Prefixes and the Entropy Rate for Long-Range Sources

Ioannis Kontoyiannis
Information Systems Laboratory,
Electrical Engineering,
Stanford University.

Yurii M. Suhov
Statistical Laboratory,
Pure Math. & Math. Statistics,
Cambridge University.

Appeared as Chapter 6 in "Probability, Statistics and Optimization; A Tribute to Peter Whittle" (F.P. Kelly, editor), Wiley, June 1994.

Abstract. The asymptotic a.s.-relation

$$H = \lim_{n\to\infty} \frac{n\log n}{\sum_{i=1}^{n} L_i^n(X)}$$

is derived for any finite-valued stationary ergodic process $X = (X_n,\ n \in \mathbf{Z})$ that satisfies a Doeblin-type condition: there exists $r \geq 1$ such that $\operatorname{ess\,inf}\ P(X_{n+r} \mid \mathcal{B}_{-\infty,n}) \geq \alpha > 0$. Here H is the entropy rate of process X and $L_i^n(X)$ is the length of a shortest prefix in X which is initiated at time i and is not repeated among the prefixes initiated at times j, $1 \leq i, j \leq n$ $(i \neq j)$. Such an asymptotic was previously established by Shields in [1] for the i.i.d. processes and the irreducible aperiodic Markov chains.

1 Introduction

This work follows a paper by P. Shields [1] concerned with a problem of a relation between the entropy rate of a finite-valued stationary ergodic process and asymptotics of the lengths of shortest new prefixes in the process. This problem arises in Information Theory, within the framework of the study of the so-called Ziv-Lempel encoding algorithms. (A comprehensive introduction into Ziv-Lempel encoding may be found in [2]. A review of concepts and results about the subject is given in [3] and various related results can be found in [1], [4], [5] and [6] - see also the references therein.)

Let us fix some preliminary notation and definitions: We shall use the standard model for a source, that is, a stationary ergodic stochastic process $X = (X_n,\ n \in \mathbf{Z})$, taking values in the finite set $\mathcal{V}$ (the 'alphabet'); let H be the entropy rate of this process. Given a realization $x = (x_n)$ of the process, by $x_{a,b}$ we denote the word $(x_a, \ldots, x_b)$ taken from x (for $a \leq b$ integers). The key object in our forthcoming discussion is $L_i^n(x)$, which is defined as the length of the shortest 'prefix' in x initiated at time i that is not repeated among the prefixes initiated at times j, $1 \leq j \leq n$, $j \neq i$ ($1 \leq i \leq n$ integers):

$$L_i^n(x) = \min\,[l \geq 1 :\ x_{i,i+l-1} \neq x_{j,j+l-1} \quad \text{for any} \quad j = 1, \ldots, n, \quad j \neq i]. \tag{1}$$

This symmetric version of the Ziv-Lempel prefix was introduced by Grassberger in [7]. In the same paper it was conjectured that the asymptotic relation

$$\frac{1}{H} = \text{a.s}-\lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^n(X)}{n \log n} \tag{2}$$

holds for every stationary ergodic process X, and it was proved in [1], for the special cases where X is an i.i.d. process, an irreducible aperiodic Markov chain, or $H = 0$. Moreover, it was shown in [4] that there exist Bernoulli processes for which (2) fails even in probability (the precise definition and an extensive treatment of the Bernoulli property for stochastic processes can be found in [8]). On the other hand, the weaker relation

$$\text{a.s}-\limsup_{n\to\infty} \frac{1}{n} \sharp \left\{ i :\quad 1 \leq i \leq n,\ \left| \frac{L_i^n(X)}{\log n} - \frac{1}{H} \right| > \epsilon \right\} \leq \epsilon \tag{3}$$

was proved in [1] for any ergodic process with $H > 0$ (for any $\epsilon > 0$).

Result (2) was suggested in [7] as an entropy-estimator: Observing the output of a source (= a stationary ergodic process) X, calculating the lengths $L_i^n(X)$ of its shortest new prefixes, and forming the quantities in the r.h.s. of (2), would (almost surely) produce, for large n, accurate estimates for the entropy H of the source. The practical significance this new result lies in its applications to non-Markov sources (when the Markov property cannot be assumed or is to be tested), and to sources with long-range correlations.

In most realistic applications, like for e.g. in image processing or English text encoding, such long-range dependence *does* in fact occur. It is, therefore, important to extend (2) to processes that may not have finite dependence on their 'past', that may possess 'infinite memory'.

In this paper we extend result (2) to a wider class of processes by exploiting a connection between one-dimensional stochastic processes, and spatial processes that arise as equilibrium states in Statistical Mechanics: We introduce a new condition on the dependence of our process X on its past. It is a version of a Doeblin-type condition in Statistical Mechanics, where in the corresponding physical model, instead of looking at the 'past' of the process, we are measuring the strength of 'interactions' between distant 'particles' on a given particle configuration.

Our condition is the following: We require that there exists an integer $r \geq 1$ and a real α such that for all $y \in \mathcal{V}$ we have

$$\operatorname*{ess\,inf}_{x} \quad P\left(X_r = y \mid x_{-\infty,0}\right) \; \geq \alpha > 0, \tag{4}$$

and in Theorem 1 we prove that (4), together with ergodicity and stationarity, imply the validity of (2). The important feature of condition (4) is, of course, that it is weaker than the Markov property, in the sense that it allows our process to (possibly) have infinite memory. To be more precise, (4) says that conditionally on (almost) any semi-infinite configuration $x_{-\infty,0}$ on the past of the process, after a fixed, finite number of steps (namely r), any finite configuration $x_{r,r+n}$ of symbols from $\mathcal{V}$ may occur, with positive probability.

Theorem 2 is a more technical result: We show that (4), together with a mild 'regularity' condition on X, implies that X is a Bernoulli process.

Finally, in order to ensure that we have produced a genuine generalization, we close Section 2 with a discussion about the existence of non-Markov processes satisfying condition (4). Examples of such processes are produced by the Theory of Gibbs processes. We briefly outline the construction of such examples and we indicate a natural analogy between our construction and the actual physical model it corresponds to.

2 Statements of Results

Given a (finite or infinite) set $\mathbf{T} \subseteq \mathbf{Z}$, we denote by $\mathcal{B}_{\mathbf{T}}$ the σ-algebra generated by the collection of random variables $X_{\mathbf{T}} = (X_n,\, n \in \mathbf{T})$; likewise, by $x_{\mathbf{T}}$ we denote the restriction of a realization x to $\mathbf{T}$. The notation (s, s'), $[s, s']$, etc., is used below for intervals on the lattice $\mathbf{Z}$ (naturally s may take value $-\infty$ and s' the value ∞). In the case of the semi-infinite interval $(-\infty, 0]$ we use the simplified notation $\mathcal{B}_-$ and x_-.

Theorem 1 *Suppose that the stationary ergodic process X satisfies the following Doeblin-type condition: there exists an integer $r \geq 1$ and a real α such that for all $y \in \mathcal{V}$ we have:*

$$\operatorname*{ess\,inf}_{x} \; [\, P\,(X_r = y \mid \mathcal{B}_-)\,]\,(x_-) \geq \alpha > 0. \tag{4}$$

Then relation (2) holds.

Using Ornstein's results (see [8]), it is easy to prove that bound (4), together with a mild 'regularity' assumption on the conditional distributions of X implies that X is a Bernoulli process:

Theorem 2 *Suppose that the stationary process X satisfies (4) and the following condition:*

$$\operatorname*{ess\,sup}_{x,x':\; x_{-k,0}=x'_{-k,0}} \left| \frac{[\, P\,(X_r = y \mid \mathcal{B}_-)\,]\,(x_-)}{[\, P\,(X_r = y \mid \mathcal{B}_-)\,]\,(x'_-)} \;-\; 1 \right| \;=\; \beta_k, \tag{5a}$$

where

$$\sum_k \beta_k < \infty. \tag{5b}$$

Then X is Bernoulli.

Examples of non-Markov processes satisfying conditions (4) and (5a,b) are given by means of the theory of Gibbs processes. The gist of the approach is in replacing the one-sided conditional probabilities

$$[\, P\,(X_r = y \mid \mathcal{B}_-)\,]\,(x_-)$$

by two-side ones

$$\left[P\left(X_{[s,s']} = y_{[s,s']} \mid \mathcal{B}_{(-\infty,s-1]} \vee \mathcal{B}_{[s'+1,\infty)}\right)\right] (x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)}). \tag{6}$$

Here $\mathcal{B}^{(1)} \vee \mathcal{B}^{(2)}$ denotes the σ- algebra generated by $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$, and $x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)}$ is the realization over $\mathbf{Z} \setminus [s,s']$ obtained by glueing realizations $x_{(-\infty,s-1]}$ and $x_{[s'+1,\infty)}$ together. A natural way to obtain a compatible system of conditional probabilities (6) is to fix a function $v(l;x,x')$ of a non-negative integer variable l and variables $x,x' \in \mathcal{V}$, which takes real values and the value $-\infty$, and is symmetric in x,x' $(v(l;x,x') = v(l;x',x))$. An important requirement on v is that it tends to zero rapidly enough as $l \to \infty$ (see below). We then define

$$\left[P\left(X_{[s,s']} = y_{[s,s']} \mid \mathcal{B}_{(-\infty,s-1]} \vee \mathcal{B}_{[s'+1,\infty)}\right)\right]\left(x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)}\right)$$

$$= \frac{\exp\left[\beta W(y_{[s,s']}) + \beta W(y_{[s,s']}|x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})\right]}{\Xi_\beta(s,s';x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})}. \tag{7}$$

Here $\beta \in \mathbf{R}$ is a parameter (an inverse temperature in Statistical Mechanics) and $W(y_{[s,s']})$ and $W(y_{[s,s']}|x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})$ denote the sums

$$W(y_{[s,s']}) = \sum_{s \le j \le k \le s'} v(|j-k|;y_j,y_k) \tag{8a}$$

and

$$W(y_{[s,s']}|x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)}) = \sum_{\substack{s \le j \le s', \\ k \in \mathbf{Z} \setminus [s,s']}} v(|j-k|;y_j,x_k), \tag{8b}$$

respectively. Finally, $\Xi_\beta(s,s';x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})$ is the appropriate normalizing factor

$$\Xi_\beta(s,s';x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})$$

$$= \sum_{\widetilde{y}_{[s,s']} = (\widetilde{y}_s,\dots,\widetilde{y}_{s'})} \exp\left[\beta W(\widetilde{y}_{[s,s']}) + \beta W(\widetilde{y}_{[s,s']}|x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})\right]. \tag{9}$$

Physically speaking, the value $v(l,y,y')$ is interpreted as the 'potential energy' of a pair of 'spins' $y,y' \in \mathcal{V}$ which are at a distance $l \ge 0$ apart. Continuing this analogy, $W(y_{[s,s']})$ can be thought of as the potential energy of a finite configuration of spins $y_{[s,s']} = (y_s,\dots,y_{s'})$ which is produced by the two-body potential v, and similarly, the quantity $W(y_{[s,s']}|x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)})$ is interpreted as the energy of interaction between the finite configuration $y_{[s,s']}$ and the infinite configuration $x_{(-\infty,s-1]} \vee x_{[s'+1,\infty)}$.

Suppose that v is dominated by a positive function $\psi(l)$:

$$|v(l;\cdot,\cdot)| \le \psi(l), \quad l \ge 0,$$

where $\psi(l)$ takes, for $l \geq l_0$, finite values and is monotonically decreasing. [For $l < l_0$, we may set $v(l, y, y') = -\infty$.] Then under the assumption that ψ is decreasing sufficiently fast, that is,

$$\sum_{l \geq l_0} l\psi(l) < \infty, \tag{10}$$

there exists, for any β, a unique process X with the two-side conditional probabilities (6) determined, P_X-a.s., by formula (7), and also it satisfies conditions (4) and (5a,b). This process is called the Gibbs process corresponding to the potential v and the inverse temperature β.

For the proofs see [9], Ch. 8, Sect. 8.3, Theorem 8.39, or the original papers [10], [11], [12] and [13].

Remark. It has been proved (see [14]) that a Gibbs process exists and is unique under a condition weaker than (10), namely that the value of

$$B = \limsup_{n \to \infty} \frac{1}{\log \log n} \sum_{l=1}^{n} l\psi(l), \tag{11}$$

is finite and sufficiently small. However, it is not clear whether or not this condition guarantees the Bernoulli property.

The proof of Theorem 1 is carried out in Section 3 and the proof of Theorem 2 in Section 4.

3 Prefixes and Doeblin's condition

We start with the following lemma:

Lemma 3.1 *Under condition (4),*

$$P\left(L_i^n(X) > N\right) \leq n(1-\alpha)^{[N/r]}. \tag{12}$$

Proof. In the sequel we mark by $\Box$ the end of the proof of an intermediate assertion. Let us assume first that $r = 1$ in (4). Condition $L_i^n(x) > N$ means that the word $x_{i,i+N-1}$ is repeated among the words $x_{j,j+N-1}$, for some $1 \leq j \leq n$, $j \neq i$. Let j_0 be the least such j and assume, without loss of generality, that $i < j_0$. Consider two cases: $i + N - 1 \geq j_0$ and $i + N - 1 < j_0$. In the first case set $k = j_0 - i + 1$ and write

$$P\left(X_{i,i+N-1} = X_{j_0,j_0+N-1}\right)$$

$$= \sum_{x_{i,i+k-1}=(x_i,\ldots,x_{i+k-1})} P\left(X_{i,i+k-1} = x_{i,i+k-1}\right) \times P\left(X_{j_0,j_0+N-1} = X_{i,i+N-1} \mid X_{i,i+k-1} = x_{i,i+k-1}\right). \tag{13}$$

Observe that $k \leq N$. Writing the conditional probability in the r.h.s. of (13) as a product gives an estimate

$$P\left(X_{j_0,j_0+N-1} = X_{i,i+N-1} \mid X_{i,i+k-1} = x_{i,i+k-1}\right) \leq (1-\alpha)^N. \tag{14}$$

Substituting (14) into (13) then yields

$$P\left(X_{i,i+N-1} = X_{j_0,j_0+N-1}\right) \leq (1-\alpha)^N. \tag{15}$$

In the second case we repeat the argument with replacing k by N; this again leads to (15). Finally, bound (12) follows after counting all possibilities for j_0.

In the case of a general $r \geq 1$, replace $(1-\alpha)^N$ in the r.h.s. of (15) by $(1-\alpha)^{[N/r]}$. In fact, the inequality

$$P\left(X_{j_0,j_0+N-1} = X_{i,i+N-1} \mid X_{i,i+k-1} = x_{i,i+k-1}\right) \leq (1-\alpha)^{[N/r]} \tag{16}$$

holds because we can bound the conditional probability in the l.h.s. of (13) from above by

$$P\left(X_{j_0+tr} = X_{i+tr}, \quad t = 0, \ldots, [N/r] \mid X_{i,i+k-1} = x_{i,i+k-1}\right), \tag{17}$$

and then write (17) as a product of r-step conditional probabilities. The rest of the argument does not differ from the previous case $r = 1$. $\square$

Corollary 3.2 *There exists a constant $c > 0$ such that, with probability one,*

$$\limsup_{n\to\infty} \max_{1\leq i\leq n} \frac{L_i^n(X)}{\log n} \leq c. \tag{18}$$

Proof. We use a standard Borel-Cantelli argument. Let

$$B_{n,c} = \left\{ x : \max_{1\leq i\leq n} \frac{L_i^n(x)}{\log n} > c \right\}.$$

It suffices to prove that, for some $c > 0$,

$$\sum_n P(B_{n,c}) < \infty. \tag{19}$$

Combining (12) with the obvious inequality

$$P\left(\max_{1\leq i\leq n} L_i^n(X) > N\right) \leq n \max_{1\leq i\leq n} P\left(L_i^n(X) > N\right), \tag{20}$$

yields

$$P(B_{n,c}) \leq n^2 (1-\alpha)^{[(c/r)\log n]},$$

and (19) follows if $c > -3r/\log(1-\alpha)$. $\square$

Proof of Theorem 1. It suffices to combine result (3) and Corollary 3.2.

4 The Bernoulli property

Proof of Theorem 2. The main point now is to find a so-called very weak Bernoulli partition. A natural candidate is the partition η of the realization space into the elements $C_i(\eta)$ of the form

$$C_i(\eta) = \{x : \ x_0 = i\}, \quad i \in \mathcal{V}.$$

According to the results of Section 7, Part 1 from [8], to prove that η is weak Bernoulli it suffices to check that

$$\lim_{N\to\infty} \text{ ess sup } \sup_{A \in \mathcal{B}_{N,\infty}} [\, P(A \mid \mathcal{B}_{-}),\, P(A)\,] \;=\; 0. \tag{21}$$

It is natural to analyze higher-order Markov chains approximating our process X. The s-th Markov chain approximating X is a Markov chain with state space $\mathcal{V}^s$, and transition probability matrix $Q_s = (q_s(x_{1,s}, x'_{1,s}))$ given by

$$q_s(x_{1,s}, x'_{1,s}) = P(X_r = x'_1, \ldots, X_{rs} = x'_s \mid X_{(-s+1)r} = x_1, \ldots, X_0 = x_s),$$

where $x_{1,s} = (x_1, \ldots, x_s)$, $x'_{1,s} = (x'_1, \ldots, x'_s)$. According to condition (4), for any $s \geq 1$, this chain is irreducible and aperiodic and has a unique equilibrium distribution μ_s :

$$\mu_s \, Q_s = \mu_s.$$

Moreover, this distribution satisfies

$$\mu(x_{1,s}) \geq \alpha^s > 0.$$

The main technical step of the proof is the following:

Lemma 4.1. *There exists a constant $\gamma \in (0,1)$ such that for any positive integers s and t and any probability distribution ν on $\mathcal{V}^s$, the variation distance* var $[\, \nu Q_s^t, \mu_s \,]$ *obeys*

$$\text{var } [\, \nu Q_s^t, \mu_s \,] \leq \gamma^t. \tag{22}$$

Proof. The proof of this lemma follows a standard argument used in the theory of Gibbs processes (see the above references and also [15], [16] and [17]). Here we give a simplified version of the for the completeness of the exposition.

Writing

$$\mathcal{S}_s = \{x_{1,s} \in \mathcal{V}^s : \ \nu(x_{1,s}) > \mu_s(x_{1,s})\}$$

and

$$\mathcal{T}_s = \{x_{1,s} \in \mathcal{V}^s : \ (\nu Q_s)(x_{1,s}) > \mu_s(x_{1,s})\}$$

we have

$$\mathrm{var}\,[\,\nu Q_s, \mu_s\,] = \tfrac{1}{2}\textstyle\sum_{x_{1,s}} \Big|\, (\nu Q_s)(x_{1,s}) - \mu_s(x_{1,s})\,\Big|$$

$$= \sum_{x_{1,s}\,\in\,\mathcal{T}_s} [\,(\nu Q_s)(x_{1,s}) - \mu_s(x_{1,s})\,] = \sum_{x_{1,s}} [\,\nu(x_{1,s}) - \mu_s(x_{1,s})\,] \sum_{x'_{1,s}\,\in\,\mathcal{T}_s} q_s(x_{1,s}, x'_{1,s})$$

$$\le \sum_{x'_{1,s}\,\in\,\mathcal{S}_s} [\,\nu(x_{1,s}) - \mu_s(x_{1,s})] \left[1 - \sum_{x'_{1,s}\,\in\,\bar{\mathcal{T}}_s} q_s(x_{1,s}, x'_{1,s}) \right], \tag{23}$$

where $\bar{\mathcal{T}}_s$ denotes the complement $\mathcal{V}^s \setminus \mathcal{T}_s$.

To estimate the term in the large square brackets in the r.h.s. of (23), we use

Lemma 4.2. *Fix a (semi-infinite) realization* $x^0_+ = (x_n, n \ge 1)$ *and denote by* $\widetilde{q}_s(x'_{1,s})$ *the value of the transition probability* $q_s(x^0_{1,s}, x'_{1,s})$. *There exists a constant* $\delta \in (0,1)$ *such that for any* s *and any* $x_{1,s}, x'_{1,s} \in \mathcal{V}^s$, *we have*

$$q_s(x_{1,s}, x'_{1,s}) \ge \delta \widetilde{q}_s(x'_{1,s}). \tag{24}$$

Proof. We write

$$q_s(x_{1,s}, x'_{1,s}) = P(X_r = x'_1 \mid X_{(-s+1)r} = x_1, \dots, X_0 = x_s)$$

$$\times P(X_{2r} = x'_2 \mid X_{(-s+2)r} = x_2, \dots, X_r = x'_1) \times \dots \times \ P(X_{sr} = x'_s \mid X_0 = x_s, \dots, X_{(s-1)r} = x'_{s-1})$$

and replace, in each term, the values x_j by x^0_j, for $j = 1, 2, \dots$. Using estimates (5a,b) leads to (24). $\square$

Returning to (23), we bound the r.h.s. from above by

$$\sum_{x'_{1,s}\,\in\,\mathcal{S}_s} [\,\nu(x_{1,s}) - \mu_s(x_{1,s})] \left[1 - \delta \sum_{x'_{1,s}\,\in\,\bar{\mathcal{T}}_s} \widetilde{q}_s(x'_{1,s}) \right]. \tag{25}$$

We also can assume that

$$\sum_{x'_{1,s}\,\in\,\bar{\mathcal{T}}_s} \widetilde{q}_s(x'_{1,s}) \ \ge \frac{1}{2},$$

for otherwise we can replace in (23) $\mathcal{S}_s$ and $\widetilde{\mathcal{T}}_s$ with their complements and arrive at an expression similar to (25), with $\mathcal{T}_s$ instead of $\bar{\mathcal{T}}_s$. Therefore,

$$\text{var}\,[\nu Q_s^t,\mu_s] \le \left(1-\frac{1}{2}\delta\right) \sum_{x'_{1,s}\,\in\,\mathcal{S}_s} [\,\nu(x_{1,s})-\mu_s(x_{1,s})]$$

$$=\frac{1}{2}\left(1-\frac{1}{2}\delta\right)\sum_{x_{1,s}} \Big|\,(\nu Q_s)(x_{1,s})-\mu_s(x_{1,s})\,\Big|$$

$$=\left(1-\frac{1}{2}\delta\right)\ \text{var}\,[\,\nu,\mu_s\,].$$

Now putting $\gamma=1-1/2\delta$, the assertion of Lemma 4.1 follows by iterating the bound obtained. □

It is now easy to complete the proof of Theorem 2. Set $N=[n/r]$ and denote by $\mathcal{B}_{\{r,2r,\ldots,lr\}}$ the σ-algebra generated by the r.v.s X_r, X_{2r}, ..., X_{lr}, $1\le l\le N$. We can write

$$[\,P(A\mid\mathcal{B}_-)\,](x_-)=\textstyle\sum_{x_1,\ldots,x_N}\prod_{l=1}^N [\,P(X_{lr}=x_l\mid\mathcal{B}_{\{r,2r,\ldots,(l-1)r\}}\vee\mathcal{B}_-)\,]((x_1,\ldots,x_l)\vee x_-)$$

$$\times[\,P(A\mid\mathcal{B}_{\{r,2r,\ldots,Nr\}}\vee\mathcal{B}_-)\,]((x_1,\ldots,x_N)\vee x_-) \tag{26}$$

and, similarly,

$$P(A)=$$

$$\sum_{x_1,\ldots,x_N}\prod_{l=1}^N [\,P(X_{lr}=x_l\mid\mathcal{B}_{\{r,2r,\ldots,(l-1)r\}})\,]((x_1,\ldots,x_l)\times[\,P(A\mid\mathcal{B}_{\{r,2r,\ldots,(l-1)r\}})\,](x_1,\ldots,x_N). \tag{27}$$

The idea is now to 'cut off' the conditional probabilities passing to the higher-order Markov chains approximating the process X. Let us denote the probability distribution on the space of the realizations $x=(x_n,n\in\mathbf{Z})$, induced by the s-th order chain, by $\widetilde{P}^{(s)}$. We can of course write down formulas similar to (26), (27) for $\widetilde{P}^{(s)}(A\mid\mathcal{B}_-)\,](x_-)$ and $\widetilde{P}^{(s)}(A)$. Comparing corresponding factors then yields

$$\left|\,\frac{[\,P(A\mid\mathcal{B}_-)\,](x_-)}{[\,\widetilde{P}^{(s)}(A\mid\mathcal{B}_-)\,](x_-)}-1\,\right|\le\frac{N}{s}\sum_{k\ge s}\beta_k \tag{28}$$

and similarly

$$\left|\,\frac{P(A)}{\widetilde{P}^{(s)}(A)}-1\,\right|\le\frac{N}{s}\sum_{k\ge s}\beta_k. \tag{29}$$

Applying Lemma 4.1 now yields

$$\left|\,\frac{[\,\widetilde{P}^{(s)}(A\mid\mathcal{B}_-)\,](x_-)}{\widetilde{P}^{(s)}(A)}-1\,\right|\le\gamma^{[(N-s)/s]}, \tag{30}$$

so that choosing $s = s(N)$ in such a way that both $[(N-s)/s] \to \infty$ and $(N/s)\sum_{k\geq s}\beta_k \to 0$, as $N \to \infty$, leads to (21). This completes the proof of Theorem 2.

References

[1] P. Shields. Entropy and prefixes. *Annals of Probability*, 20:403–409, 1992.

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J.Wiley: New York, 1991.

[3] A. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inf. Theory*, 35:1250–1258, 1989.

[4] P. Shields. String-matching: the ergodic case. *Annals of Probability*, 20:1199–1203, 1992.

[5] D. Ornstein and B. Weiss. How sampling reveals a process. *Annals of Probability*, 18:905–930, 1990.

[6] D. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Trans. Inf. Theory*, 39:78–83, 1993.

[7] P. Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Trans. Inf. Theory*, 35:669–675, 1989.

[8] D. Ornstein. *Ergodic Theory, Randomness and Dynamical Systems.* Yale University Press, 1974.

[9] H.-O. Georgii. *Gibbs Measures and Phase Transitions.* W. de Gruyter: Berlin et al, 1989.

[10] R.L. Dobrushin. The problem of uniqueness of a Gibbs random field and the problem of phase transition. *Funct. Anal. Appl.*, 2:302–312, 1968.

[11] D. Ruelle. Statistical mechanics of a one-dimensional lattice gas. *Math. Phys.*, 9:267–278, 1968.

[12] G. Gallavotti and S. Miracle-Sole. Absence of phase transitions in hard-core one dimensional systems with long-range interactions. *Journ. Math. Phys.*, 11:147–154, 1970.

[13] Ju.M. Suhov. The matrix method for continuous systems in classical statistical mechanics. *Trans. Moscow Math. Soc.*, 24:185–212, 1971.

[14] R.A. Minlos and G.M. Natapov. Uniqueness of the Gibbs distribution in one-dimansional classical systems. *Theor. Math. Phys.*, 24:697–703, 1975.

[15] G. Gallavotti. Ising model and bernoulli schemes in one dimension. *Commun. Math. Phys.*, 32:183–190, 1973.

[16] G. Caldiera and E. Presutti. Absence of phase transitions in hard-core one dimensional systems with long-range interactions. *Commun. Math. Phys.*, 35:279–286, 1973.

[17] Yu.M. Suhov. Random point processes and DLR equations. *Commun. Math. Phys.*, 50:112–132, 1976.