\ell+1})}{k-k_0}. \end{aligned} \tag{3.6}$$

The key to the remainder of the proof is the observation that, for any $m \in (k_0, k]$, the sequence

$$\left\{ N_n(a_1^m) - N_{n-1}(a_1^{m-1}) Q\left(a_m|a_{m-k_0}^{m-1}\right) : n \ge m \right\}$$

is a martingale. Indeed, $N_n(a_1^k) = \sum_{i=1}^{n-k+1} \mathbb{1}(X_i^{i+k-1} = a_1^k)$ implies that

$$N_n(a_1^m) - N_{n-1}(a_1^{m-1}) Q\left(a_m|a_{m-k_0}^{m-1}\right) = \sum_{j=m}^{n} \Delta_j,$$

where

$$\Delta_j = \mathbb{1}\left(X_{j-m+1}^{j} = a_1^m\right) - \mathbb{1}\left(X_{j-m+1}^{j-1} = a_1^{m-1}\right) Q\left(a_m|a_{m-k_0}^{m-1}\right),$$

which satisfies $E(\Delta_j|X_1^{j-1}) = 0$. Furthermore,

$$E(\Delta_j^2) = Q(a_1^m)\left(1 - Q(a_m|a_{m-k_o}^{m-1})\right) \le Q(a_1^m),$$

so Kolmogorov's inequality for martingales gives that

$$\text{Prob}\left\{ \max_{n\le 2^i} \left| N_n(a_1^m) - N_{n-1}(a_1^{m-1}) Q(a_m|a_{m-k_0}^{m-1}) \right| > \lambda \right\} < \frac{2^i Q(a_1^m)}{\lambda^2}.$$

Applying this with $m = k - \ell + 1$ and $\lambda = 2^{k-i\beta-2}Q(a_1^{k-\ell+1})/(k-k_0)$, it follows that the sum, for $1 \le \ell \le k-k_0$, of the probabilities of the events (3.6) is upper bounded by

$$\sum_{\ell=1}^{k-k_0} \frac{2^i Q(a_1^{k-\ell+1})(k-k_0)^2}{[2^{i-i\beta-2}Q(a_1^{k-\ell+1})]^2} \le \frac{2^4 k^3}{2^{(1-2\beta)i}Q(a_1^k)}. \tag{3.7}$$

Next let γ be the minimum of $Q(a_1^{k_0})$, over all $a_1^{k_0} \in \mathcal{S}_{k_0}$, and let δ be the minimum of $Q(a_{k_0+1}|a_1^{k_0})$, over all $a_1^{k_0+1} \in \mathcal{S}_{k_0+1}$. For all k and all $a_1^k \in \mathcal{S}_k$, we then have $Q(a_1^k) \ge \gamma\delta^k$, and hence our argument shows that the probability that (3.2) holds for some fixed $k \in (k_0, \alpha \log 2^i]$ and $a_1^k \in \mathcal{S}_k$, is bounded above by $2^4k^3/2^{(1-2\beta)i}\gamma\delta^k$. Summing over $k \in (k_0, \alpha \log 2^i]$ and $a_1^k \in \mathcal{S}_k$ produces

$$\begin{aligned}\text{Prob}(B_i(\alpha,\beta)) &\le \sum_{k_0<k\le\alpha\log 2^i} \frac{2^4k^3|A|^k}{2^{(1-2\beta)i}\gamma\delta^k} \\ &< \frac{2^4(\alpha\log 2^i)^4(|A|/\delta)^{\alpha\log 2^i}}{\gamma 2^{(1-2\beta)i}} = \frac{(2\alpha\log 2)^4}{\gamma} i^4 2^{(2\beta-1+\alpha\log\frac{|A|}{\delta})i}.\end{aligned}$$

It follows from this final bound that if $\alpha > 0$ is chosen to satisfy $2\beta + \alpha \log \frac{|A|}{\delta} < 1$, then the sum of the probabilities of the events $B_i(\alpha, \beta)$ is indeed finite, that is, the desired result (3.3) holds. This completes the proof of the typicality theorem. □

4. Proof of the inconsistency theorem. In this section, Q denotes the uniform i.i.d. process,

$$Q(x_1^n) = |A|^{-n}, \qquad x_1^n \in A^n.$$

We prove that the Bayesian Markov order estimator with Dirichlet priors, viz.

$$\hat{k}_{\text{KT}}(x_1^n) = \arg\min_k(-\log p_k - \log \text{KT}_k(x_1^n))$$

fails to put out 0, the order of Q, eventually almost surely, provided that $\{p_k\}$ is slowly decreasing, that is, $\log p_k = o(k)$. We prove that, in fact,

$$\hat{k}_{\text{KT}}(x_1^n) \to \infty \quad \text{almost surely.} \tag{4.1}$$

The same inconsistency will be established also for the non-Bayesian MDL estimator defined by replacing $\text{KT}_k(x_1^n)$ in the definition of $\hat{k}_{\text{KT}}$ with the normalized maximum likelihood

$$\text{NML}_k(x_1^n) = P_{\text{ML}(k)}(x_1^n) \Big/ \sum_{y_1^n \in A^n} P_{\text{ML}(k)}(y_1^n). \tag{4.2}$$

The key to inconsistency is the observation that for k large enough it is likely that each k-block appears only once in x_1^n, which forces $-\log \text{KT}_k(x_1^n) = n\log|A|$. On the other hand $-\log Q(x_1^n) + (1/2)\log n = n\log|A| + ((|A|-1)/2)\log n$. These two facts mean that $\hat{k}_{\text{KT}} > 0$, and it only takes a bit more to establish (4.1). The formal proof is organized as the following two propositions.

PROPOSITION 5. *For $k_n = \alpha \log n$, with a sufficiently large constant α,*

$$-\log p_{k_n} - \log KT_{k_n}(x_1^n) + \log KT_0(x_1^n) \to -\infty \quad a.s.$$

PROPOSITION 6. *For every fixed $k > 0$,*

$$-\log KT_k(x_1^n) + \log KT_0(x_1^n) \to \infty \quad a.s.$$

Indeed, Proposition 5 implies that $\hat{k}_{\mathrm{KT}}(x_1^n) \neq 0$, and Proposition 6 implies that $\hat{k}_{\mathrm{KT}}(x_1^n) \neq k$, for any fixed $k > 0$, both eventually almost surely.

PROOF OF PROPOSITION 5. Note first that if no k-block occurs in x_1^{n-1} more than once then $\mathrm{KT}_k(x_1^n) = |A|^{-n}$. This follows from the representation (2.11) of $\mathrm{KT}_k(x_1^n)$ since the assumption on x_1^{n-1} implies that exactly $n-k$ different k-blocks a_1^k occur in x_1^{n-1}; to each of these there is one $a_{k+1} \in A$ with $a_1^{k+1} \in x_1^n$, so that $N(a_1^k|x_1^{n-1}) = N(a_1^{k+1}|x_1^n) = 1$.

The probability that some k-block occurs in x_1^{n-1} more than once is less than $n^2|A|^{-k}$. Indeed, for any ℓ, m with $1 \leq \ell < m \leq n-k+1$, the probability of $x_\ell^{\ell+k-1} = x_m^{m+k-1}$ is $|A|^{-k}$ because, arbitrarily fixing $x_1^{m-1} \in A^{m-1}$, for exactly one of the $|A|^k$ equiprobable choices of $x_m^{m+k-1} \in A^k$ will $x_\ell^{\ell+k-1} = x_m^{m+k-1}$ hold. In particular, for $k_n = \alpha \log n$ with $\alpha = 4/\log|A|$, the probability that some k_n-block occurs in x_1^{n-1} more than once is less than n^{-2}. This and the previous observation gives that

$$-\log \mathrm{KT}_{k_n}(x_1^n) = n \log|A| \quad \text{eventually a.s.} \tag{4.3}$$

Next we use the fact that for any fixed $k \geq 0$ there is a constant K such that,

$$|\log P_{\mathrm{ML}(k)}(x_1^n) + n \log|A|| < K \log\log n \quad \text{eventually a.s.}, \tag{4.4}$$

see Proposition A.2 in the Appendix, where now $\log Q(x_1^n) = -n \log|A|$. Since

$$\left| \log \mathrm{KT}_0(x_1^n) - \log P_{\mathrm{ML}(0)}(x_1^n) + \frac{|A|-1}{2} \log n \right| \leq C,$$

by the bound (2.12) in Lemma 3, it follows that $\log \mathrm{KT}_0(x_1^n)$ differs from $-n \log|A| - \frac{|A|-1}{2} \log n$ by less than (constant)$\log\log n$, eventually almost surely. This, together with (4.3), proves Proposition 5, because the hypothesis on $\{p_k\}$ implies that $\log p_{k_n} = o(\log n)$. □

PROOF OF PROPOSITION 6. For fixed $k > 0$, we have eventually almost surely

$$\frac{n}{2}|A|^{-k} < N(a_1^k|x_1^{n-1}) < n \quad \text{for all } a_1^k \in A^k,$$

so that, by (2.14), $-\log \mathrm{KT}_k(x_1^n)$ differs from $-P_{\mathrm{ML}(k)}(x_1^n) + \frac{|A|^k(|A|-1)}{2} \log n$ by less than a constant (depending on k). Hence, using (4.4), $-\log \mathrm{KT}_k(x_1^n)$ differs from $n \log|A| + \frac{|A|^k(|A|-1)}{2} \log n$ by less than (constant)$\log\log n$, eventually almost surely. Combining this with the previous result on $\log \mathrm{KT}_0(x_1^n)$ gives Proposition 6. □

Finally, to check the analogues of Propositions 5 and 6 for $\mathrm{NML}_k(x_1^n)$, we need a result from the theory of universal coding (see [18]), namely, for fixed $k \geq 0$ and $n \to \infty$,

$$\log \sum_{y_1^n \in A^n} P_{\mathrm{ML}(k)}(y_1^n) \sim \frac{|A|^k(|A|-1)}{2} \log n.$$

This and (4.4) immediately give the analogue of Proposition 6 for the normalized maximum likelihood $\mathrm{NML}_k(x_1^n)$ defined by (4.2). To get the analogue of Proposition 5, we additionally need an analogue of (4.3), namely with k_n as there,

$$-\log \mathrm{NML}_{k_n}(x_1^n) \leq n \log |A| \quad \text{eventually a.s.}$$

The latter follows because if no k-block occurs in x_1^{n-1} more than once then clearly $P_{\mathrm{ML}(k)}(x_1^n) = 1$, and consequently,

$$\mathrm{NML}_k(x_1^n) = 1 \Big/ \sum_{y_1^n \in A^n} P_{\mathrm{ML}(k)}(y_1^n) \ \geq |A|^{-n}.$$

APPENDIX

Here we prove, for completeness, that for a stationary process $Q \in \mathscr{M}$ of order $k_0 < k^*$,

$$k_{\mathrm{BIC}}(x_1^n) \notin [0, k_0) \cup (k_0, k^*] \quad \text{eventually a.s.}$$

Clearly, this is a consequence of the following two propositions.

PROPOSITION A.1. *In the case $k < k_0$, there is a positive constant C such that*

$$-\log P_{\mathrm{ML}(k)}(x_1^n) \geq -\log P_{\mathrm{ML}(k_0)}(x_1^n) + Cn \quad \textit{eventually a.s.}$$

PROPOSITION A.2. *For any fixed $k > k_0$, there is a positive constant C such that*

$$-\log P_{\mathrm{ML}(k_0)}(x_1^n) \leq -\log P_{\mathrm{ML}(k)}(x_1^n) + C \log\log n \quad \textit{eventually a.s.}$$

PROOF OF PROPOSITION A.1. Recall [see (2.2)] that

$$\log P_{\mathrm{ML}(k)}(x_1^n) = \sum_{a_1^{k+1} \in x_1^n} N(a_1^{k+1}|x_1^n) \log \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}.$$

The ergodic theorem implies that $N(a_1^{k+1}|x_1^n)/(n-k)$ converges almost surely to $Q(a_1^{k+1})$, for each $a_1^{k+1} \in \mathscr{S}_{k+1}$, as $n \to \infty$, and hence

$$\lim_{n\to\infty} \left(-\frac{1}{n} \log P_{\mathrm{ML}(k)}(x_1^n) \right) = -\sum_{a_1^{k+1} \in \mathscr{S}_{k+1}} Q(a_1^{k+1}) \log \frac{Q(a_1^{k+1})}{Q(a_1^k)} \quad \text{a.s.}, \tag{A.1}$$

the limit being the conditional entropy H_k of X_k, given X_1^{k-1}. (Here $Q(a_1^k)$ is replaced by 1 if $k = 0$.) It is well known (see [17], Theorem I.6.11), that H_k is strictly greater than H_{k_0} if $k < k_0$. Hence (A.1) implies the proposition. □

PROOF OF PROPOSITION A.2. Fix $k > k_0$. Since Q is a stationary process in $\mathscr{M}_k$, the product formula (2.1) yields

$$\log Q(x_1^n) = \log Q(x_1^k) + \sum_{a_1^{k+1} \in x_1^n} N(a_1^{k+1}|x_1^n) \log Q(a_{k+1}|a_1^k)$$

whenever $Q(x_1^n) > 0$. This and the expression (2.2) for $\log P_{\mathrm{ML}(k)}(x_1^n)$, yield

$$\begin{aligned} &\log P_{\mathrm{ML}(k)}(x_1^n) - \log Q(x_1^n) \\ &= -\log Q(x_1^k) - \sum_{a_1^{k+1} \in x_1^n} N\left(a_1^{k+1}|x_1^n\right) \log \frac{Q(a_{k+1}|a_1^k)}{N(a_1^{k+1}|x_1^n)/N(a_1^k|x_1^{n-1})} \end{aligned} \tag{A.2}$$

(with an obvious modification if $k = k_0 = 0$).

Recall from Proposition 3 in Section 3 that the ratio

$$\frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} = \frac{N(a_1^k|x_1^n)}{(n-k+1)Q(a_1^k)}$$

differs from 1 by less than (constant) $\sqrt{\frac{\log\log n}{n}}$, eventually almost surely. Applying this with k replaced by $k+1$, and also with n replaced by $n-1$, it follows, since $Q(a_{k+1}|a_1^k) = Q(a_1^{k+1})/Q(a_1^k)$, that the quantity

$$\varepsilon(a_1^{k+1}|x_1^n) = \frac{Q(a_{k+1}|a_1^k)}{N(a_1^{k+1}|x_1^n)/N(a_1^k|x_1^{n-1})} - 1$$

is bounded in absolute value by (constant) $\sqrt{\frac{\log\log n}{n}}$, eventually almost surely. Hence, using that $\log(1+\varepsilon) = \varepsilon + O(\varepsilon^2)$, the summation in (A.2) becomes

$$\sum_{a_1^{k+1} \in x_1^n} N\left(a_1^{k+1}|x_1^n\right) \varepsilon\left(a_1^{k+1}|x_1^n\right) + \eta_n,$$

where $|\eta_n| \le$ (constant) $\log\log n$, eventually almost surely. Eventually almost surely, however, the last summation is over all $a_1^{k+1} \in \mathscr{S}_{k+1}$, so that it takes the form

$$\sum_{a_1^{k+1} \in \mathscr{S}_{k+1}} \left[Q\left(a_{k+1}|a_1^k\right) N\left(a_1^k|x_1^{n-1}\right) - N\left(a_1^{k+1}|x_1^n\right)\right].$$

This sum is equal to 0, since summing for a_{k+1} with $a_1^k \in \mathscr{S}_k$ fixed gives zero. This completes the proof of Proposition A.2. □

REFERENCES

[1] BARRON, A. (1985). Logically smooth density estimation. Ph.D. dissertation, Dept. Electrical Engineering, Stanford Univ.

[2] BARRON, A. RISSANEN, J. and YU, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** 2743–2760.
[3] BÜHLMAN, P. and WYNER, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513.
[4] CSISZÁR, I. and KÖRNER, J. (1981). *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest.
[5] DIACONIS, P. and FREEDMAN, D. (1993). Nonparametric binary regression: a Bayesian approach. *Ann. Statist.* **21** 2108–2137.
[6] FINESSO, L. (1992). Estimation of the order of a finite Markov chain. In *Recent Advances in the Mathematical Theory of Systems, Control, and Network Signals* (H. Kimura and S. Kodama, eds.) 643–645. Mita Press.
[7] FINESSO, L., LIU, C.-C. and NARAYAN, P. (1996). The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory* **42** 1488–1497.
[8] FLAJOLET, P., KIRSCHENHOFER, P. and TICHY, R. F. (1988). Deviations from uniformity in random strings. *Probab. Theory Related Fields* **80** 139–150.
[9] HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statis. Soc. Ser. B* **41** 190–195.
[10] HAUGHTON, D. (1988). On the choice of model to fit data from an exponential family. *Ann. Statist.* **16** 342–355.
[11] KIEFFER, J. (1993). Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory* **39** 803–902.
[12] KRICHEVSKY, R. E. and TROFIMOV, V. K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **IT-27** 199–207.
[13] MARTON, K. and SHIELDS, P.(1994). Entropy and the consistent estimation of joint distributions. *Ann. Probab.* **22** 960–977. [Correction (1996). *Ann. Probab.* **24** 541–545.]
[14] PAPANGELOU, F. (1996). Large deviations and the Bayesian estimation of higher Markov transition functions. *J. Appl. Probab.* **33** 18–27.
[15] RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore.
[16] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
[17] SHIELDS, P. C. (1996). *The Ergodic Theory of Discrete Sample Paths.* Amer. Math. Soc., Providence, RI.
[18] SHTARKOV, J. (1977). Coding of discrete sources with unknown statistics. In *Topics in Information Theory* (I. Csiszár and P. Elias, eds.) 559-574. North-Holland, Amsterdam.
[19] WEINBERGER, M., RISSANEN, J. and FEDER, M. (1995). A universal finite memory source. *IEEE Trans. Inform. Theory* **41** 643–652.
[20] WILLEMS, F. M. J., SHTARKOV, Y. M. and TJALKENS, T. J. (1995). The context tree weighting method: Basic properties. *IEEE Trans. Inform. Theory* **41** 653–664.
[21] WILLEMS, F. M. J., SHTARKOV, Y. M. and TJALKENS, T. J. (1994). Technical report, Electrical Engineering Dept., Eindhoven Univ., an earlier unabridged version of [20].

A. RÉNYI INSTITUTE OF MATHEMATICS
HUNGARIAN ACADEMY OF SCIENCES
POB 127
1364 BUDAPEST
HUNGARY
E-MAIL: csiszar@math-inst.hu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606
E-MAIL: paul.shields@utoledo.edu