Institute of Physics Publishing Nonlinearity

Nonlinearity **18** (2005) 2545–2563 doi:10.1088/0951-7715/18/6/007

Large deviations for empirical entropies of g-measures

J-R Chazottes[1] **and D Gabrielli**[2]

[1] Centre de Physique Théorique, CNRS-Ecole Polytechnique, F-91128 Palaiseau, Cedex, France
[2] Dipartimento di Matematica Pura e Applicata, Università Dell'Aquila, Via Vetoio Loc. Coppito, 67100 L'Aquila, Italy

E-mail: jeanrene@cpht.polytechnique.fr and gabriell@univaq.it

Received 8 November 2004, in final form 20 June 2005
Published 31 August 2005
Online at stacks.iop.org/Non/18/2545

Recommended by C Liverani

Abstract

The entropy of an ergodic finite-alphabet process can be computed from a single typical sample path x_1^n using the entropy of the k-block empirical probability and letting k grow with n roughly like $\log n$. We further assume that the distribution of the process is a g-measure. We prove large deviation principles for conditional, non-conditional and relative $k(n)$-block empirical entropies.

Mathematics Subject Classification: 37D35, 60F10

1. Introduction

A problem of interest is the entropy-estimation problem. Given a sample path $x_1, x_2, \ldots, x_n$ (where the x_i's are drawn from a finite alphabet A) *typical* for an unknown ergodic source, the question is: how do we estimate its entropy? The simplest idea is to use a 'plug-in' estimator. First one computes for *each* block of length k, the k-marginals of the source as the limit, when $n \to \infty$, of the *k-block empirical probability* of the sample x_1^n; then one can compute the k-block entropy of the source and let $k \to \infty$ to get the entropy of the source. A natural question is thus: how is it possible to choose $k = k(n)$ to do these two steps at the same time? Ornstein and Weiss [21] (see also [23]) proved that this is indeed possible for any ergodic source of positive entropy if k does not grow 'too fast' with n, loosely like $\log n$. The proof is based on an 'empirical version' of the Shannon–McMillan–Breiman theorem.

A first result on fluctuations of $k(n)$-block empirical entropies, refining Ornstein–Weiss' almost-sure result, was obtained in [16]. In that paper the authors consider chains of infinite order which lose memory exponentially fast. Under additional restrictions on the sequence $k(n)$ they prove a central limit theorem for the conditional $k(n)$-block empirical entropy, and they prove also that the rescaled $k(n)$-block empirical entropy cannot have Gaussian fluctuations.

In this paper, we are interested in large deviations for $k(n)$-block empirical entropies. To this end we assume that the distribution of the process generating the sample path x_1^n is a g-measure for the potential $\phi = \log g$ (see below for definitions and references). Such a process can be viewed as (a special case of) a chain with complete connections or a chain of infinite order, see, e.g. [13, 14]. Another way, especially useful for our concern, to characterize and describe a g-measure is as a one-dimensional equilibrium state [15, 20].

In this setting, we prove large deviation principles for conditional, non-conditional and relative entropies of the $k(n)$-block empirical probability of the sample path x_1^n *when* $k(n)$ *grows, roughly speaking, like* $\log n$. This is done for *any* g-measure.

When the block length k is fixed, it is easy to obtain a large deviation principle for k-block empirical entropies by 'contraction' of the large deviation principle for the empirical process [6]. This is possible because k-block entropies are continuous in the weak topology. To prove the result when $k(n)$ grows with n we will generalize some classical combinatorial techniques. We will use the *combinatorics of types* to see 'how fast we can let k grow with n' and get a condition close to the Ornstein–Weiss one.

The rate functions we obtain are convex and we will also compute their Legendre transform which coincides with the corresponding scaled cumulant generating functions. This will allow us to derive some properties of the rate functions and an explicit representation in some cases.

Let us note that the rate function we obtain for conditional and rescaled non-conditional empirical entropy can have a linear part. This unexpected feature is related to the entropy of zero-temperature limit of equilibrium states which can be in general non-zero.

Let us briefly mention that around the problem of entropy estimation other techniques and ideas have been developed. The 'plug-in' estimator is only one among several entropy estimators, see, e.g. [7, 9, 22, 23]. We point out that we could have worked in the context of one-dimensional Gibbs measures. An interesting issue is the case of multi-dimensional Gibbs measures since we can no longer use the combinatorics of types.

This paper is organized as follows. In the next section we record preliminary definitions and notions, in particular, on g-measures and the various entropies under study. In section 3 we present our main results. In section 4 we discuss our results, in particular, the form of the rate functions that we obtain for empirical entropies. Section 5 is devoted to the collection of combinatorial tools needed to understand 'how fast k can grow with n' later on. Section 6 contains the proof of the main results.

2. Preliminary definitions and notions

Let A be a finite alphabet. We will denote by $a_1^\infty \stackrel{\text{def}}{=} (a_1, a_2, \ldots)$ the elements of $A^{\mathbb{N}}$ and by a_1^k the finite string $(a_1, \ldots, a_k)$. We will use the notation x_1^n for a 'sample path' $(x_1, x_2, \ldots, x_n)$, $x_i \in A$. We denote by T the 'shift' operator defined as $Tx_1^\infty = x_2^\infty$. The cylinder set $[a_1^n]$ is the set of infinite strings b_1^∞ drawn from $A^{\mathbb{N}}$ such that $b_1^n = a_1^n$.

We call $\mathcal{M}^k$ the set of probability measures ν_k on A^k and $\mathcal{M}_s^k$ the set of probability measures ν_k on A^k which satisfy the following stationarity condition

$$\sum_{b\in A} \nu_k(a_1^{k-1}b) = \sum_{b\in A} \nu_k(ba_1^{k-1}) \qquad \forall a_1^{k-1} \in A^{k-1}. \tag{2.1}$$

The subset $\mathcal{M}_s^k$ is convex and $\mathcal{E}^k$ denotes the set of its extremal elements.

We call $\mathcal{M}$ the set of probability measures ν on $A^{\mathbb{N}}$ with the usual sigma-algebra of cylinders. The subset of shift-invariant (or stationary) measures is denoted by $\mathcal{M}_s$. The set of ergodic measures (the extremal points of $\mathcal{M}_s$) is denoted by $\mathcal{E}$.

Given a measure $\nu \in \mathcal{M}_s$ we will write ν_k for its k-marginals. Of course we have the identity $\nu_k(a_1^k) = \nu([a_1^k])$ for any $a_1^k \in A^k$ and consequently $\nu_k \in \mathcal{M}_s^k$.

2.1. g-measures and equilibrium states

In this paper we deal with g-measures associated to continuous and regular g-functions. We refer the reader to [19,20,26] for full details about the following material.

Let g be a continuous function on $A^{\mathbb{N}}$ satisfying

$$\sum_{b_1^\infty : Tb_1^\infty = a_1^\infty} g(b_1^\infty) = 1 \qquad \text{for all } a_1^\infty \in A^{\mathbb{N}}. \tag{2.2}$$

We further assume that g is strictly positive (this implies $g < 1$ by (2.2)). We associate to such a function a potential, normalized according to (2.2), by setting

$$\phi \overset{\text{def}}{=} \log g. \tag{2.3}$$

Observe that $\phi < 0$. A g-measure can be defined as an equilibrium state for the potential ϕ. We measure the continuity of ϕ by the sequence of its variations $(\mathrm{var}_m(\phi))_{m\in\mathbb{N}}$:

$$\mathrm{var}_m(\phi) \overset{\text{def}}{=} \sup\{|\phi(a_1^\infty) - \phi(b_1^\infty)| : a_1^m = b_1^m\}. \tag{2.4}$$

Notice that (uniform) continuity of ϕ (with respect to the canonical distance metrizing product topology) is equivalent to $\mathrm{var}_m(\phi) \to 0$ as $m \to \infty$.

It is well known that if $\mathrm{var}_m(\phi)$ decreases to 0 fast enough, then there is a unique g-measure which is the unique equilibrium state for ϕ. For instance, if this decreasing is exponential [1] or more generally summable [26]. On the other hand, an example of non-uniqueness was given by Bramson and Kalikow [4]. In that example, $\mathrm{var}_m(\phi) \geqslant (C/\log m)$. Very recently the authors of [2] showed that square-summability of variations, ensuring uniqueness [18], is tight. Let us mention a uniqueness criterion based on a 'one-sided' Dobrushin condition involving oscillations of the potential instead of variations [14].

From now on, we fix one of the g-measures associated to ϕ and denote it by ρ. For all $n \geqslant 1$ and $a_1^\infty \in A^{\mathbb{N}}$, we have the following property

$$\mathrm{e}^{-n\varepsilon_n} \leqslant \frac{\rho([a_1^n])}{\exp\left(\sum_{j=1}^{n-1} \phi(a_j^\infty)\right)} \leqslant \mathrm{e}^{n\varepsilon_n}, \tag{2.5}$$

where $(\varepsilon_n)_n$ is a sequence of non-negative real numbers decreasing to 0.

For $k \geqslant 2$, let $\rho^{(k)}$ be the $(k-1)$-step Markov approximation of ρ, that is, the (unique) equilibrium state of the cylindrical potential

$$\phi_k(a_1^\infty) = \phi_k(a_1^k) \overset{\text{def}}{=} \log \frac{\rho([a_1^k])}{\rho([a_2^k])}.$$

When $k = 1$, $\rho^{(1)}$ is the Bernoulli measure for the potential $\phi_1(a_1^\infty) = \phi_1(a_1) \overset{\text{def}}{=} \log \rho(a_1)$. We can see ϕ_k also as a function on A^k.

We have the following property

$$\|\phi - \phi_k\|_\infty \leqslant \mathrm{var}_k(\phi). \tag{2.6}$$

This implies the statement that for all $a_1^\infty \in A^{\mathbb{N}}$

$$\lim_{k\to\infty} \log \frac{\rho([a_1^k])}{\rho([a_2^k])} = \phi(a_1^\infty)$$

uniformly.

We shall use the variational principle repeatedly. Let $\psi : A^{\mathbb{N}} \to \mathbb{R}$ be a continuous function. Then

$$\sup\{\mathbb{E}_\eta[\psi] + h(\eta) : \eta \in \mathcal{M}_s\} = P_{\text{top}}(\psi). \tag{2.7}$$

Moreover, the supremum is attained if and only if η is an equilibrium state of ψ. $P_{\mathrm{top}}(\psi)$ is the topological pressure of ψ. It is defined as

$$P_{\mathrm{top}}(\psi) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{a_1^n} \exp\left(\sup \left\{ \sum_{j=1}^{n} \psi(b_j^\infty) : b \in [a_1^n] \right\} \right). \tag{2.8}$$

Coming back to a normalized potential $\phi = \log g$, we have $P_{\mathrm{top}}(\phi) = 0$. This can be seen, for instance, by plugging (2.5) in (2.8). The variational principle then tells us that

$$h(\rho) = -\mathbb{E}_\rho[\phi]. \tag{2.9}$$

In particular, the entropy of a g-measure is always strictly positive.

We shall also consider multiples of the potential ϕ, that is, potentials of the form $\beta\phi$, $\beta \in \mathbb{R}$. When $\beta \neq 1$, such potentials have no reason to be normalized as ϕ is, i.e. the corresponding equilibrium states are not g-measures. But this does not matter for us in the sense that we will only deal with equilibrium states of $\beta\phi$ which we will indicate with $\rho_{\beta\phi}$.

Remark. A g-measure is also called a chain of infinite order or a chain with complete connections, see, e.g. [13, 14] for recent accounts. See also [17]. In probabilistic terms, a chain of infinite order, or a chain with complete connections, is a process characterized by transition probabilities that depend on the whole past in a continuous manner. A g-measure can also be interpreted as a one-dimensional Gibbs measure if the variations go to 0 exponentially fast [15].

2.2. Entropies

The k-block ($k \geqslant 1$) Shannon entropy is defined as

$$H_k(\nu) \overset{\mathrm{def}}{=} -\sum_{a_1^k} \nu([a_1^k]) \log \nu([a_1^k]) = H_k(\nu_k) \overset{\mathrm{def}}{=} -\sum_{a_1^k} \nu_k(a_1^k) \log \nu_k(a_1^k).$$

The conditional k-block ($k \geqslant 2$) entropy is defined as

$$h_k(\nu) \overset{\mathrm{def}}{=} -\sum_{a_1^k} \nu([a_1^k]) \log \frac{\nu([a_1^k])}{\nu([a_1^{k-1}])} = h_k(\nu_k) \overset{\mathrm{def}}{=} -\sum_{a_1^k} \nu_k(a_1^k) \log \nu_k(a_k|a_1^{k-1}),$$

where $\nu_k(a_k|a_1^{k-1})$ is the conditional probability $\nu_k(a_1^k)/\sum_b \nu_k(a_1^{k-1}b)$. We have the relation

$$h_k(\nu) = H_k(\nu) - H_{k-1}(\nu), \qquad k \geqslant 1,$$

where by convention we set $H_0(\nu) \overset{\mathrm{def}}{=} 0$. Hence $h_1(\nu) \overset{\mathrm{def}}{=} H_1(\nu)$. Note that $h_k(\cdot)$ is a concave function on $\mathcal{M}^k$.

It is well known that if ν is a stationary measure, then

$$\lim_{k\to\infty} h_k(\nu) = \lim_{k\to\infty} \frac{H_k(\nu)}{k} = h(\nu),$$

where $h(\nu)$ is the (Shannon–Kolmogorov–Sinai) entropy of ν.

The k-block ($k \geqslant 1$) relative entropy of a stationary measure ν with respect to a g-measure ρ is defined as

$$D_k(\nu|\rho) \overset{\mathrm{def}}{=} \sum_{a_1^k} \nu([a_1^k]) \log \frac{\nu([a_1^k])}{\rho([a_1^k])} = D_k(\nu_k|\rho_k) \overset{\mathrm{def}}{=} \sum_{a_1^k} \nu_k(a_1^k) \log \frac{\nu_k(a_1^k)}{\rho_k(a_1^k)} \cdot$$

The map $D_k(\cdot|\rho_k)$ is convex on $\mathcal{M}^k$. The conditional k-block ($k \geqslant 1$) relative entropy is defined as

$$\Delta_k(\nu|\rho) \overset{\text{def}}{=} D_k(\nu|\rho) - D_{k-1}(\nu|\rho) = \Delta_k(\nu_k|\rho_k) \overset{\text{def}}{=} D_k(\nu_k|\rho_k) - D_{k-1}(\nu_{k-1}|\rho_{k-1}),$$

where we set $D_0(\nu|\rho) \overset{\text{def}}{=} 0$. This imposes $\Delta_1(\nu|\rho) \overset{\text{def}}{=} D_1(\nu|\rho)$.

The relative entropy $h(\nu|\rho)$ between $\nu \in \mathcal{M}_s$ and a g-measure ρ is defined as

$$h(\nu|\rho) \overset{\text{def}}{=} \lim_{k\to\infty} \frac{1}{k} D_k(\nu|\rho) = \lim_{k\to\infty} \Delta_k(\nu|\rho) \qquad \text{and} \qquad h(\nu|\rho) = -\mathbb{E}_\nu[\phi] - h(\nu). \tag{2.10}$$

By the variational principle, it is obvious that $h(\nu|\rho) = 0$ if, and only if, ν is an equilibrium state of ϕ (see [6] for more details).

2.3. *Empirical measures and entropies*

Given a finite string (a 'sample path') x_1^n we define the empirical measures

$$\pi_k(a_1^k; x_1^n) = \pi_{k,n}(a_1^k) \overset{\text{def}}{=} \frac{\sum_{i=1}^n \mathbb{I}(\tilde{x}_i^{i+k-1} = a_1^k)}{n}, \qquad k \in \mathbb{N},$$

where $\tilde{x}_1^\infty \in A^{\mathbb{N}}$ is the periodic, with period n, sample path $(x_1^n x_1^n x_1^n \cdots)$.

It is easy to see that $\pi_k(\cdot; x_1^n) \in \mathcal{M}_s^k$. The family of probability measures $(\pi_k(\cdot; x_1^n))_{k\in\mathbb{N}}$ is consistent in the sense that

$$\sum_{a_j} \pi_j(a_1^j; x_1^n) = \pi_{j-1}(a_1^{j-1}; x_1^n), \qquad j \in \mathbb{N}$$

and these are the marginals of the empirical process $\pi(\cdot; x_1^n)$ defined as

$$\pi(S; x_1^n) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{T^i \tilde{x}_1^\infty}(S), \tag{2.11}$$

where S is any measurable subset of $A^{\mathbb{N}}$.

We can now define the following plug-in estimators for entropies.

Definition 2.1. *Let $x_1^n \in A^n$ be a sample path. The k-block empirical entropy is defined as*

$$\hat{H}_k(x_1^n) \overset{\text{def}}{=} H_k(\pi_k(\cdot; x_1^n)).$$

The conditional k-block empirical entropy is defined as

$$\hat{h}_k(x_1^n) \overset{\text{def}}{=} h_k(\pi_k(\cdot; x_1^n)).$$

The relative k-block empirical entropy with respect to a measure ρ is defined as

$$\hat{D}_k(x_1^n|\rho) \overset{\text{def}}{=} D_k(\pi_k(\cdot; x_1^n)|\rho_k).$$

The relative conditional k-block empirical entropy with respect to a measure ρ is defined as

$$\hat{\Delta}_k(x_1^n|\rho) \overset{\text{def}}{=} \Delta_k(\pi_k(\cdot; x_1^n)|\rho_k).$$

3. Main results

We are now ready to state the main results of this paper.

Theorem 3.1 (large deviation principles for empirical entropies). *Let x_1^n be a sample path distributed according to a g-measure ρ. Assume that $(k(n))_{n\in\mathbb{N}}$ diverges and eventually satisfies*

$$k(n) \leqslant \frac{1-\varepsilon}{\log|A|}\log n \tag{3.1}$$

for some $0 < \varepsilon < 1$. Then the conditional empirical entropy $\hat{h}_{k(n)}(x_1^n)$ satisfies the following large deviation principle:

for any closed set $C \subset \mathbb{R}$

$$\limsup_{n\to\infty}\frac{1}{n}\log\rho\{x_1^n : \hat{h}_{k(n)}(x_1^n) \in C\} \leqslant -\inf\{\mathbf{I}(u) : u \in C\},$$

for any open set $O \subset \mathbb{R}$

$$\liminf_{n\to\infty}\frac{1}{n}\log\rho\{x_1^n : \hat{h}_{k(n)}(x_1^n) \in O\} \geqslant -\inf\{\mathbf{I}(u) : u \in O\},$$

where the convex rate function $\mathbf{I}$ is defined as

$$\mathbf{I}(u) = \begin{cases}\inf\{h(\nu|\rho) : \nu \in \mathcal{M}_s : h(\nu) = u\} & u \in [0, \log|A|],\\ +\infty & \text{otherwise.}\end{cases} \tag{3.2}$$

The same large deviation principle holds if we replace $\hat{h}_{k(n)}(x_1^n)$ by the rescaled empirical entropy $\hat{H}_{k(n)}(x_1^n)/k(n)$.

Theorem 3.2 (large deviations for empirical relative entropies). *Let x_1^n be a sample path distributed according to a g-measure ρ. Suppose that $(k(n))_{n\in\mathbb{N}}$ diverges and eventually satisfies $k(n) \leqslant ((1-\varepsilon)/\log|A|)\log n$, for some $0 < \epsilon < 1$. Then the empirical relative entropies $\hat{\Delta}_{k(n)}(x_1^n|\rho)$ and $(1/k(n))\hat{D}_{k(n)}(x_1^n|\rho)$ satisfy a large deviation principle as in theorem 3.1 but with the rate function*

$$\mathbf{J}(u) = \begin{cases}u & u \in [0, -\inf\{\mathbb{E}_\eta[\phi] : \eta \in \mathcal{E}\}],\\ +\infty & \text{otherwise.}\end{cases} \tag{3.3}$$

These theorems are proved in section 6. Their proof relies in an essential way upon combinatorial properties of types and a continuity property of entropy which are established in section 5.

The following proposition deals with the case of fixed block length. The preceding theorems extend this proposition to the case when $k(n)$ is allowed to grow with n according to (3.1).

Proposition 3.3 (large deviations for fixed block length). *Let x_1^n be a sample path distributed according to a g-measure ρ. Then, for each $k \geqslant 1$, the empirical entropies $(1/k)\hat{H}_k(x_1^n)$, $\hat{h}_k(x_1^n)$, $(1/k)\hat{D}_k(x_1^n)$ and $\hat{\Delta}_k(x_1^n)$ satisfy an LDP with normalizing factor $1/n$ and rate functions, respectively, given by*

$$\mathbf{I}_k^H(u) = \inf\{h(\nu|\rho) : H_k(\nu)/k = u\}, \qquad \mathbf{I}_k^h(u) = \inf\{h(\nu|\rho) : h_k(\nu) = u\},$$

$$\mathbf{I}_k^D(u) = \inf\{h(\nu|\rho) : D(\nu_k|\rho_k)/k = u\}, \qquad \mathbf{I}_k^\Delta(u) = \inf\{h(\nu|\rho) : \Delta(\nu_k|\rho_k) = u\},$$

where the infima are taken over $\nu \in \mathcal{M}_s$. The infimum over an empty set is taken equal to $+\infty$ following the usual convention.

This proposition is a direct consequence of the contraction principle and suggests that the rate functions we can expect when we consider $k(n)$ growing with n are 'contracted' relative entropies. Note that the rate functions of proposition 3.3 need not be convex.

From the convexity of $\mathbf{I}$ and $\mathbf{J}$ we know that they are in Legendre duality with the corresponding scaled cumulant generating function for the different empirical entropies. In the next two propositions we give the expression of the scaled cumulant generating function for empirical entropies and empirical relative entropies.

Proposition 3.4. *Assume that the hypotheses of theorem 3.1 hold. Then the rate function* $\mathbf{I}$ *is in Legendre duality with the convex function* $t \mapsto \mathbf{R}(t)$, $t \in \mathbb{R}$, *defined as*

$$\mathbf{R}(t) = \begin{cases} (t+1)P_{\mathrm{top}}(\phi/(t+1)) & \text{for } t > -1, \\ \sup\{\mathbb{E}_\eta[\phi] : \eta \in \mathcal{E}\} & \text{for } t \leqslant -1. \end{cases} \tag{3.4}$$

Moreover,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\rho[\mathrm{e}^{nt\hat{h}_{k(n)}(x_1^n)}] = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\rho[\mathrm{e}^{nt(\hat{H}_{k(n)}(x_1^n)/k(n))}] = \mathbf{R}(t). \tag{3.5}$$

Using (2.5) it is easy to check that

$$\mathbf{R}(t) = (t+1) \lim_{n\to\infty} \frac{1}{n} \log \sum_{a_1^n \in A^n} \rho([a_1^n])^{\frac{1}{t+1}} \qquad \text{for } t > -1. \tag{3.6}$$

This resembles a Rényi entropy.

Proposition 3.5. *Assume that the hypotheses of theorem 3.2 hold. Then the rate function* $\mathbf{J}$ *is in Legendre duality with the convex function* $t \mapsto \mathbf{P}^\Delta(t)$, $t \in \mathbb{R}$, *defined as*

$$\mathbf{P}^\Delta(t) \stackrel{\text{def}}{=} \begin{cases} (1-t)\inf\{\mathbb{E}_\nu[\phi] : \nu \in \mathcal{E}\} & t > 1, \\ 0 & t \leqslant 1. \end{cases}$$

Moreover,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\rho[\mathrm{e}^{nt\hat{\Delta}_{k(n)}(x_1^n|\rho)}] = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\rho[\mathrm{e}^{nt(\hat{D}_{k(n)}(x_1^n|\rho)/k(n))}] = \mathbf{P}^\Delta(t). \tag{3.7}$$

Let us introduce

$$h_\infty \stackrel{\text{def}}{=} \lim_{\beta\to\infty} h(\rho_{\beta\phi}). \tag{3.8}$$

The existence of this limit will be shown below (lemma 6.1). In general h_∞ can be strictly positive and we stress that it is equal to $\log|A|$ for the uniform Bernoulli measure.

In the case when $\mathbf{R}$ is a strictly convex, continuously differentiable function on $]-1, +\infty[$, we can improve the results of theorem 3.1. A large class of g-measures satisfies this property, namely those associated to potentials with square summable variations.

Proposition 3.6 (more on large deviations). *In addition to assumptions of theorem 3.1, assume that the variations of* ϕ *are square summable. Then* $\mathbf{I}$ *is strictly convex on* $[h_\infty, \log|A|]$, *with a unique minimum, where it assumes the value* 0, *at* $u = h(\rho)$. *Moreover it admits the following representation:*

$$\mathbf{I}(u) = h(\rho_{\beta_u\phi}|\rho_\phi) \qquad \text{for } u \in [h_\infty, \log|A|], \tag{3.9}$$

where $\beta_u \geqslant 0$ *is the unique solution of the equation* $h(\rho_{\beta\phi}) = u$. *On the interval* $[0, h_\infty]$ *the function* $\mathbf{I}$ *is linear*

$$\mathbf{I}(u) = -u - \sup\{\mathbb{E}_\eta[\phi] : \eta \in \mathcal{E}\}.$$

4. Comments on the results

We make some comments on the above results.

Zero-temperature limit and non-differentiability of **R** *at* -1. By using a classical formula for the derivative of the pressure [19], it is straightforward to see that the right derivative of $t \mapsto \mathbf{R}(t)$ at -1, when the variations of ϕ are square summable, is equal to

$$\lim_{\beta\to+\infty} (P_{\text{top}}(\beta\phi) - \beta\mathbb{E}_{\rho_\beta}[\phi]),$$

where we recall that $\rho_{\beta\phi}$ is the equilibrium state of the potential $\beta\phi$. By the variational principle, we thus get

$$\lim_{t\downarrow-1} \frac{\mathrm{d}\mathbf{R}(t)}{\mathrm{d}t} = \lim_{\beta\to+\infty} h(\rho_{\beta\phi}) = h_\infty.$$

This limit is not zero, in general, therefore the function **R** is not differentiable at $t = -1$. Notice that this is related to zero-temperature limit of equilibrium states.

About the route to large deviations. Let us emphasize that we prove our large deviation bounds directly. Another way to prove large deviation principles is to first prove the existence of the corresponding scaled cumulant generating function and then to apply the Gärtner–Ellis theorem (see, e.g. [11]). To that end one needs to prove, e.g. that the scaled cumulant generating function is differentiable and strictly convex. We could do that under the assumption that the potential of the g-measure has square-summable variations. But, as (3.4) shows, the scaled cumulant generating function is not differentiable at -1 in the case $h_\infty \neq 0$. Therefore one cannot apply the Gärtner–Ellis theorem. Notice also that the rate functionals of proposition 3.3 can be in general non-convex. This means that even in the case when k is fixed the Gärtner–Ellis theorem may not apply.

We want to stress that with our approach we need not assume anything on the rate of convergence to zero of the variations of the potential.

On the growth condition (3.1). A look at the proof of theorem 3.1 reveals that we actually have a slightly more general condition on $k(n)$. In fact we could impose, e.g.

$$\frac{(\log n)^2 |A|^{k(n)}}{n} \to 0 \qquad \text{as } n \to \infty.$$

We feel that condition (3.1) is more appealing and it is related to the condition which appears in the laws of large numbers for empirical entropies (see below).

Flatness of **I**. If ρ is not the unique equilibrium state of ϕ, it is easy to see that the rate function **I** can be identically zero in some interval containing $h(\rho)$. Indeed, the set of equilibrium states of ϕ forms a Choquet simplex, and the map $\nu \mapsto h(\nu)$ is convex affine [19] on the set of shift-invariant measures. Hence, there is an equilibrium state ρ_1 (maybe equal to ρ) such that $h(\rho_1)$ minimizes the entropy among all equilibrium states of ϕ. It may not be unique but this does not matter: we call this minimal entropy h_1. We do the same for the maximal entropy and call the corresponding value (maybe equal to $h(\rho)$) h_2. Then, it is easy to verify that $\mathbf{I}(u) = 0$ for all $u \in [h_1, h_2]$ since $\mathbf{I}(h_1) = \mathbf{I}(h_2) = 0$ (by the variational principle), and **I** is convex and positive.

Strong laws of large numbers for empirical entropies. If ρ is the unique equilibrium state of ϕ (e.g. when ϕ has square-summable variations), then 0 is the minimum of **I** and it is attained only at $u = h(\rho)$ (this is an immediate consequence of the variational principle). We can use theorem 3.1 and apply the Borel–Cantelli lemma to obtain

$$\lim_{n\to+\infty} \frac{1}{k(n)} \hat{H}_{k(n)}(x_1^n) = \lim_{n\to+\infty} \hat{h}_{k(n)}(x_1^n) = h(\rho) \qquad \rho\text{-a.s.}$$

Therefore, we recover in our context the almost-sure result obtained by Ornstein and Weiss [21, 23] that we mentioned in the introduction. Note that our $k(n)$ is allowed to grow a little less fast and that we make much stronger hypotheses on the source ρ. A similar statement, in probability, can be deduced from the results of [16]. The almost-sure convergence of conditional empirical entropy in the case of an ergodic measure ν with positive entropy can be proved under the condition that $k(n) \leqslant ((1-\varepsilon)/h(\nu))\log n$ (and $k(n) \to \infty$), for some $0 < \epsilon < 1$. If $\epsilon = 0$, this almost-sure convergence fails in general [24].

The same argument when applied to the statement of theorem 3.2 leads to the almost-sure convergence of empirical relative entropies to zero

$$\lim_{n\to\infty} \hat{\Delta}_{k(n)}(x_1^n|\rho) = \lim_{n\to\infty} \frac{1}{k(n)} \hat{D}_{k(n)}(x_1^n|\rho) = 0 \qquad \rho\text{-a.s.}$$

A similar result in probability for $\sqrt{n}\hat{D}_{k(n)}(x_1^n|\rho)$ appears in [16] with more assumptions on $k(n)$.

Connection with central limit asymptotics. Theorem 3.2 has its own interest, but it is also connected with the central limit asymptotics of conditional empirical entropy [16] as follows. The following decomposition holds (see [16]):

$$\hat{h}_{k(n)}(x_1^n) - h(\rho) = -\frac{1}{n}\sum_{j=0}^{n-1}(\phi(T^j x_1^\infty) - \mathbb{E}_\rho[\phi]) - \hat{\Delta}_{k(n)}(x_1^n|\rho) + \mathcal{C}_n, \tag{4.1}$$

where the correction term $\mathcal{C}_n$ is such that $|\mathcal{C}_n| \leqslant C\, \mathrm{var}_{k(n)}(\phi)$ and $x_1^\infty \in [x_1^n]$. In words, the conditional empirical entropy is equal to the empirical average of the potential $-\phi$ plus a term due to the conditional empirical relative entropy between the empirical measure and the 'true' measure plus a correction.

In [16], the authors assume that the variations of ϕ decrease exponentially fast. They show, under appropriate assumptions on the way $k(n)$ is allowed to grow, that $\sqrt{n}\hat{\Delta}_{k(n)}(x_1^n|\rho)$ goes to zero in ρ-probability, as well as $\sqrt{n}\mathcal{C}_n$. Therefore, they can conclude that the central limit theorem for $\hat{h}_{k(n)}(x_1^n) - h(\rho)$ is equivalent to the central limit theorem for $-(1/n)\sum_{j=0}^{n-1}\phi(T^j x_1^\infty) - \mathbb{E}_\rho[-\phi]$. In particular, the variance is given by

$$\sigma^2 = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}_\rho\left[\left(\sum_{j=0}^{n-1}\phi(T^j x_1^\infty) - n\mathbb{E}_\rho[\phi]\right)^2\right]. \tag{4.2}$$

At large deviation scale it is possible to see that term $\mathcal{C}_n$ is irrelevant but not $\hat{\Delta}_{k(n)}(x_1^n|\rho)$.

In fact large deviations for $\hat{h}_{k(n)}(x_1^n)$ are different from large deviations for $-(1/n)\sum_{j=0}^{n-1}\phi(T^j x_1^\infty)$. The latter have the same large deviations as $-(1/n)\log\rho([x_1^n])$. Indeed, it is easy to check (using (2.5) and (2.8)) that for any real t

$$\Phi(t) \overset{\text{def}}{=} \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\rho[\mathrm{e}^{-t\sum_{j=0}^{n-1}\phi\circ T^j}] = \lim_{n\to\infty}\frac{1}{n}\log\sum_{a_1^n\in A^n}\rho([a_1^n])^{1-t} = P_{\mathrm{top}}((1-t)\phi).$$

The common rate function for $(-(1/n)\log\rho([x_1^n]))_n$ and $(-(1/n)\sum_{j=1}^{n}\phi(T^j x_1^\infty))_n$ is then given by the Legendre transform of Φ.

In [8], it is proved that $\sigma^2 = (\mathrm{d}^2\Phi/\mathrm{d}t^2)(0) = (\mathrm{d}^2 P_{\mathrm{top}}(t\phi)/\mathrm{d}t^2)(0)$. On the other hand, one expects that the second derivative of the scaled cumulant generating function at 0 (or, equivalently, the inverse of the second derivative at $h(\rho)$ of the rate function) equals the variance[3]. Though $\mathbf{R}(t) \neq \Phi(t)$ for all $t \neq 0$, a simple computation shows that $(\mathrm{d}^2\mathbf{R}/\mathrm{d}t^2)(0) = (\mathrm{d}^2 P_{\mathrm{top}}(t\phi)/\mathrm{d}t^2)(0) = \sigma^2$.

[3] Note that this does not imply a central limit theorem even under real analyticity (see [5]).

Therefore, we have distinct rate functions (because the conditional empirical relative entropy 'correction' contributes at large deviation scale) but their second derivatives at 0 coincide.

Remark. Using (4.1), the fact that $\hat{\Delta}_{k(n)}(x_1^n|\rho) \geqslant 0$, and the fact that $\mathcal{C}_n$ is irrelevant at large deviation scale, it is easy to get

$$\mathbf{R}(t) \leqslant \Phi(t) \quad \forall t > 0, \qquad \mathbf{R}(t) \geqslant \Phi(t) \quad \forall t < 0.$$

5. Some combinatorial tools

In this section we collect some definitions and lemmas about types, as well as a continuity lemma for conditional entropy. These are essential ingredients for the proofs of our main results which are in the next section. The proofs of the following lemmas are given in section 7.

We call $\mathcal{U}^k(A^n)$ the subset of $\mathcal{M}_s^k$ whose elements can be obtained as empirical measure of sample paths of length n. Formally we set

$$\mathcal{U}^k(A^n) = \{\nu_k \in \mathcal{M}_s^k : \exists x_1^n \in A^n \text{ s.t. } \nu_k(\cdot) = \pi_k(\cdot; x_1^n)\}. \tag{5.1}$$

The set A^n of sample paths x_1^n can be partitioned into equivalence classes called types. The equivalence relation $\sim_k$ is defined as

$$x_1^n \sim_k y_1^n \Longleftrightarrow \pi_k(\cdot; x_1^n) = \pi_k(\cdot; y_1^n). \tag{5.2}$$

Let us call $\mathcal{T}^k(A^n) = A^n/\sim_k$ the quotient space. Elements of $\mathcal{T}^k(A^n)$ are labelled with the corresponding empirical measure $\pi_k(\cdot; x_1^n)$; this means that there is a bijective correspondence between $\mathcal{T}^k(A^n)$ and $\mathcal{U}^k(A^n)$. We call $\tau_{\pi_{k,n}} \in \mathcal{T}^k(A^n)$ the type corresponding to $\pi_k(\cdot; x_1^n)$.

We recall that with $\mathcal{E}^k$ we indicate the extremal elements of $\mathcal{M}_s^k$.

Lemma 5.1. *Given a measure $\nu_k \in \mathcal{E}^k$ then $h_k(\nu_k) = 0$.*

Lemma 5.2. *Given a measure $\nu_k \in \mathcal{M}_s^k$ there exists a measure $\mu_k \in \mathcal{U}^k(A^n)$ such that*

$$\|\mu_k - \nu_k\|_{\mathrm{tv}} = \sum_{a_1^k \in A^k} |\mu_k(a_1^k) - \nu_k(a_1^k)| \leqslant \frac{(k+2)|A|^k}{n}. \tag{5.3}$$

Lemma 5.3. *The following inequalities hold*

$$|\mathcal{T}^k(A^n)| = |\mathcal{U}^k(A^n)| \leqslant (n+1)^{|A|^k}, \tag{5.4}$$

$$|\{x_1^n \in \tau_{\pi_{k,n}}\}| \leqslant (n-1)\mathrm{e}^{nh_k(\pi_{k,n})}, \tag{5.5}$$

$$|\{x_1^n \in \tau_{\pi_{k,n}}\}| \geqslant (\mathrm{e}n)^{-2|A|^k}\mathrm{e}^{nh_k(\pi_{k,n})}. \tag{5.6}$$

Lemma 5.4. *We have the following continuity property of the conditional k-block entropy:*

$$\sup_{\{\nu_k,\mu_k : \|\nu_k-\mu_k\|_{\mathrm{tv}} \leqslant \delta\}} |h_k(\nu_k) - h_k(\mu_k)| \leqslant -2\delta \log \frac{\delta}{|A|^k} \tag{5.7}$$

provided that $\delta \leqslant \mathrm{e}^{-1}$.

6. Proofs of main results

6.1. Proof of theorem 3.1

Consider a closed set $C \subseteq \mathbb{R}$. We have

$$\rho\{x_1^n : \hat{h}_k(x_1^n) \in C\} = \sum_{\{x_1^n : \hat{h}_k(x_1^n) \in C\}} \rho([x_1^n]).$$

From (2.5) and (2.6) we get

$$\rho([x_1^n]) = \mathrm{e}^{n\{\mathbb{E}_{\pi_k(\cdot, x_1^n)}[\phi_k]\}} \vartheta_{k,n}(x_1^n), \tag{6.1}$$

where

$$\mathrm{e}^{-n(\varepsilon_n + \mathrm{var}_k(\phi))} \leqslant \vartheta_{k,n}(x_1^n) \leqslant \mathrm{e}^{n(\varepsilon_n + \mathrm{var}_k(\phi))}. \tag{6.2}$$

Hence we have

$$\sum_{\{x_1^n : \hat{h}_k(x_1^n) \in C\}} \rho([x_1^n]) \leqslant \mathrm{e}^{n(\varepsilon_n + \mathrm{var}_k(\phi))} \times \sum_{\{\pi_{k,n} \in \mathcal{U}^k(A^n) : h_k(\pi_{k,n}) \in C\}} |\{x_1^n \in \tau_{\pi_{k,n}}\}| \, \mathrm{e}^{n\{\mathbb{E}_{\pi_{k,n}}[\phi_k]\}},$$

where we have used types defined in section 5. Let us call

$$h_k^{-1}(C) \overset{\text{def}}{=} \{\nu_k \in \mathcal{M}_s^k : h_k(\nu_k) \in C\} \qquad \text{and} \qquad h^{-1}(C) \overset{\text{def}}{=} \{\mu \in \mathcal{M}_s : h(\mu) \in C\}.$$

Using inequalities (5.4) and (5.5) we obtain the following upper bound

$$\sum_{\{x_1^n : \hat{h}_k(x_1^n) \in C\}} \rho([x_1^n]) \leqslant \mathrm{e}^{n(\varepsilon_n + \mathrm{var}_k(\phi))} (n+1)^{|A|^k} (n-1) \exp\left(n \left\{ \sup_{\nu_k \in h_k^{-1}(C)} (\mathbb{E}_{\nu_k}[\phi_k] + h_k(\nu_k)) \right\}\right). \tag{6.3}$$

If we consider sequences $(k(n))_{n \in \mathbb{N}}$ that satisfy the growth condition (3.1) we obtain

$$\limsup_{n \to \infty} \frac{1}{n} \log \rho\{x_1^n : \hat{h}_{k(n)}(x_1^n) \in C\} \leqslant \limsup_{k \to \infty} \sup_{\nu_k \in h_k^{-1}(C)} (\mathbb{E}_{\nu_k}[\phi_k] + h_k(\nu_k)).$$

We will prove that for any $\varepsilon > 0$ there exists an integer K such that for any $k > K$ and for any $\nu_k \in h_k^{-1}(C)$ there exists a $\mu \in h^{-1}(C)$ such that

$$\mathbb{E}_{\nu_k}[\phi_k] + h_k(\nu_k) \leqslant h(\mu) + \mathbb{E}_{\mu}[\phi] + \varepsilon. \tag{6.4}$$

The arbitrariness of ε will imply the first statement of the theorem.

To prove formula (6.4) we only have to take μ as the unique $(k-1)$-step Markov extension of ν_k and K such that $\mathrm{var}_K(\phi) < \varepsilon$.

Let us now prove the lower bound. Consider an open set $O \subseteq \mathbb{R}$.

$$\sum_{\{x_1^n : \hat{h}_k(x_1^n) \in O\}} \rho([x_1^n]) \geqslant \mathrm{e}^{-n(\varepsilon_n + \mathrm{var}_k(\phi))} \times \sum_{\{\pi_{k,n} \in \mathcal{U}^k(A^n) : h_k(\pi_{k,n}) \in O\}} |\{x_1^n \in \tau_{\pi_{k,n}}\}| \mathrm{e}^{n\{\mathbb{E}_{\pi_{k,n}}[\phi_k]\}}. \tag{6.5}$$

Using inequality (5.6) we obtain

$$\begin{aligned}\sum_{\{x_1^n : \hat{h}_k(x_1^n) \in O\}} \rho([x_1^n]) &\geqslant \mathrm{e}^{-n(\varepsilon_n + \mathrm{var}_k(\phi))} (\mathrm{e}n)^{-2|A|^k} \sum_{\{\pi_{k,n} \in \mathcal{U}^k(A^n) : h_k(\pi_{k,n}) \in O\}} \mathrm{e}^{n\{h_k(\pi_{k,n}) + \mathbb{E}_{\pi_{k,n}}[\phi_k]\}} \\ &\geqslant \mathrm{e}^{-n(\varepsilon_n + \mathrm{var}_k(\phi))} (\mathrm{e}n)^{-2|A|^k} \exp\left(n \left\{ \sup_{\{\nu_k \in h_k^{-1}(O) \cap \mathcal{U}^k(A^n)\}} (h_k(\nu_k) + \mathbb{E}_{\nu_k}[\phi_k]) \right\}\right).\end{aligned}$$

If we consider sequences $(k(n))_{n\in\mathbb{N}}$ which satisfy the growth condition (3.1) we obtain

$$\liminf_{n\to\infty}\frac{1}{n}\log\rho\{x_1^n:\hat{h}_{k(n)}(x_1^n)\in O\}\geqslant\liminf_{n\to\infty}\sup_{\{\nu_{k(n)}\in h_{k(n)}^{-1}(O)\cap\mathcal{U}^{k(n)}(A^n)\}}(\mathbb{E}_{\nu_{k(n)}}[\phi_{k(n)}]+h_{k(n)}(\nu_{k(n)})).$$

We will prove that for any $\varepsilon>0$ and for any $\mu\in h^{-1}(O)$ there exists a $\pi_{k(n),n}\in h_{k(n)}^{-1}(O)\cap\mathcal{U}^{k(n)}(A^n)$ such that

$$\mathbb{E}_{\pi_{k(n),n}}[\phi_{k(n)}]+h_{k(n)}(\pi_{k(n),n})\geqslant h(\mu)+\mathbb{E}_\mu[\phi]-\varepsilon.$$

The arbitrariness of ε implies the second statement of theorem 3.1.

When n is large enough $|h_{k(n)}(\mu_{k(n)})-h(\mu)|$ can become arbitrarily small and from lemmas 5.2 and 5.4, if $d_n\overset{\text{def}}{=}(k(n)+2)(|A|^{k(n)}/n)$, there exists a measure $\pi_{k(n),n}\in\mathcal{U}^{k(n)}(A^n)$ such that

$$|h_{k(n)}(\mu_{k(n)})-h_{k(n)}(\pi_{k(n),n})|\leqslant-2d_n\log\frac{d_n}{|A|^{k(n)}}\,.$$

For a sequence $(k(n))_{n\in\mathbb{N}}$ which satisfy the growth condition (3.1) both d_n and $-2d_n\log(d_n/|A|^{k(n)})$ converge to zero. Since O is an open set we obtain that if n is large enough there exists a $\pi_{k(n),n}\in h_{k(n)}^{-1}(O)\cap\mathcal{U}^{k(n)}(A^n))$ and such that $|h_{k(n)}(\pi_{k(n),n})-h(\mu)|$ is arbitrarily small. It is also easy to show that

$$|\mathbb{E}_\mu(\phi)-\mathbb{E}_{\pi_{k(n),n}}(\phi_{k(n)})|\leqslant\mathrm{var}_{k(n)}(\phi)+d_n\|\phi\|_\infty.$$

The statement easily follows.

The proof for the estimator $\hat{H}_{k(n)}(x_1^n)/k(n)$ is analogous; we will only point out the differences.

For the upper bound we need to prove that for any $\varepsilon>0$ there exists a K such that for any $k>K$ and for any $\nu_k\in\mathcal{M}_s^k$ with $(H_k(\nu_k)/k)\in C$, there exists $\mu\in\mathcal{M}_s$ with $h(\mu)\in C$ and such that inequality (6.4) holds. This can be done considering $\mu=(\nu_1^M+\cdots+\nu_k^M)/k$, where $\nu_i^M\in\mathcal{M}_s$ is the unique $(i-1)$-step Markov extension of ν_i. Due to the fact that h is affine on $\mathcal{M}_s$, we have in fact $h(\mu)=H_k(\nu_k)/k$.

The proof of the lower bound is similar. We omit the details.

The convexity of $\mathbf{I}$ follows from the fact that the maps $h(\cdot),h(\cdot|\rho):\mathcal{M}_s\to\mathbb{R}$ are affine. Given $\nu\in\mathcal{M}_s$ such that $h(\nu)=x$ and $\mu\in\mathcal{M}_s$ such that $h(\mu)=y$, then for any $c\in[0,1]$

$$\begin{aligned}&h(c\nu+(1-c)\mu)=cx+(1-c)y,\\&h(c\nu+(1-c)\mu|\rho)=ch(\nu|\rho)+(1-c)h(\mu|\rho).\end{aligned}$$

This implies that

$$\mathbf{I}(cx+(1-c)y)\leqslant h(c\nu+(1-c)\mu|\rho)=ch(\nu|\rho)+(1-c)h(\mu|\rho).\qquad(6.6)$$

If we take the infimum over all $\nu\in\mathcal{M}_s$ such that $h(\nu)=x$ and $\mu\in\mathcal{M}_s$ such that $h(\mu)=y$ from (6.6) one obtains the convexity of $\mathbf{I}$.

Theorem 3.1 is proved.

6.2. Proof of theorem 3.2

The proof of theorem 3.2 is similar to that of theorem 3.1, so we leave the details to the reader.

6.3. *Proof of proposition 3.3*

Let us recall the following large deviation principle [6]. Let x_1^n be a sample path distributed according to a g-measure ρ. Then the empirical process $\pi(\cdot; x_1^n)$ defined at (2.11) satisfies a large deviation principle in $(\mathcal{M}_s, d_w)$ with normalizing factor $1/n$ and rate function

$$I^\pi(\nu) = h(\nu|\rho). \tag{6.7}$$

Here d_w is a distance that metrizes weak convergence.

Now we observe that for every fixed k the entropies upon consideration are continuous in $(\mathcal{M}_s, d_w)$. Therefore, the contraction principle [11] immediately yields the proposition.

6.4. *Proof of proposition 3.4*

We prove that the Legendre transform of **I** is **R**. We know from theorem 3.1 that **I** is a convex function and this implies the Legendre duality.

We have

$$\sup_{u\in[0,\log|A|]}\left\{tu - \inf_{\{\nu\in\mathcal{M}_s:h(\nu)=u\}} h(\nu|\rho)\right\} = \sup_{\nu\in\mathcal{M}_s}\{\mathbb{E}_\nu[\phi] + th(\nu) + h(\nu)\}. \tag{6.8}$$

If $t > -1$, then we get by applying the variational principle

$$(6.8) = (t+1)\sup_{\nu\in\mathcal{M}_s}\left\{\mathbb{E}_\nu\left[\frac{\phi}{t+1}\right] + h(\nu)\right\} = (t+1)P_{\text{top}}\left(\frac{\phi}{t+1}\right).$$

If $t < -1$, we get

$$(6.8) = (t+1)\inf_{\nu\in\mathcal{M}_s}\left\{\mathbb{E}_\nu\left[\frac{\phi}{t+1}\right] + h(\nu)\right\}.$$

Observe that $h(\nu) \geqslant 0$ for all $\nu \in \mathcal{M}_s$. Moreover, the set of measures with entropy 0 is dense in $\mathcal{M}_s$ (w.r.t. weak topology), see, e.g. [12]. Hence, for $t < -1$, (6.8) $= (t+1)\inf\{\mathbb{E}_\eta[\phi/(t+1)] : \eta \in \mathcal{M}_s\}$. The case $t = -1$ is trivial.

The identification of $\mathbf{R}(t)$ with the scaled cumulant generating functions (formula (3.5)) follows from general arguments [11].

It is interesting to note that using the combinatorial properties of types and the results of section 5 it is possible to prove (3.5) directly. We just sketch the proof.

Following the arguments already used in the proof of theorem 3.1 we can obtain

$$\frac{1}{n}\log\sum_{x_1^n\in A^n} \mathrm{e}^{nt\hat{h}_{k(n)}(x_1^n)}\rho([x_1^n]) \leqslant \sup_{\nu_{k(n)}\in\mathcal{M}_s^{k(n)}}\{\mathbb{E}_{\nu_{k(n)}}[\phi_{k(n)}] + (t+1)h_{k(n)}(\nu_{k(n)})\} + \overline{R}_n \tag{6.9}$$

and

$$\frac{1}{n}\log\sum_{x_1^n\in A^n} \mathrm{e}^{nt\hat{h}_{k(n)}(x_1^n)}\rho([x_1^n]) \geqslant \sup_{\nu_{k(n)}\in\mathcal{U}^{k(n)}(A^n)}\{\mathbb{E}_{\nu_{k(n)}}[\phi_{k(n)}] + (t+1)h_{k(n)}(\nu_{k(n)})\} + \underline{R}_n, \tag{6.10}$$

where $\overline{R}_n$ and $\underline{R}_n$ are correcting terms converging to zero.

We now compute the supremum in (6.9).

If $t \leqslant -1$, the function to be maximized is convex and the supremum is attained at one of the extremal points of $\mathcal{M}_s^k$, which has entropy zero by virtue of lemma 5.1. Hence the supremum in question equals

$$\sup\{\mathbb{E}_{\nu_{k(n)}}[\phi_{k(n)}] : \nu_{k(n)} \in \mathcal{E}^{k(n)}\}. \tag{6.11}$$

If $t > -1$, the supremum in (6.9) is equal to

$$(t+1)\sup_{\nu\in\mathcal{M}_s}\left\{\mathbb{E}_\nu\left[\frac{\phi_{k(n)}}{t+1}\right]+h(\nu)\right\}=(t+1)\ P_{\text{top}}\left(\frac{\phi_{k(n)}}{t+1}\right).$$

To see this, we first notice that if ν is the $(k(n)-1)$-step Markov measure having $\nu_{k(n)}$ as $k(n)$-marginals, then $h_k(\nu_{k(n)}) = h(\nu)$. On the other hand, the variational principle tells us that $\mathbb{E}_\nu[\phi_{k(n)}/(t+1)]+h(\nu)$ attains its supremum precisely at a unique $(k(n)-1)$-step Markov measure because $\phi_{k(n)}$ is a $k(n)$-cylindrical function. This supremum equals $P_{\text{top}}(\phi_{k(n)}/(t+1))$.

It is not difficult to prove now that the limit when $n\to\infty$ of the upper bound coincides with $\mathbf{R}(t)$. Using the results of section 5 it is also possible to prove that the lower bound has the same limit.

The result for the estimator $\hat{H}_{k(n)}(x_1^n)/k(n)$ can be deduced from the previous result using the fact that $(\hat{h}_i(x_1^n))_i$ is a bounded decreasing sequence and

$$\hat{H}_k(x_1^n)=\sum_{i=1}^{k}\hat{h}_i(x_1^n). \tag{6.12}$$

6.5. *Proof of proposition 3.5*

The proof of this proposition is very simple and left to the reader. It is possible to get (3.7) directly using the combinatorics of types.

6.6. *Proof of proposition 3.6*

When the variations of ϕ are square summable the map $\beta\mapsto P_{\text{top}}(\beta\phi)$, $\beta\in\mathbb{R}$, is continuously differentiable and strictly convex. This can be deduced from [25]. The extension of their proofs to the square summable case is straightforward. This implies that the map $\mathbf{R}$ is continuously differentiable and strictly convex in the interval $(-1,\infty)$. Moreover $\mathbf{R}(0)=0$ and $(\mathrm{d}\mathbf{R}/\mathrm{d}t)(0)=h(\rho)$. This establishes the first part of the proposition.

We now turn to prove the representation formula (3.9). First introduce the following auxiliary function of $\beta\in[0,+\infty)$:

$$\mathcal{I}(\beta)\stackrel{\text{def}}{=}\inf\{h(\nu|\rho):\nu\in\mathcal{M}_s, h(\nu)=h(\rho_{\beta\phi})\}.$$

We now claim that $\mathcal{I}(\beta)=h(\rho_{\beta\phi}|\rho)$. The proof is by contradiction of the variational principle.

Assume that $\eta\neq\rho_{\beta\phi}$ is such that

$$h(\eta|\rho)\leqslant h(\rho_{\beta\phi}|\rho)\qquad\text{and}\qquad h(\eta)=h(\rho_{\beta\phi}).$$

This means that (remember (2.10))

$$\mathbb{E}_\eta[\phi]\geqslant\mathbb{E}_{\rho_{\beta\phi}}[\phi].$$

Multiplying this inequality by $\beta>0$ and adding $h(\eta)$ to the lhs and $h(\rho_{\beta\phi})$ to the rhs (since these two quantities are indeed equal by hypothesis) yields

$$\mathbb{E}_\eta[\beta\phi]+h(\eta)\geqslant\mathbb{E}_{\rho_{\beta\phi}}[\beta\phi]+h(\rho_{\beta\phi}).$$

But the variational principle states that the rhs is equal to the supremum over all shift-invariant measures ν of $\mathbb{E}_\nu[\beta\phi]+h(\nu)$ and is attained only for $\nu=\rho_{\beta\phi}$. Therefore η must be equal to $\rho_{\beta\phi}$. In this instance of the variational principle, we used the fact that if a potential ϕ has square summable variations then $\beta\phi$ also has square summable variations, in particular, for any $\beta>0$.[4]

[4] In the case of non-uniqueness, the claim still holds, but $\rho_{\beta\phi}$ is *any* equilibrium state associated to $\beta\phi$ since relative entropy only depends on $\beta\phi$.

We now invoke lemma 6.1 hereafter to define a map $\mathcal{H} : [0, +\infty[\to]h_\infty, \log|A|]$ defined as $\mathcal{H}(\beta) = h(\rho_{\beta\phi})$. Since this map is continuous, strictly decreasing, to each $u \in]h_\infty, \log|A|]$ we can associate a unique β_u such that $h(\rho_{\beta_u}) = u$.

The last statement of the proposition follows from the first comment in section 4. □

We state and prove the lemma used just above.

Lemma 6.1. *Assume that ϕ has square summable variations (hence so has $\beta\phi$ for all $\beta \in \mathbb{R}$) and is not cohomologous to a constant*[5]. *Then the map $\beta \mapsto h(\rho_{\beta\phi})$ is continuous, strictly decreasing on* $[0, +\infty[$ *and* $h(\rho_{\beta\phi}) \in]h_\infty, \log|A|]$.

Proof. By the variational principle, $h(\rho_{\beta\phi}) = P_{\text{top}}(\beta\phi) - \beta\mathbb{E}_{\rho_{\beta\phi}}[\phi]$. (This shows continuity.) $\beta \mapsto P_{\text{top}}(\beta\phi)$ is strictly decreasing (since $\phi < 0$) and strictly convex (see above). This strict convexity of the pressure can be translated as follows [19]

$$\beta_1 < \beta_2 \quad \Rightarrow \quad \mathbb{E}_{\rho_{\beta_1\phi}}[\phi] < \mathbb{E}_{\rho_{\beta_2\phi}}[\phi].$$

Therefore we get that $\beta \mapsto h(\rho_{\beta\phi})$ is strictly decreasing when $\beta > 0$. It is obvious from the variational principle that $h(\rho_{\beta\phi}) = \log|A|$ when $\beta = 0$. Since $h(\rho_{\beta\phi})$ is bounded from below by 0, $h_\infty = \lim_{\beta\to+\infty} h(\rho_{\beta\phi})$ exists. This ends the proof of the lemma. □

7. Proofs of some lemmas

This section contains the proof of the lemmas of section 5.

Let us introduce the following graph theoretical representations that we will use in the proofs of the lemmas. We call $\mathcal{N}_n^k$ the set of integer-valued maps $N_n^k : A^k \to \mathbb{N}$ such that

$$\sum_{a_1^k \in A^k} N_n^k(a_1^k) = n \tag{7.1}$$

and

$$\sum_{b\in A} N_n^k(a_1^{k-1}b) = \sum_{b\in A} N_n^k(ba_1^{k-1}) \qquad \forall a_1^{k-1} \in A^{k-1}. \tag{7.2}$$

Let $\mathcal{L}_n^k$ be the subset of $\mathcal{M}_s^k$ whose elements are obtained by normalizing elements in $\mathcal{N}_n^k$, i.e.

$$\mathcal{L}_n^k = \left\{ \nu_k \in \mathcal{M}_s^k : \exists N_n^k \in \mathcal{N}_n^k \text{ s.t. } \nu_k(\cdot) = \frac{N_n^k(\cdot)}{n} \right\}. \tag{7.3}$$

If $k = 1$ then $\mathcal{U}^1(A^n) = \mathcal{L}_n^1$, otherwise a strict inclusion holds $\mathcal{U}^k(A^n) \subset \mathcal{L}_n^k$ $(n > 1)$.

We will call a k-order compatible balanced directed multigraph (k-multigraph, k-M, for short) a directed multigraph with the following properties: the vertices are labelled with elements of A^{k-1}; for each vertex the number of outgoing arrows is equal to the number of ingoing arrows; an arrow can go from the vertex a_1^{k-1} to the vertex b_1^{k-1} if and only if $a_2^{k-1} = b_1^{k-2}$. This arrow inherits the natural label $a_1^{k-1}b^{k-1}$ (note that several arrows can have the same label).

Given an element $N_n^k \in \mathcal{N}_n^k$ we represent it with a k-M containing n arrows ([11], section II.2) drawing $N_n^k(b_1^k)$ directed edges from the vertex associated to b_1^{k-1} to the one associated to b_2^k.

[5] This means that ϕ is not the equilibrium measure for a potential of the form $V - V \circ T + c$, where V is a measurable function, $c \in \mathbb{R}$. In this case the equilibrium measure would coincide with the measure of maximal entropy, the uniform Bernoulli measure.

Conversely, given a k-M containing n arrows, it is possible to associate to it an element of $\mathcal{N}_n^k$ defining $N_n^k(a_1^k)$ as the number of arrows going from a_1^{k-1} to a_2^k. This gives a bijective correspondence.

To each element $\nu_k \in \mathcal{U}^k(A^n)$, we associate the element $N_n^k = n\nu_k \in \mathcal{N}_n^k$. Then we construct a k-M as before, which is connected (note that we are not considering vertices without ingoing/outgoing arrows). Given two vertices a_1^{k-1} and b_1^{k-1} which have some ingoing/outgoing arrows, there exist $i < j$ with $|i - i| < n$ such that $\tilde{x}_i^{i+k-2} = a_1^{k-1}$ and $\tilde{x}_j^{j+k-2} = b_1^{k-1}$. This means that for any $i \leqslant l < j$ there exists at least one arrow with label $\tilde{x}_l^{l+k-1}$, i.e. at least one path from the vertex a_1^{k-1} to the vertex b_1^{k-1}.

Conversely, given a connected k-M we associate to it an element of $\mathcal{U}^k(A^n)$. A connected k-M has at least one Eulerian circuit (see, e.g. [3], section I.3). One follows the circuit generating a sample path in the following way: every time one goes through an arrow with label a_1^k, one concatenates the element a_k. The sample path x_1^n that one obtains in this way is such that $n\pi(\cdot; x_1^n)$ has associated the connected k-M one started with.

This is a bijective correspondence between $\mathcal{U}^k(A^n)$ and the subclass of connected k-M containing n arrows. This correspondence says that it is possible, starting from the k-M associated to an element $\pi_{k,n} \in \mathcal{U}^k(A^n)$, to construct an element $x_1^n \in \tau_{\pi_{k,n}}$ by simply following an Eulerian circuit.

Some classical combinatorial arguments allow us to estimate the number of Eulerian circuits of a k-M, and this gives an estimate on the number of samples $x_1^n \in \tau_{\pi_{k,n}}$ (see [11], section II.2):

$$\frac{\prod_{a_1^{k-1}}(n\sum_b \pi_{k,n}(a_1^{k-1}b) - 1)!}{\prod_{a_1^k}(n\pi_{k,n}(a_1^k))!} \leqslant |\{x_1^n \in \tau_{\pi_{k,n}}\}| \leqslant n\frac{\prod_{a_1^{k-1}}(n\sum_b \pi_{k,n}(a_1^{k-1}b))!}{\prod_{a_1^k}(n\pi_{k,n}(a_1^k))!}. \tag{7.4}$$

We will call a k-order weighted compatible balanced directed graph (k-weighted graph, k-WG, for short) a directed graph with the following properties: the vertices are labelled with elements of A^{k-1}; to each arrow is associated a non-negative weight; for each vertex, the sum of the weights associated to outgoing arrows is equal to the sum of the weights associated to ingoing arrows; an arrow can go from the vertex a_1^{k-1} to the vertex b_1^{k-1} if and only if $a_2^{k-1} = b_1^{k-2}$; the total sum of the weights is 1.

Given a measure $\nu_k \in \mathcal{M}_s^k$ we can represent it by a k-WG, and conversely given a k-WG we can associate to it an element of $\mathcal{M}_s^k$.

7.1. Proof of lemma 5.1

A convex combination of measures corresponds to a k-WG with a convex combination of weights. Therefore the extremality property in $\mathcal{M}_s^k$ corresponds to the extremality property in the set of k-WGs. Consider a k-WG having non-zero weights only on arrows forming a single cycle (a loop of successive arrows visiting a vertex no more than once). All the non-zero weights are equal to $1/\ell$, where ℓ is the length of the cycle. Every such k-WG cannot be obtained as a convex combination of other k-WGs. Otherwise at least one of them would violate one of the conditions to be a k-WG. Moreover any k-WG can be obtained as a convex combination of a finite number of k-WGs consisting of a single cycle. A decomposition can be obtained by iterating a finite number of times the following procedure. Take the (an) arrow to which is associated the minimum weight and consider a cycle containing it. Substract the minimum weight to all the arrows belonging to the cycle and add the k-WG consisting of the single cycle weighted by m.w.$/\ell$, where m.w. $=$ minimum weight, to the convex decomposition. This gives a complete characterization of $\mathcal{E}^k$. A direct consequence is that $h_k(\nu_k) = 0$ for every

$\nu_k \in \mathcal{E}^k$. This is because for every measure ν_k with an associated k-WG consisting of a single cycle $\nu_k(a_k|a_1^{k-1})$ can be only zero or one. The lemma is proved.

7.2. *Proof of lemma 5.2*

Given a measure $\nu_k \in \mathcal{M}_s^k$ it is possible to construct a measure $\tilde{\mu}_k \in \mathcal{L}_n^k$ such that $\|\tilde{\mu}_k - \nu_k\|_{\mathrm{tv}} \leqslant (2A^k/n)$. This is trivial when $k = 1$ and a little more tricky when $k > 1$ because of the stationarity condition (2.1). Consider for any arrow from a_1^{k-1} to a_2^k the following parameter

$$\gamma(a_1^k) = \min\left\{\left|\nu_k(a_1^k) - \frac{[n\nu_k(a_1^k)] + 1}{n}\right|, \left|\nu_k(a_1^k) - \frac{[n\nu_k(a_1^k)]}{n}\right|\right\}. \tag{7.5}$$

where $[\cdot]$ represents the integer part. Take the (an) arrow with the associated minimum value of γ. Consider an elementary cycle containing a_1^k and add or subtract (depending if the minimum value in (7.5) was obtained with the first or the second argument) the value $\gamma(a_1^k)$ to all the elements of the cycle. Fix the values of all the weights whose value is i/n with i some integer number $0 \leqslant i \leqslant n$, and remove them from the k-WG. It is easy to see that one can iterate this procedure up to fix all the values of the weights. One ends up with some weights which satisfy the stationarity condition but are not necessarily normalized to one. One concludes the procedure by adding or subtracting the weight necessary to have the wanted normalization. One can do this sequentially by using an elementary unit of weight $1/n$ and adding or subtracting one unit of weight at the same time in elementary cycles, so that the stationarity condition is preserved. This is always possible. The measure $\tilde{\mu}_k$ that is obtained in this way belongs to $\mathcal{L}_n^k$ and is such that

$$\|\tilde{\mu}_k - \nu_k\|_{\mathrm{tv}} \leqslant \frac{2|A|^k}{n}.$$

If the k-M corresponding to $\tilde{\mu}_k$ is connected then the proof is finished. If the k-M associated to $\tilde{\mu}_k$ is not connected let $m > 1$ be the number of connected components containing, respectively, $e(1), \ldots, e(m)$ directed edges with $\sum_{j=1}^m e(j) = n$. Considering a Eulerian circuit for every component one can associate a sample path $s(i)$ of length $e(i)$ to the ith component for $i = 1, \ldots, m$. The measure $\tilde{\mu}_k$ has the following expression

$$\tilde{\mu}_k(\cdot) = \sum_{j=1}^{m} \frac{e(j)}{n} \pi_k(\cdot; s(j)). \tag{7.6}$$

Let us now consider the sample path $s = s(1)s(2)\cdots s(m)$ of length n and construct μ_k as the k empirical measure $\mu_k(\cdot) = \pi_k(\cdot; s) \in \mathcal{U}^k(A^n)$. Both μ_k and $\tilde{\mu}_k$ are constructed by sliding windows of width k along cyclicized samples and computing frequencies in these windows. Every time the window of size k is overlapped to the sample s and does not cross points of separation among different $s(i)$ the k-sequence that is matched contributes both in $\tilde{\mu}_k$ and μ_k. Using the fact that $m \leqslant A^{k-1}$ we deduce

$$\|\mu_k - \tilde{\mu}_k\|_{\mathrm{tv}} \leqslant \frac{k|A|^{k-1}}{n}. \tag{7.7}$$

Using (7.6), (7.7) and the triangle inequality yields the statement of the lemma.

7.3. *Proof of lemma 5.3*

The proof of inequalities (5.4) and (5.5) is very simple and elegant and can be found in [23], more precisely in section I.6.d in the case of non-cyclicized samples and in section II.1.a in the case of cyclicized samples, which is our case. The proof of inequality (5.6) is obtained from estimate (7.4) and the Stirling formula. Inequality (5.5) can be proved in an analogous way.

7.4. Proof of lemma 5.4

This lemma can be found in [10] but we give its proof for the sake of completeness. Consider μ_k and ν_k two measures in $\mathcal{M}^k$ such that $\|\nu_k - \mu_k\|_{\mathrm{tv}} \leqslant \delta$. Let us set $\delta_k(a_1^k) = |\nu_k(a_1^k) - \mu_k(a_1^k)|$. Obviously

$$\sum_{a_1^{k-1}} \delta_{k-1}(a_1^{k-1}) \leqslant \sum_{a_1^k} \delta_k(a_1^k) \leqslant \delta.$$

Using the triangle inequality one obtains

$$\begin{aligned}|h_k(\nu_k) - h_k(\mu_k)| \leqslant \sum_{a_1^k} &|\nu_k(a_1^k)\log\nu_k(a_1^k) - \mu_k(a_1^k)\log\mu_k(a_1^k)| \\ &+ \sum_{a_1^{k-1}} |\nu_{k-1}(a_1^{k-1})\log\nu_{k-1}(a_1^{k-1}) - \mu_{k-1}(a_1^{k-1})\log\mu_{k-1}(a_1^{k-1})|.\end{aligned}$$

By a simple computation it is possible to obtain the modulus of continuity of the function $-x\log x$ on the interval [0, 1] when δ is small enough

$$\sup_{\{x,y\in[0,1]:|x-y|\leqslant\delta\}} |x\log x - y\log y| = -\delta\log\delta.$$

Using this result we get

$$|h_k(\nu_k) - h_k(\mu_k)| \leqslant -\sum_{a_1^k} \delta_k(a_1^k)\log\delta_k(a_1^k) - \sum_{a_1^{k-1}} \delta_{k-1}(a_1^{k-1})\log\delta_{k-1}(a_1^{k-1}). \tag{7.8}$$

We write the rhs of (7.8) as

$$-|A|^k \sum_{a_1^k} \frac{\delta_k(a_1^k)}{|A|^k}\log\delta_k(a_1^k) - |A|^{k-1}\sum_{a_1^{k-1}} \frac{\delta_{k-1}(a_1^{k-1})}{|A|^{k-1}}\log\delta_{k-1}(a_1^{k-1}), \tag{7.9}$$

and apply Jensen inequality using the fact that $-x\log x$ is a concave function. When δ is small enough we finally obtain

$$\begin{aligned}|h_k(\nu_k) - h_k(\mu_k)| &\leqslant -\delta\log\frac{\delta}{|A|^k} - \delta\log\frac{\delta}{|A|^{k-1}} \\ &\leqslant -2\delta\log\frac{\delta}{|A|^k}\cdot \end{aligned}\tag{7.10}$$

The lemma is proved.

Acknowledgments

We acknowledge the financial support of the Cofin MIUR 2002 prot. 2002027798_005. We thank the CIC, Cuernavaca, México, for its warm hospitality, where part of this work was done. We also thank A Galves and D Guiol for discussions about the entropy estimation problem. This project started when both authors were postdoctoral fellows at IME-USP (Saõ Paulo) with a FAPESP grant.

References

[1] Bowen R 1975 *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms* (*Lecture Notes in Mathematics* vol 470) (Berlin: Springer)

[2] Berger N, Hoffman C and Sidoravicius V 2003 Nonuniqueness for specifications in $\ell^{2+\varepsilon}$ *Preprint* math.PR/0312334

[3] Bollobas B 1998 *Modern Graph Theory* (*Graduate Texts in Mathematics* vol 184) (New York: Springer)

[4] Bramson M and Kalikow S A 1993 Nonuniqueness in g-functions *Israel J. Math.* **84** 153–60

[5] Bryc W 1993 A remark on the connection between the large deviation principle and the central limit theorem *Stat. Probab. Lett.* **18** 253–6

[6] Chazottes J-R and Olivier E 2000 Relative entropy, dimensions and large deviations for g-measures *J. Phys. A: Math. Gen.* **33** 675–89

[7] Chazottes J-R and Ugalde E 2005 Entropy estimation and fluctuations of hitting and recurrence times for Gibbsian sources *Discrete Continuous Dyn. Syst.* B **5** 565–86

[8] Coelho Z and Parry W 1990 Central limit asymptotics for shifts of finite type *Israel J. Math.* **69** 235–49

[9] Collet P, Galves A and Schmitt B 1999 Repetition times for Gibbsian sources *Nonlinearity* **12** 1225–37

[10] Csiszár I and Körner J 1981 Information theory. Coding theorems for discrete memoryless systems *Probability and Mathematical Statistics* (New York: Academic)

[11] Den Hollander F 2000 Large deviations *Fields Inst. Monogr.* vol 14

[12] Denker M, Grillenberger C and Sigmund K 1976 *Ergodic Theory on Compact Spaces* (*Lecture Notes in Mathematics* vol 527) (Berlin: Springer)

[13] Fernández R, Ferrari P A and Galves A 2001 *Coupling, Renewal and Perfect Simulation of Chains of Infinite Order* (*Lecture Notes for the 5th Brazilian School of Probability*) (*Ubatuba, August 2001*) pp 92, available at http://www.ime.br/˜pablo/publications.html

[14] Fernández R and Maillard G 2005 Chains with complete connections: general theory, uniqueness, loss of memory and mixing properties *J. Stat. Phys.* **118** 555–88

[15] Fernández R and Maillard G 2004 Chains with complete connections and one-dimensional Gibbs measures *Electron. J. Probab.* **9** 145–76

[16] Gabrielli D, Galves A and Guiol D 2003 Fluctuations of the empirical entropies for a chain of infinite order *Math. Phys. Electron. J.* **9** No 5

[17] Iosifescu M and Grigorescu S 1990 *Dependence with Complete Connections and its Applications* (*Cambridge Tracts in Mathematics* vol 96) (Cambridge: Cambridge University Press)

[18] Johansson A and Öberg A 2003 Square summability of variations of g-functions and uniqueness of g-measures *Math. Res. Lett.* **10** 587–601

[19] Keller G 1998 *Equilibrium States in Ergodic Theory* (*London Mathematical Society Student Texts* vol 42) (Cambridge: Cambridge University Press)

[20] Ledrappier F 1974 Principe variationnel et systèmes symboliques *Z. Wahr. Verw. Geb.* **30** 185–202

[21] Ornstein D S and Weiss B 1990 How sampling reveals a process *Ann. Probab.* **18** 905–30

[22] Schurmann T and Grassberger P 1996 Entropy estimation of symbol sequences *Chaos* **6** 414–27

[23] Shields P C 1996 *The Ergodic Theory of Discrete Sample Paths* (*Graduate Studies in Mathematics* vol 13) (Providence, RI: American Mathematical Society)

[24] Shields P C 2001 private communication

[25] Takens F and Verbitskiy E 1999 Multifractal analysis of local entropies for expansive homeomorphisms with specification *Commun. Math. Phys.* **203** 593–612

[26] Walters P 1978 Invariant measures and equilibrium states for some mappings which expand distances *Trans. Am. Math. Soc.* **236** 121–53