Stochastics and Dynamics
Vol. 12, No. 4 (2012) 1250006 (31 pages)

DOI: 10.1142/S0219493712500062

THE CENTRAL LIMIT THEOREM FOR UNIFORMLY STRONG MIXING MEASURES

NICOLAI HAYDN

Mathematics Department,
USC, LA 90089-1113, USA
nhaydn@math.usc.edu

Received 5 December 2010
Accepted 3 May 2011
Published 6 July 2012

The theorem of Shannon–McMillan–Breiman states that for every generating partition on an ergodic system, the exponential decay rate of the measure of cylinder sets equals the metric entropy almost everywhere (provided the entropy is finite). In this paper we prove that the measure of cylinder sets are lognormally distributed for strongly mixing systems and infinite partitions and show that the rate of convergence is polynomial provided the fourth moment of the information function is finite. Also, unlike previous results by Ibragimov and others which only apply to finite partitions, here we do not require any regularity of the conditional entropy function. We also obtain the law of the iterated logarithm and the weak invariance principle for the information function.

Keywords: Mixing dynamical systems; invariant measures; central limit theorem; Stein's method; weak invariance principle.

AMS Subject Classification: 37A50, 60F05

1. Introduction

Let μ be a T-invariant probability measure on a space Ω on which the map T acts measurably. For a measurable partition $\mathcal{A}$ one forms the nth join $\mathcal{A}^n = \bigvee_{j=0}^{n-1} T^{-j}\mathcal{A}$ which forms a finer partition of Ω. (The atoms of $\mathcal{A}^n$ are traditionally called *n-cylinders*.) For $x \in \Omega$ we denote by $A_n(x) \in \mathcal{A}^n$ the n-cylinder which contains x. The theorem of Shannon–McMillan–Breiman (see e.g. [21, 28]) then states that for μ-almost every x in Ω the limit

$$\lim_{n\to\infty} \frac{-\log \mu(A_n(x))}{n}$$

exists and equals the metric entropy $h(\mu)$ provided the entropy is finite in the case of a countable infinite partition. It is easy to see that this convergence is not uniform (not even for Bernoulli measures with weights that are not all equal). This theorem was proved for finite partitions in increasing degrees of generality in the years

1948 to 1957 and then was by Carleson [8] and Chung [10] generalized to infinite partitions. Similar results (for finite partitions) for the recurrence and waiting times were later proved by Ornstein and Weiss [24] and Nobel and Wyner [23] respectively. The limiting behavior for recurrence times was generalized in 2002 by Ornstein and Weiss [25] to countably infinite partitions. In this paper we are concerned with the limiting distribution of the information function $I_n(x) = -\log \mu(A_n(x))$ around its mean value.

The statistical properties of I_n are of great interest in information theory where they are connected to the efficiency of compression schemes. Let us also note that in dynamical systems the analog of Shannon–McMillan–Breiman's (SMB) theorem for compact metric spaces is the Brin–Katok local entropy formula [5] which states that for ergodic invariant measures, the exponential decay rate of dynamical balls is almost everywhere equal to the entropy.

There is a large classical body of work on the Central Limit Theorem (CLT) for independent random variables. For dependent random variables the first CLTs are due to Markov (for Markov chains) and Bernstein [2] for random variables that are allowed to have some short range dependency but have to be independent if separated by a suitable time difference (for more than a power of the length n of the partial sums S_n). In 1956, Rosenblatt [34] then introduced the notions of uniform mixing and strong mixing (see below) and proved a CLT for the partial sums S_n of random variables that satisfy the strong mixing property. In [35] he then proved a more general CLT for random variables on systems that satisfy an L^2 norm condition.[a] Around the same time Nagaev [22] proved a convergence theorem for the stable law for strongly mixing systems. His result covers the case of the CLT and formed the basis for Ibragimov's famous 1962 paper [18] in which he proved for finite partitions "a refinement to SMB's theorem" by showing that $I_n(x) = -\log \mu(A_n(x))$ is in the limit lognormally distributed for systems that are strongly mixing and satisfy a regularity condition akin to a Gibbs property. Based on his results and methods, Philipp and Stout [30] proved the almost sure invariance principle for the information function I_n under similar conditions as Ibragimov used (requiring faster decay rates). This result in turn was then used by Kontoyiannis [19] to prove the almost sure invariance principle, CLT and the law of the iterated logarithm LIL for recurrence and waiting times, thus strengthening the result of Nobel and Wyner [23] who showed that for strongly mixing systems (without regularity condition) the exponential growth rate of waiting times equals the metric entropy.

Various improvements and refinements to the CLT for the information function have been successively done mainly for measures that satisfy a genuine Gibbs property. For instance, Collet, Galves and Schmitt [11] in order to prove the lognormal

[a]The map T satisfies an L^2 *norm condition* if $\sup_{f:\mu(f)=0} \frac{\|T^n f\|^2}{\|f\|^2}$ decays exponentially fast as $n \to \infty$. This is a somewhat stronger mixing condition than the strong mixing condition.

distribution of entry times for exponentially ψ-mixing Gibbs measures[b] needed to know that I_n is in the limit lognormally distributed. A more general result is due to Paccaut [26] for maps on the interval where he had to assume some topological covering properties. For some non-uniformly hyperbolic maps on the interval similar results were formulated in [13, 7]. However, all those results use explicitly the Gibbs property of the invariant measure μ to approximate the information function I_n by an ergodic sum and then to invoke standard results on the CLT for sufficiently regular observables (see for instance [16, 20, 9]). (Of course the variance has to be nonzero because otherwise the limiting distribution might not be normal as an example in [11] illustrates.)

Results that do not require the explicit Gibbs characterization of the measure like Kontoyiannis' paper [19], all ultimately rely on the original paper of Ibragimov [18] and require apart from the strong mixing condition the regularity of the Radon–Nikodym derivative of the measure under the local inverse maps. In [17] we went beyond his regularity constraint and proved a CLT with error bounds for the lognormal distribution of the information function for (ψ, f)-mixing systems which included traditional ψ-mixing maps and also equilibrium states for rational maps with critical points in the Julia set.

This paper is significant in two aspects: (i) we allow for the partition to be countably infinite instead of finite and (ii) unlike Ibragimov (and all who followed him) we do not require an L^1-regularity condition for the Radon–Nikodym derivative for local inverses of the map. This condition which was introduced in [18] is the L^1 equivalent of what would otherwise allow a transfer operator approach to analyze the invariant measure and imply the Gibbs property.[c] We moreover prove that the rate of convergence is polynomial (Theorem 2) and the variance is always positive for genuinely infinite partitions.

Let us note that convergence rates for the CLT have previously been obtained by Broise [6] for a large class of expanding maps on the interval for which the Perron–Frobenius operator has a "spectral gap". Similar estimates were obtained by Pène [27] for Gibbs measures for dispersing billiards.

This paper is structured as follows: In Sec. 2, we introduce uniform strong mixing systems and in Sec. 3, we prove the existence of the variance σ^2 of strongly mixing probability measures (Proposition 14) as well as the growth rate of higher order moments (Proposition 15). This is the main part of the proof (note that Ibragimov's regularity condition was previously needed precisely to obtain the variance of the

[b]We say an invariant probability measures μ is *Gibbs* for a potential f with pressure $P(f)$ if there exists a constant $c > 0$ so that

$$\frac{1}{c} \le \frac{\mu(A_n(x))}{e^{f(x)+f(Tx)+\cdots+f(T^{n-1}x)-nP(f)}} \le c$$

for every $x \in \Omega$ and $n = 1, 2, \ldots$.

[c]More precisely, Ibragimov's condition requires that the L^1-norms of the differences $f - f_n$ decay polynomially, where $f = \lim_{n\to\infty} f_n$ and $f_n(x) = \log \mathbb{P}(x_0 \mid x_{-1}x_{-2}\ldots x_{-n})$.

measure). In Sec. 4, we then prove the CLT using Stein's method of exchangeable pairs. In Sec. 5, we prove the Weak Invariance Principle for I_n using the CLT and the convergence rate obtained in Sec. 3.

2. Main Results

Let T be a map on a space Ω and μ a probability measure on Ω. Moreover, let $\mathcal{A}$ be a (possibly infinite) measurable partition of Ω and denote by $\mathcal{A}^n = \bigvee_{j=0}^{n-1} T^{-j}\mathcal{A}$ its *nth join* which is also a measurable partition of Ω for every $n \geq 1$. The atoms of $\mathcal{A}^n$ are called *n-cylinders*. Let us put $\mathcal{A}^* = \bigcup_{n=1}^{\infty} \mathcal{A}^n$ for the collection of all cylinders in Ω and put $|A|$ for the length of a cylinder $A \in \mathcal{A}^*$, i.e. $|A| = n$ if $A \in \mathcal{A}^n$.

We shall assume that $\mathcal{A}$ is generating, i.e. that the atoms of $\mathcal{A}^\infty$ are single points in Ω.

2.1. *Mixing*

Definition 1. We say the invariant probability measure μ is *uniformly strong mixing* if there exists a decreasing function $\psi : \mathbb{N} \to \mathbb{R}^+$ which satisfies $\psi(\Delta) \to 0$ as $\Delta \to \infty$ such that

$$\left| \sum_{(B,C)\in S} (\mu(B \cap C) - \mu(B)\mu(C)) \right| \leq \psi(\Delta)$$

for every subset S of $\mathcal{A}^n \times T^{-\Delta-n}\mathcal{A}^m$ and every $n, m, \Delta > 0$.

Various kinds of mixing:[d] In the following list of different mixing properties U is always in the σ-algebra generated by $\mathcal{A}^n$ and V lies in the σ-algebra generated by $\mathcal{A}^*$ (see also [12]). The limiting behavior is as the length of the "gap" $\Delta \to \infty$:

(1) *ψ-mixing*: $\sup_n \sup_{U,V} \left| \frac{\mu(U \cap T^{-\Delta-n}V)}{\mu(U)\mu(V)} - 1 \right| \to 0.$

(2) *Left ϕ-mixing*: $\sup_n \sup_{U,V} \left| \frac{\mu(U \cap T^{-\Delta-n}V)}{\mu(U)} - \mu(V) \right| \to 0.$

(3) *Strong mixing* [34, 18] (also called α-mixing): $\sup_n \sup_{U,V} |\mu(U \cap T^{-\Delta-n}V) - \mu(U)\mu(V)| \to 0.$

(4) *Uniform mixing* [34, 35]: $\sup_n \sup_{U,V} \left| \frac{1}{k} \sum_{j=1}^{k} \mu(U \cap T^{-n-j}V) - \mu(U)\mu(V) \right| \to 0$ as $k \to \infty$.

One can also have *right ϕ-mixing* when $\sup_n \sup_{U,V} |\frac{\mu(U \cap T^{-\Delta-n}V)}{\mu(V)} - \mu(U)| \to 0$ as $\Delta \to \infty$. Clearly ψ-mixing implies all the other kinds of mixing. The next strongest

[d] Here we adopt probabilistic terminology which differs from the one used in the dynamical systems community.

mixing property is ϕ-mixing, then comes strong mixing and uniform mixing is the weakest. The uniform strong mixing property is stronger than the strong mixing property but is implied by the dynamical ϕ-mixing property as we will see in Lemma 6. In fact if μ is strong mixing then the sets S in Definition 1 have to be of product form.

Let us note that μ has the weak Bernoulli property (with respect to the partition $\mathcal{A}$) if for every $\varepsilon > 0$ there exists an $N(\varepsilon)$ such that

$$\sum_{B\in\mathcal{A}^n} |\mu(B\cap C) - \mu(B)\mu(C)| \le \varepsilon\mu(C)$$

for every $C \in T^{-\Delta-n}\mathcal{A}^m$, $\Delta > N$ and $n, m \in \mathbb{N}$ (see e.g. [28]). Summing over C shows that the weak Bernoulli property implies the uniform strong mixing property where the rate ψ depends upon how fast the function $N(\varepsilon)$ grows as ε goes to zero.

For a partition $\mathcal{A}$ we have the (nth) *information function* $I_n(x) = -\log\mu(A_n(x))$, where $A_n(x)$ denotes the unique n-cylinder that contains the point $x \in \Omega$, whose moments are

$$K_w(\mathcal{A}) = \sum_{A\in\mathcal{A}} \mu(A)|\log\mu(A)|^w = \mathbb{E}(I_n^w),$$

$w \ge 0$ not necessarily integer. (For $w = 1$ one traditionally writes $H(\mathcal{A}) = K_1(\mathcal{A}) = \sum_{A\in\mathcal{A}} -\mu(A)\log\mu(A)$.) If $\mathcal{A}$ is finite then $K_w(\mathcal{A}) < \infty$ for all w. For infinite partitions the theorem of Shannon–McMillan–Breiman requires that $H(\mathcal{A})$ be finite [8, 10]. In order to prove that the information function is lognormally distributed we will require a larger than fourth moment $K_w(\mathcal{A})$ for some $w > 4$ (not necessarily integer) be finite.

2.2. *Results*

For $x \in \Omega$ we denote $A_n(x)$ the n-cylinder in $\mathcal{A}^n$ which contains the point x. We are interested in the limiting behavior of the distribution function

$$\Xi_n(t) = \mu\left(\left\{x \in \Omega : \frac{-\log\mu(A_n(x)) - nh}{\sigma\sqrt{n}} \le t\right\}\right)$$

for real-valued t and a suitable positive σ, where h is the metric entropy of μ. The Central Limit Theorem states that this quantity converges to the normal distribution $N(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-s^2/2}ds$ as n goes to infinity if there exists a suitable σ which is positive. Our main result is the following theorem:

Theorem 2. *Let μ be a uniformly strong mixing probability measure on Ω with respect to a countably finite, measurable and generating partition $\mathcal{A}$ which satisfies $K_w(\mathcal{A}) < \infty$ for some $w > 4$. Assume that ψ decays at least polynomially with power $> 8 + \frac{24}{w-4}$.*

Then

(I) *The limit*

$$\sigma^2 = \lim_{n\to\infty} \frac{K_2(\mathcal{A}^n) - H^2(\mathcal{A}^n)}{n}$$

exists and defines the variance of μ. *Moreover, if the partition is infinite then* σ *is strictly positive.*

(II) *If* $\sigma > 0$:

$$|\Xi_n(t) - N(t)| \le C_0 \frac{1}{n^\kappa}$$

for all t *and all*

(i) $\kappa < \frac{1}{10} - \frac{3}{5}\frac{w}{(p+2)(w-2)+6}$ *if* ψ *decays polynomially with power* p,

(ii) $\kappa < \frac{1}{10}$ *if* ψ *decays faster than any power.*

The variance σ^2 is determined in Proposition 14 and essentially only requires finiteness of the second moment $K_2(\mathcal{A})$. In order to obtain the rate of convergence one usually needs a higher than second moment of I_n. Since we use Stein's method we require the fourth moment to be finite (unlike in [17] where for finite partitions and (ψ, f)-mixing measures we only needed bounds on the third moment).

Throughout the paper we shall assume that $K_w(\mathcal{A}) < \infty$ for some finite $w > 4$. The case in which w can be arbitrarily large (e.g. for finite partitions) is done with minor modifications and yields the obvious result for the rate of convergence. For simplicity's sake we assume in the proofs that the decay rate of ψ is polynomial at some finite power p. The case of hyper-polynomial decay can be traced out with minor modifications and yields the stated result.

If the partition $\mathcal{A}$ is finite then $K_w(\mathcal{A}) < \infty$ for all w and we obtain the following corollary:

Corollary 3. *Let* μ *be a uniformly strong mixing probability measure on* Ω *with respect to a finite, measurable and generating partition* $\mathcal{A}$ *and* ψ *decays at least polynomially with power* $> 8 + \frac{24}{w-4}$.

Then

(I) *The limit* $\sigma^2 = \lim_{n\to\infty} \frac{1}{n}(K_2(\mathcal{A}^n) - H^2(\mathcal{A}^n))$ *exists (variance of* μ*).*

(II) *If* $\sigma > 0 : \Xi_n(t) = N(t) + \mathcal{O}(n^{-\kappa})$ *for all* t *and*

$$\begin{cases} \kappa < \dfrac{1}{10} - \dfrac{3}{5(p+2)} & \text{if } \psi(\Delta) = \mathcal{O}(\Delta^{-p}),\ \Delta \in \mathbb{N}, \\[2ex] \kappa < \dfrac{1}{10} & \text{if } \psi \text{ decays hyper polynomially.} \end{cases}$$

By a result of Petrov [29] we now obtain the Law of the Iterated Logarithm from Theorem 2 by virtue of the error bound (better than $\frac{1}{(\log n)^{1-\varepsilon}}$ (some $\varepsilon > 0$) which are the ones required in [29]).

Corollary 4. *Under the assumptions of Theorem* 2,

$$\limsup_{n\to\infty} \frac{I_n(x) - nh}{\sigma\sqrt{2n\log\log n}} = 1$$

almost everywhere.

A similar statement is true for the lim inf where the limit is then equal to -1 almost everywhere.

Based on the CLT we also get the weak invariance principle (WIP) (see Sec. 4). Recently, there has been a great interest in the WIP in relation to mixing properties of dynamical systems. For instance it has been obtained for a large class of observables and for a large class of dynamical systems by Chernov in [9]. Other recent results are [14, 13, 30]. Those results however are typically for sums of sufficiently regular observables. Here we prove the WIP for $I_n(x)$.

Theorem 5. *Under the assumption of Theorem* 2 *the information function* I_n *satisfies the Weak Invariance Principle* (*provided the variance* σ^2 *is positive*).

2.3. *Examples*

(I) Bernoulli shift: Let Σ be the full shift space over the infinite alphabet $\mathbb{N}$ and let μ be the Bernoulli measure generated by the positive weights $p_1, p_2, \dots$ ($\sum_j p_j = 1$). The entropy is then $h(\mu) = \sum_j p_j|\log p_j|$ and since $K_2(\mathcal{A}) = \sum_i p_i \log^2 p_i = \frac{1}{2}\sum_{i,j} p_i p_j(\log^2 p_i + \log^2 p_j)$ we obtain that the variance is given by the following expression which is familiar from finite alphabet Bernoulli shifts:

$$\sigma^2 = K_2(\mathcal{A}) - h(\mu)^2 = \frac{1}{2}\sum_{i,j} p_i p_j \log^2 \frac{p_i}{p_j}.$$

We have used that the partition $\mathcal{A}$ is given by the cylinder sets whose first symbols are fixed. Here we naturally assume that $\sum_i p_i \log^2 p_i < \infty$. If moreover $\sum_i p_i \log^4 p_i < \infty$ then

$$\mathbb{P}\left(\frac{-\log\mu(A_n(x)) - nh}{\sigma\sqrt{n}} \le t\right) = N(t) + \mathcal{O}(r^{-1/4})$$

with exponent $\frac{1}{4}$ which is a well-known result for unbounded iid random variables. With other techniques one can however weaken the moment requirement in this case.

(II) Markov shift: Again let Σ be the shift space over the infinite alphabet $\mathbb{N}$ and μ the Markov measure generated by an infinite probability vector $\mathbf{p} = (p_1, p_2, \dots)$ ($p_j > 0$, $\sum_j p_j = 1$) and an infinite stochastic matrix P ($\mathbf{p}P = \mathbf{p}$, $P\mathbf{1} = \mathbf{1}$). The partition $\mathcal{A}$ is again the partition of single element cylinder sets. If $\mathbf{x} = x_1x_2\cdots x_n$ is a word of length n (we write $\mathbf{x} \in \mathcal{A}^n$) then the measure of its cylinder set is $\mu(\mathbf{x}) = p_{x_1}P_{x_1x_2}P_{x_2x_3}\cdots P_{x_{n-1}x_n}$. The metric entropy is

$h(\mu) = \sum_{i,j} -p_i P_{ij} \log P_{ij}$ [38] and the variance [33, 39] (see also the Appendix) is

$$\sigma^2 = \frac{1}{2} \sum_{ijk\ell} p_i P_{ij} p_k P_{k\ell} \log^2 \frac{P_{ij}}{P_{k\ell}} + 4 \sum_{k=2}^{\infty} \sum_{\mathbf{x} \in \mathcal{A}^k} \mu(\mathbf{x}) (\log P_{x_1 x_2} \log P_{x_{k-1} x_k} - h^2).$$

(III) Equilibrium states: The measure μ is an equilibrium state for the potential function f if it realizes the supremum in the variational principle $P(f) = \sup_\nu (h(\nu) + \int f d\nu)$ where the supremum is over all invariant probability measures ν and $P(f)$ is the pressure of f (see e.g. [38]). If f is Hölder continuous and T is an Axiom A map, then the CLT has been studied a great deal in particular for finite partitions since standard techniques for sums of random variables can be applied (see e.g. [6, 16, 20, 27]). The existence of equilibrium states for infinite alphabet Markov shifts was proven contemporaneously by Bressaud [4] and Sarig [36] where additional conditions on the potential had to be imposed to ensure that the transfer operator is bounded and the pressure is finite. Note that (I) and (II) are special cases of equilibrium states.

3. Variance and Higher Moments

3.1. *Some basic properties*

Let us begin by showing that the uniform strong mixing property is implied by the ϕ-mixing property.

Lemma 6. *ϕ-mixing implies uniformly strong mixing.*

Proof. Let μ be a left ϕ-mixing probability measure (the right ϕ-mixing case is done in the same way). That means, there exists a decreasing $\phi(\Delta) \to 0$ as $\Delta \to \infty$ so that

$$|(\mu(B \cap C) - \mu(B)\mu(C))| \le \phi(\Delta)\mu(B)$$

for every C in the σ-algebra generated by $\mathcal{C} = T^{-n-\Delta}\mathcal{A}^m$ and every cylinder $B \in \mathcal{B} = \mathcal{A}^n$ for all n and Δ. Let $S \subset \mathcal{B} \times \mathcal{C}$ and put S_B for the intersection of $\{B\} \times \mathcal{C}$ with S. Then $|(\mu(B \cap S_B) - \mu(B)\mu(S_B))| \le \phi(\Delta)\mu(B)$ and

$$\left| \sum_{(B,C)\in S} (\mu(B \cap C) - \mu(B)\mu(C)) \right|$$
$$\le \sum_{B\in\mathcal{B}} |(\mu(B \cap S_B) - \mu(B)\mu(S_B))| \le \sum_{B\in\mathcal{B}} \phi(\Delta)\mu(B) \le \phi(\Delta)$$

implies that μ is uniformly strong mixing with $\psi = \phi$. □

The following estimate has previously been shown for ψ-mixing measures (in which case they are exponential) in [15] and for ϕ-mixing measures in [1]. Denote by $A_n(x)$ the atom in $\mathcal{A}^n$ $(n = 1, 2, \ldots)$ which contains the point $x \in \Omega$. (Abadi [1] also showed that in case (II) the decay cannot in general be exponential.)

Lemma 7. *Let μ be uniform strong mixing. Then there exists a constant C_1 so that for all $A \in \mathcal{A}_n$, $n = 1, 2, \ldots$:*

(I) $\mu(A) \le C_1 n^{-p}$ *if ψ is polynomially decreasing with exponent $p > 0$;*
(II) $\mu(A) \le C_1 \theta^{\sqrt{n}}$ *for some $\theta \in (0, 1)$ if ψ is exponentially decreasing.*

Proof. Fix $m \ge 1$ so that $a = \max_{A \in \mathcal{A}^m} \mu(A)$ is less than $\frac{1}{4}$ and let $\Delta_1, \Delta_2, \ldots$ be integers which will be determined below. We put $n_j = jm + \sum_{i=1}^{j-1} \Delta_i$ (put $\Delta_0 = 0$) and for $x \in \Omega$ let $B_j = A_m(T^{n_{j-1}+m}x)$ and put $C_k = \bigcap_{j=1}^k B_j$. Then $A_{n_k}(x) \subset C_k$ and

$$\mu(C_{k+1}) = \mu(C_k \cap B_{k+1}) = \mu(C_k)\mu(B_{k+1}) + \rho(C_k, B_{k+1}),$$

where the remainder term $\rho(C_k, B_{k+1})$ is by the mixing property in absolute value bounded by $\psi(\Delta_k)$. Now we choose Δ_j so that $\psi(\Delta_j) \le a^{\frac{j}{2}+1}$. Then $\mu(C_{k+1}) \le \mu(C_k)a + a^{\frac{k}{2}+1}$ implies that $\mu(C_k) \le c_0 a^{\frac{k}{2}}$ (as $\sqrt{a} \le \frac{1}{2}$) for some $c_0 > 0$.

(I) If ψ decays polynomially with power p, i.e. $\psi(t) \le c_1 t^{-p}$, then the condition $\psi(\Delta_j) \le a^{\frac{j}{2}+1}$ is satisfied if we put $\Delta_j = [c_2 a^{-\frac{j}{2p}}]$ for a suitable constant $c_2 > 0$. Consequently, $n_k \le c_3 a^{-\frac{k}{2p}}$ $(c_3 \ge 1)$ and therefore $k \ge 2p \frac{\log n_k}{|\log a|}$. Hence

$$\mu(A_{n_k}(x)) \le c_0 a^{\frac{k}{2}} \le c_0 a^{p\frac{\log n_k}{|\log a|}} \le c_4 n_k^{-p}$$

and from this one obtains $\mu(A_n(x)) \le c_5 n^{-p}$ for all integers n (and some larger constant c_5).

(II) If ψ decays exponentially, i.e. $\psi(t) \le c_6 \vartheta^t$ for some $\vartheta \in (0, 1)$, then we choose $\Delta_j = [\frac{j}{2} \frac{\log a}{\log \vartheta}]$ and obtain $n_k \le mk + c_7 k^2$, which gives us $k \ge c_8 \sqrt{n_k}$ $(c_8 > 0)$ and the stretched exponential decay of the measure of cylinder sets:

$$\mu(A_n(x)) \le c_9 a^{c_8 \sqrt{n}}.$$

Now put $\theta = a^{c_8}$. □

3.2. *The information function and mixing properties*

The metric entropy h for the invariant measure μ is $h = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{A}^n)$, where $\mathcal{A}$ is a generating partition of Ω (cf. [21]), provided $H(\mathcal{A}) < \infty$. For $w \ge 1$ put $\eta_w(t) = t \log^w \frac{1}{t}$ $(\eta_w(0) = 0)$. Then

$$K_w(\mathcal{B}) = \sum_{B \in \mathcal{B}} \mu(B) |\log \mu(B)|^w = \sum_{B \in \mathcal{B}} \eta_w(\mu(B))$$

for partitions $\mathcal{B}$. Similarly one has the conditional quantity ($\mathcal{C}$ is a partition):

$$K_w(\mathcal{C}|\mathcal{B}) = \sum_{B \in \mathcal{B}, C \in \mathcal{C}} \mu(B) \eta_w\left(\frac{\mu(B \cap C)}{\mu(B)}\right) = \sum_{B,C} \mu(B \cap C) \left|\log \frac{\mu(B \cap C)}{\mu(B)}\right|^w.$$

Lemma 8. ([17]) *For any two partitions $\mathcal{B}, \mathcal{C}$ for which $K_w(\mathcal{B}), K_w(\mathcal{C}) < \infty$ and $\mu(C) \le e^{-w}\, \forall\, C \in \mathcal{C}$:*

(i) $K_w(\mathcal{C}|\mathcal{B}) \le K_w(\mathcal{C})$,
(ii) $K_w(\mathcal{B} \vee \mathcal{C})^{1/w} \le K_w(\mathcal{C}|\mathcal{B})^{1/w} + K_w(\mathcal{B})^{1/w}$,
(iii) $K_w(\mathcal{B} \vee \mathcal{C})^{1/w} \le K_w(\mathcal{C})^{1/w} + K_w(\mathcal{B})^{1/w}$.

Proof. (i) Since $\eta_w(t)$ is convex and increasing on $[0, e^{-w}]$ and decreasing to zero on $(e^{-w}, 1]$ we have $\sum_i x_i \eta_w(\alpha_i) \le \eta_w\left(\sum_i x_i \alpha_i\right)$ for weights $x_i \ge 0$ ($\sum_i x_i = 1$) and numbers $\alpha_i \in [0,1]$ which satisfy $\sum_i x_i \alpha_i \le e^{-w}$. Hence

$$K_w(\mathcal{C}|\mathcal{B}) = \sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B)\eta_w\left(\frac{\mu(B\cap C)}{\mu(B)}\right) \le \sum_C \eta_w\left(\sum_B \mu(B)\frac{\mu(B\cap C)}{\mu(B)}\right)$$
$$= \sum_C \eta_w\left(\mu(C)\right) = K_w(\mathcal{C}).$$

(ii) The second statement follows from Minkowski's inequality on L^w-spaces:

$$K_w(\mathcal{B}\vee\mathcal{C})^{\frac{1}{w}} = \left(\sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)|\log\mu(B\cap C)|^w\right)^{\frac{1}{w}}$$
$$\le \left(\sum_{B,C}\mu(B\cap C)\left|\log\frac{\mu(B\cap C)}{\mu(B)}\right|^w\right)^{\frac{1}{w}} + \left(\sum_{B,C}\mu(B\cap C)|\log\mu(B)|^w\right)^{\frac{1}{w}}$$
$$= K_w(\mathcal{C}|\mathcal{B})^{\frac{1}{w}} + K_w(\mathcal{B})^{\frac{1}{w}}.$$

(iii) This follows from (ii) and (i). □

Corollary 9. *Let $w \ge 1$ and $\mathcal{A}$ so that $K_w(\mathcal{A}) < \infty$ and $\mu(A) \le e^{-w}\,\forall\, A \in \mathcal{A}$. Then there exists a constant C_2 (depending on w) so that for all n*

$$K_w(\mathcal{A}^n) \le C_2 n^w.$$

Proof. We want to use Lemma 8(iii) to show that the sequence $a_n = K_w(\mathcal{A}^n)^{1/w}$, $n = 1, 2, \ldots$, is subadditive. The hypothesis of Lemma 8 is satisfied since $\mu(A) \le e^{-w}$ for all $A \in \mathcal{A}$. We thus obtain $K_w(\mathcal{A}^{n+m})^{\frac{1}{w}} \le K_w(\mathcal{A}^n)^{\frac{1}{w}} + K_w(\mathcal{A}^m)^{\frac{1}{w}}$ for all $n, m \ge 1$ and therefore subadditivity of the sequence a_n. Since by assumption $K_w(\mathcal{A}) < \infty$ we get that the limit $\lim_{n\to\infty}\frac{1}{n}K_w(\mathcal{A}^n)^{1/w}$ exists, is finite and equals the inf (see e.g. [38]). □

The function I_n has expected value $\mathbb{E}(I_n) = H(\mathcal{A}^n)$, for which we also write H_n, and variance $\sigma_n^2 = \sigma^2(I_n) = K_2(\mathcal{A}^n) - H_n^2$. In general, if $\mathcal{B}$ is a partition then we write $\sigma^2(\mathcal{B}) = K_2(\mathcal{B}) - H^2(\mathcal{B})$ and similarly for the conditional variance

$\sigma^2(\mathcal{C}|\mathcal{B})$. Let us define the function $J_\mathcal{B}$ by $J_\mathcal{B}(B) = -\log\mu(B) - H(\mathcal{B})$ $(B \in \mathcal{B})$ then $\sigma^2(\mathcal{B}) = \sum_{B\in\mathcal{B}} \mu(B)J_\mathcal{B}(B)^2$ and $\int J_\mathcal{B} d\mu = 0$. For two partitions $\mathcal{B}$ and $\mathcal{C}$ we put

$$J_{\mathcal{C}|\mathcal{B}}(B \cap C) = \log \frac{\mu(B)}{\mu(B \cap C)} - H(\mathcal{C}|\mathcal{B})$$

for $(B, C) \in \mathcal{B} \times \mathcal{C}$. (This means $J_{\mathcal{C}|\mathcal{B}} = J_{\mathcal{B}\vee\mathcal{C}} - J_\mathcal{B}$ and $\sigma(\mathcal{C}|\mathcal{B}) = \sigma(J_{\mathcal{C}|\mathcal{B}})$.)

Lemma 10. *Let $\mathcal{B}$ and $\mathcal{C}$ be two partitions. Then*

$$\sigma(\mathcal{B} \vee \mathcal{C}) \le \sigma(\mathcal{C}|\mathcal{B}) + \sigma(\mathcal{B}).$$

Proof. This follows from Minkowski's inequality

$$\sigma(\mathcal{B} \vee \mathcal{C}) = \sqrt{\mu(J_{\mathcal{C}|\mathcal{B}} + J_\mathcal{B})^2} \le \sqrt{\mu(J^2_{\mathcal{C}|\mathcal{B}})} + \sqrt{\mu(J^2_\mathcal{B})} = \sigma(\mathcal{C}|\mathcal{B}) + \sigma(\mathcal{B}). \qquad \square$$

As a consequence of Lemma 8(i) one also has $K_w(\mathcal{B}\vee\mathcal{C}|\mathcal{B}) = K_w(\mathcal{C}|\mathcal{B}) \le K_w(\mathcal{C})$ which in particular implies $\sigma(\mathcal{B} \vee \mathcal{C}|\mathcal{B}) = \sigma(\mathcal{C}|\mathcal{B}) \le \sqrt{K_2(\mathcal{C})}$. As before we put $\rho(B, C) = \mu(B \cap C) - \mu(B)\mu(C)$ (and in the following we often write $\mathcal{B} = \mathcal{A}^n$ and $\mathcal{C} = T^{-\Delta-n}\mathcal{A}^n$ for integers n, Δ).

The following technical lemma is central to get the variance of μ and bounds on the higher moments of $J_n = I_n - H_n$.

Lemma 11. *Let μ be uniformly strong mixing and assume that $K_w(\mathcal{A}) < \infty$ and $\mu(A) \le e^{-w}\,\forall\, A \in \mathcal{A}$ for some $w \ge 1$. Then for every $\beta > 1$ and $a \in [0, w)$ there exists a constant C_4 so that*

$$\sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B \cap C) \left|\log\left(1 + \frac{\rho(B, C)}{\mu(B)\mu(C)}\right)\right|^a$$
$$\le C_4(\psi(\Delta)(m + n)^{(1+a)\beta} + (m + n)^{a\beta-w(\beta-1)})$$

for $\Delta < \min(n, m)$ and for all $n = 1, 2, \ldots$. (As before $\mathcal{B} = \mathcal{A}^m$, $\mathcal{C} = T^{-\Delta-m}\mathcal{A}^n$).

Proof. Let m, n and Δ be as in the statement and put

$$\mathcal{L}_\ell = \left\{(B, C) \in \mathcal{B} \times \mathcal{C} : 2^{\ell-1} < 1 + \frac{\rho(B, C)}{\mu(B)\mu(C)} \le 2^\ell\right\}$$

$\ell \in \mathbf{Z}$. Using the strong mixing property we obtain

$$\sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B \cap C) \left|\log\left(1 + \frac{\rho(B, C)}{\mu(B)\mu(C)}\right)\right|^a = \sum_{\ell=-\infty}^{\infty} L_\ell(|\ell| + \mathcal{O}(1))^a,$$

where $L_\ell = \sum_{(B,C)\in\mathcal{L}_\ell} \mu(B \cap C)$. Since $\rho(B, C) = \mathcal{O}(1)(2^\ell - 1)\mu(B)\mu(C)$ we get $\mathcal{O}(\psi(\Delta)) = \sum_{(B,C)\in\mathcal{L}_\ell} \rho(B, C) = \mathcal{O}(1)(2^\ell - 1)L^\times_\ell$ where $L^\times_\ell = \sum_{(B,C)\in\mathcal{L}_\ell} \mu(B)\mu(C)$. Hence for $\ell > 0$ one obtains $L^\times_\ell = \mathcal{O}(\psi(\Delta))2^{-\ell}$ and if $\ell < 0$

then $L_\ell^\times = \mathcal{O}(\psi(\Delta))$. Also note that if $\ell = 0$ then $|\log(1+\frac{\rho(B,C)}{\mu(B)\mu(C)})| = \mathcal{O}(\frac{\rho(B,C)}{\mu(B)\mu(C)})$ and

$$\sum_{(B,C)\in\mathcal{L}_0} \mu(B\cap C)\left|\log\left(1+\frac{\rho(B,C)}{\mu(B)\mu(C)}\right)\right|^a = \mathcal{O}(1)\sum_{(B,C)\in\mathcal{L}_0}\rho(B,C) = \mathcal{O}(\psi(\Delta)).$$

We separately estimate (i) for $\ell \geq 1$ and (ii) for $\ell \leq -1$:

(i) Since $\mu(B\cap C) = (1+\frac{\rho(B,C)}{\mu(B)\mu(C)})\mu(B)\mu(C)$ we get for $\ell \geq 1$:

$$2^{\ell-1}L_\ell^\times = \sum_{(B,C)\in\mathcal{L}_\ell}\mu(B)\mu(C)2^{\ell-1} \leq L_\ell \leq \sum_{(B,C)\in\mathcal{L}_\ell}\mu(B)\mu(C)2^\ell = 2^\ell L_\ell^\times.$$

Thus

$$\sum_{\ell=1}^{(m+n)^\beta}\ell^a L_\ell \leq \sum_{\ell=1}^{(m+n)^\beta}\ell^a 2^\ell L_\ell^\times = \sum_{\ell=1}^{(m+n)^\beta}\ell^a\frac{2^\ell}{2^\ell-1}\mathcal{O}(\psi(\Delta))$$
$$\leq c_1\psi(\Delta)(m+n)^{(1+a)\beta}.$$

For $\ell > (m+n)^\beta$ we use that $\mu(B\cap C) \geq 2^{\ell-1}\mu(B)\mu(C)$ on $\mathcal{L}_\ell$ which implies $\mu(B)\mu(C) \leq 2^{1-\ell}$ and $\mu(B\cap C) \leq \min(\mu(B),\mu(C)) \leq 2^{-\frac{\ell-1}{2}}$. Hence, on $\mathcal{L}_\ell$ one has $|\log\mu(B\cap C)| \geq (\ell-1)\log\sqrt{2}$. Similarly to the previous lemma put

$$D_k = \bigcup_{(B,C)\in\mathcal{B}\times\mathcal{C},\,k-1<|\log\mu(B\cap C)|\leq k}(B\cap C)$$

and use Corollary 9 to get (as $K_w(\mathcal{A}^{n+m+\Delta}) \geq K_w(\mathcal{B}\vee\mathcal{C})$)

$$C_2(w)(n+m+\Delta)^w \geq K_w(\mathcal{B}\vee\mathcal{C}) \geq \sum_{k=1}^\infty \mu(D_k)(k-1)^w$$
$$\geq c_2(n+m)^{\beta(w-a)}\sum_{k=[(n+m)^\beta]+1}^\infty \mu(D_k)k^a.$$

We thus obtain (using that $\Delta < \min(n,m)$)

$$\sum_{\ell=(n+m)^\beta}^\infty \ell^a L_\ell \leq \frac{1}{\log\sqrt{2}}\sum_{(B,C)\in\mathcal{B}\times\mathcal{C},\,|\log\mu(B\cap C)|\geq(n+m)^\beta}|\log\mu(B\cap C)|^a\mu(B\cap C)$$
$$\leq \frac{1}{\log\sqrt{2}}\sum_{k=[(n+m)^\beta]+1}^\infty k^a\mu(D_k)$$
$$\leq c_3\frac{(n+m)^{a\beta}}{(n+m)^{(\beta-1)w}}$$

for some c_3 (which depends on w).

(ii) For negative values of ℓ we use $L_\ell \leq 2^\ell L_\ell^\times \leq c_4 2^\ell \psi(\Delta)$ which gives

$$\sum_{\ell=-\infty}^{0} |\ell|^a L_\ell \leq c_4 \sum_{\ell=0}^{\infty} \ell^a 2^{-\ell}\psi(\Delta) \leq c_5\psi(\Delta).$$

Combining (i) and (ii) yields

$$\sum_{\ell=-\infty}^{\infty} L_\ell(|\ell| + \mathcal{O}(1))^a \leq (c_1 + c_4)\psi(\Delta)(m+n)^{(1+a)\beta} + c_3(m+n)^{a\beta - w(\beta-1)}$$

which concludes the proof. □

3.3. *Entropy*

The main purpose of this section is to obtain rates of convergence for the entropy (Lemma 13).

Lemma 12. *Under the assumptions of Lemma* 11 *for every* $\beta > 1$ *there exists a constant* C_5 *so that for all* n:

$$|H(\mathcal{B} \vee \mathcal{C}) - (H(\mathcal{B}) + H(\mathcal{C}))| \leq C_5(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w}),$$

where $\mathcal{B} = \mathcal{A}^n, \mathcal{C} = T^{-\Delta-n}\mathcal{A}^n$.

Proof. Using the uniform strong mixing property $\mu(B\cap C) = \mu(B)\mu(C) + \rho(B, C)$ we obtain

$$\begin{aligned} H(\mathcal{B} \vee \mathcal{C}) &= \sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)\log\frac{1}{\mu(B\cap C)} \\ &= \sum_{B,C} \mu(B\cap C)\left(\log\frac{1}{\mu(B)} + \log\frac{1}{\mu(C)} - \log\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right)\right) \\ &= H(\mathcal{B}) + H(\mathcal{C}) + E, \end{aligned}$$

where by Lemma 11 (with $a = 1$)

$$E = -\sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)\log\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right) = \mathcal{O}(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w}).$$

This proves the lemma. □

Lemma 13. *Under the assumptions of Lemma* 11 *there exists a constant* C_6 *so that* ($H_m = H(\mathcal{A}^m)$)

$$\left|\frac{H_m}{m} - h\right| \leq C_6 \frac{1}{m^\gamma}$$

for all m, *where* $\gamma \in (0, 1 - \frac{2w}{p(w-1)})$ *if* ψ *decays polynomially with power* $p > \frac{2w}{w-1}$ *and* $\gamma \in (0,1)$ *if* ψ *decays faster than polynomially.*

Proof. Let m be an integer. Let $\mathcal{B} = \mathcal{A}^{u-\Delta}$, $\mathcal{C} = T^{-u}\mathcal{A}^{u-\Delta}$ and $\mathcal{D} = T^{-u}\mathcal{A}^{2\Delta}$, then by Lemma 12:

$$H_{2u} = 2H_{u-\Delta} + \mathcal{O}(H_{2\Delta}) + \mathcal{O}(\psi(\Delta)u^{2\beta} + n^{\beta-(\beta-1)w}).$$

If we choose $\delta \in (\frac{2w}{p(w-1)}, 1)$ and put $\beta = \frac{w}{w-1}$ then $\Delta = \mathcal{O}(u^\delta)$ implies that $\psi(\Delta)u^{2\beta} + u^{\beta-(\beta-1)w} = \mathcal{O}(1)$. With $\Delta = [u^\delta]$ we thus obtain $H_{2u} = 2H_u + \mathcal{O}(\Delta) = 2H_u + \mathcal{O}(u^\delta)$ as $H_{2\Delta} = \mathcal{O}(\Delta)$ and $H_{u-\Delta} = H_u + \mathcal{O}(\Delta)$. Iterating this estimate yields the following bound along exponential progression:

$$H_{2^i m} = 2^i H_m + \sum_{j=0}^{i-1} 2^{i-1-j}\mathcal{O}((2^j m)^\delta) = 2^i H_m + \mathcal{O}(m^\delta 2^i).$$

To get bounds for arbitrary (large) integers n we do the following dyadic argument: Let $n = km + r$ where $0 \le r < m$ and consider the binary expansion of: $k = \sum_{i=0}^{\ell} \epsilon_i 2^i$, where $\epsilon_i = 0, 1$ ($\epsilon_\ell = 1$, $\ell = [\log_2 k]$). We also put $k_j = \sum_{i=0}^{j} \epsilon_i 2^i$ ($k_\ell = k$). Obviously $k_j = k_{j-1} + \epsilon_j 2^j \le 2^{j+1}$. If $\epsilon_j = 1$ then we separate the "first" block of length $k_{j-1}m$ from the "second" block of length $2^j m$ by a gap of length $2[(k_{j-1}m)^\delta]$ which we cut away in equal parts from the two adjacent blocks. We thus obtain ($H_0 = 0$)

$$\begin{aligned} H_{mk_j} &= H_{\epsilon_j 2^j m + k_{j-1}m} = H_{\epsilon_j 2^j m} + H_{k_{j-1}m} + \mathcal{O}(\epsilon_j (k_{j-1}m)^\delta) \\ &= H_{\epsilon_j 2^j m} + H_{k_{j-1}m} + \mathcal{O}(\epsilon_j (2^j m)^\delta) \end{aligned}$$

for $j = 0, 1, \ldots, \ell - 1$. Iterating this formula and summing over j yields

$$H_{km} = \sum_{j=0}^{\ell} \epsilon_j (2^j H_m + \mathcal{O}(m^\delta 2^j)) = kH_m + \mathcal{O}(m^\delta 2^\ell).$$

The contribution made by the remainder of length r is easily bounded by

$$|H_n - H_{km}| \le \sigma(\mathcal{A}^n | \mathcal{A}^{km}) \le c_1 r \le c_1 m.$$

Consequently,

$$H_n = kH_m + \mathcal{O}(m^\delta 2^\ell) + \mathcal{O}(m) = kH_m + \mathcal{O}(m^\delta k)$$

as $2^\ell \le k \le 2^{\ell+1}$. Dividing by n and letting n go to infinity ($k \to \infty$) yields

$$h = \liminf_{n\to\infty} \frac{H_n}{n} = \frac{H_m}{m} + \mathcal{O}(m^{\delta-1})$$

for all m large enough. □

3.4. *The variance*

In this section we prove part (I) of Theorem 2 and moreover obtain convergence rates which will be needed to prove part (II) in Sec. 4.

Proposition 14. *Let μ be uniformly strong mixing and assume that $K_w(\mathcal{A}) < \infty$ and $\mu(A) \leq e^{-w}\,\forall\, A \in \mathcal{A}$ for some $w > 2$. Assume that ψ is at least polynomially decaying with power $p > 6 + \frac{8}{w-2}$. Then the limit*

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\sigma^2(\mathcal{A}^n)$$

exists and is finite. Moreover, for every $\eta < \eta_0 = 2\frac{(p-2)(w-2)}{(w-2)(p+2)+8}$ there exists a constant C_7 so that for all $n \in \mathbb{N}$:

$$\left|\sigma^2 - \frac{\sigma^2(\mathcal{A}^n)}{n}\right| \leq \frac{C_7}{n^\eta}.$$

Moreover, if the partition $\mathcal{A}$ is infinite, then σ is strictly positive.

Proof. With $\mathcal{B} = \mathcal{A}^n, \mathcal{C} = T^{-n-\Delta}\mathcal{A}^n$ we have by Lemma 12 $H(\mathcal{B} \vee \mathcal{C}) = H(\mathcal{B}) + H(\mathcal{C}) + \mathcal{O}(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w})$, and get for the variance

$$\begin{aligned}\sigma^2(\mathcal{B} \vee \mathcal{C}) &= \sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B\cap C)\left(\log\frac{1}{\mu(B\cap C)} - H(\mathcal{B}\vee\mathcal{C})\right)^2\\ &= \sum_{B,C}\mu(B\cap C)\bigg(J_{\mathcal{B}}(B) + J_{\mathcal{C}}(C) + \mathcal{O}(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w})\\ &\qquad - \log\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right)\bigg)^2.\end{aligned}$$

By Minkowski's inequality:

$$|\sigma(\mathcal{B}\vee\mathcal{C}) - \sqrt{E(\mathcal{B},\mathcal{C})}| \leq c_1(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w}) + \sqrt{F(\mathcal{B},\mathcal{C})}$$

($c_1 > 0$) where (by Lemma 11 with $a = 2$)

$$\begin{aligned}F(\mathcal{B},\mathcal{C}) &= \sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B\cap C)\log^2\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right)\\ &\leq c_2(\psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w})\end{aligned}$$

and

$$\begin{aligned}E(\mathcal{B},\mathcal{C}) &= \sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B\cap C)(J_{\mathcal{B}}(B) + J_{\mathcal{C}}(C))^2\\ &= \sum_{B,C}\mu(B\cap C)(J_{\mathcal{B}}(B)^2 + J_{\mathcal{C}}(C)^2) + 2G(\mathcal{B},\mathcal{C})\\ &= \sigma^2(\mathcal{B}) + \sigma^2(\mathcal{C}) + 2G(\mathcal{B},\mathcal{C}).\end{aligned}$$

Since $J_{\mathcal{B}}$ and $J_{\mathcal{C}}$ have average zero, the remainder term

$$\begin{aligned} G(\mathcal{B},\mathcal{C}) &= \sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B\cap C)J_{\mathcal{B}}(B)J_{\mathcal{C}}(C) \\ &= \sum_{B,C}(\mu(B)\mu(C)+\rho(B,C))J_{\mathcal{B}}(B)J_{\mathcal{C}}(C) \\ &= \sum_{B,C}\rho(B,C)J_{\mathcal{B}}(B)J_{\mathcal{C}}(C) \end{aligned}$$

which is estimated using Schwarz' inequality is as follows:

$$|G(\mathcal{B},\mathcal{C})| \le \sum_{B,C}|\rho(B,C)|\cdot|J_{\mathcal{B}}(B)|\cdot|J_{\mathcal{C}}(C)| \le \psi(\Delta)\sigma(\mathcal{B})\sigma(\mathcal{C}).$$

Hence

$$\sigma(\mathcal{B}\vee\mathcal{C}) \le \sqrt{\sigma^2(\mathcal{C})+\sigma^2(\mathcal{B})+\psi(\Delta)\sigma(\mathcal{B})\sigma(\mathcal{C})} + c_4\sqrt{\psi(\Delta)n^{3\beta}+n^{2\beta-(\beta-1)w}}. \tag{1}$$

Next we fill the gap of length Δ for which we use Lemma 10 and Corollary 9

$$\begin{aligned} |\sigma(\mathcal{A}^{2n+\Delta})-\sigma(\mathcal{B}\vee\mathcal{C})| &\le \sigma(T^{-n}\mathcal{A}^{\Delta}\,|\,\mathcal{B}\vee\mathcal{C}) \le \sqrt{K_2(T^{-n}\mathcal{A}^{\Delta})} \\ &= \sqrt{K_2(\mathcal{A}^{\Delta})} \le c_5\Delta. \end{aligned}$$

Since by assumption $\psi(\Delta) \le c_6\Delta^{-p}$ for some $p > 6+\frac{8}{w-2}$ we take can $\delta = \frac{4w}{(p+2)(w-2)+8}$ and $\beta = \frac{2+p}{4}\delta$ (in particular $\delta < \frac{1}{2}$). Then, with $\Delta = [n^{\delta}]$ we get $\psi(\Delta)n^{4\beta}+n^{2\beta-(\beta-1)w} \le \Delta^2$. Therefore, as $\sigma(\mathcal{B}) = \sigma(\mathcal{C}) = \sigma_n$ (where $\sigma_n = \sigma(\mathcal{A}^n)$), one has

$$\sigma_{2n+[n^{\delta}]} \le \sqrt{(2+\psi(\Delta))\sigma_n^2+c_7n^{2\delta}} \le \sqrt{2\sigma_n^2+c_7n^{2\delta}},$$

where in the last step we took advantage of the *a priori* estimates from Corollary 9 $\sigma^2(\mathcal{A}^n) \le K_2(\mathcal{A}^n) \le C_2n^2$ and the choice of δ which implies that $\psi(\Delta)n^2 = \mathcal{O}(1)$. Since $2\delta < 1$ one has $\sigma_k^2 \le c_8k$ for all k and some constant c_8. Given n_0 let us put recursively $n_{j+1} = 2n_j+[n_j^{\delta}]$ $(j = 0,1,2,\ldots)$. Then $2^jn_0 \le n_j \le 2^jn_0\prod_{i=0}^{j-1}(1+\frac{1}{2}n_i^{\delta-1})$ where the product is bounded by

$$\prod_{i=0}^{j-1}\left(1+\frac{1}{2}n_i^{\delta-1}\right) \le \prod_{i=0}^{j-1}\left(1+\frac{1}{n_0^{1-\delta}2^{(1-\delta)i+1}}\right) \le \exp\frac{c_9}{n_0^{1-\delta}}.$$

In the same fashion one shows that $|\sigma_{n_{j+1}}^2-2\sigma_{n_j}^2| \le c_7n_j^{2\delta}$ implies

$$2^j\sigma_{n_0}^2\exp-\frac{c_{10}}{n_0^{1-2\delta}} \le \sigma_{n_j}^2 \le 2^j\sigma_{n_0}^2\exp\frac{c_{10}}{n_0^{1-2\delta}}.$$

Hence

$$\frac{2^j\sigma_{n_0}^2}{2^jn_0}\exp-\left(\frac{c_{10}}{n_0^{1-2\delta}}+\frac{c_9}{n_0^{1-\delta}}\right) \le \frac{\sigma_{n_j}^2}{n_j} \le \frac{2^j\sigma_{n_0}^2}{2^jn_0}\exp\frac{c_{10}}{n_0^{1-2\delta}},$$

which simplifies to

$$\frac{\sigma_{n_j}^2}{n_j} = \frac{\sigma_{n_0}^2}{n_0}\left(1 + \mathcal{O}\left(\frac{1}{n_0^{1-2\delta}}\right)\right) = \frac{\sigma_{n_0}^2}{n_0} + \mathcal{O}\left(\frac{1}{n_0^{2-2\delta}}\right). \qquad (2)$$

As $w > 2$ one has $\sigma_{n_0} < \infty$. Taking $\limsup$ as $j \to \infty$ and $n_0 \to \infty$ shows that the limit $\sigma^2 = \lim_n \frac{\sigma_n^2}{n}$ exists and satisfies moreover $|\sigma^2 - \frac{\sigma_n^2}{n}| \le C_7 n^{-(2-2\delta)}$ for some C_7. Now we obtain the statement in the proposition for all $\eta < 2 - 2\delta = 2\frac{(p-2)(w-2)}{(p+2)(w-2)+8}$.

In order to prove the last statement of the proposition let $\mathcal{A}$ be an infinite partition. If we choose n_0 large enough so that the error term $\mathcal{O}(n_0^{-(1-2\delta)})$ in Eq. (2) is $< \frac{1}{2}$, then $\sigma_{n_j}^2 > \frac{1}{2} n_j \sigma_{n_0}^2$ for all j. Since

$$\begin{aligned}\sigma_{n_0}^2 &= \sum_{A\in\mathcal{A}^{n_0}} \mu(A)\log^2\mu(A) - \sum_{A,B\in\mathcal{A}^{n_0}} \mu(A)\mu(B)\log\mu(A)\log\mu(B)\\ &= \frac{1}{2}\sum_{A,B}\mu(A)\mu(B)(\log^2\mu(A) + \log^2\mu(B)) - \sum_{A,B}\mu(A)\mu(B)\log\mu(A)\log\mu(B)\\ &= \frac{1}{2}\sum_{A,B\in\mathcal{A}^{n_0}}\mu(A)\mu(B)\log^2\frac{\mu(A)}{\mu(B)}\end{aligned}$$

we conclude that $\sigma_{n_0}^2 > 0$. Hence $\sigma^2 = \lim_n \frac{\sigma_n^2}{n}$ is strictly positive. □

Remarks. (i) It is well known that for finite partitions the measure has variance zero if it is a Gibbs state for a potential which is a coboundary.

(ii) This proposition implies in particular that the limit $\lim_{n\to\infty} \frac{1}{n^2} K_2(\mathcal{A}^n)$ exists and is equal to h^2.

(iii) An application of Chebychev's inequality gives the large deviation type estimate ($\sigma_n = \sigma(J_n)$)

$$\mathbb{P}\left(\frac{1}{n}J_n(x) \ge t\right) \le \frac{\sigma_n^2}{n^2t^2} = \mathcal{O}\left(\frac{1}{nt^2}\right).$$

3.5. *Higher order moments*

In the proof of Theorem 2 part (II) we will need estimates on the third and fourth moments of J_n. We first estimate the fourth moment and then use Hölder's inequality to bound the third moment. Denote by

$$M_w(\mathcal{B}) = \sum_{B\in\mathcal{B}} \mu(B)|J_{\mathcal{B}}(B)|^w,$$

the wth (absolute) moment of the function $J_{\mathcal{B}}$. By Minkowski's inequality

$$M_4^{\frac{1}{4}}(\mathcal{B}\vee\mathcal{C}) = \sqrt[4]{\mu(J_{\mathcal{C}|\mathcal{B}} + J_{\mathcal{B}})^4} \le \sqrt[4]{\mu(J_{\mathcal{C}|\mathcal{B}}^4)} + \sqrt[4]{\mu(J_{\mathcal{B}}^4)} = M_4^{\frac{1}{4}}(\mathcal{C}|\mathcal{B}) + M_4^{\frac{1}{4}}(\mathcal{B}),$$

where $M_w(\mathcal{C}|\mathcal{B}) = \sum_{B\in\mathcal{B},C\in\mathcal{C}} \mu(B\cap C)|J_{\mathcal{C}|\mathcal{B}}(B\cap C)|^w$ are the conditional moments. It follows from Corollary 9 that the absolute moments for the joins $\mathcal{A}^n$ can roughly be bounded by $M_w(\mathcal{A}^n) \le K_w(\mathcal{A}^n) \le C_2 n^w$. This estimate however is useless to prove Theorem 2 and the purpose of the next proposition is to reduce the exponent w to $\frac{1}{2}w$ in the cases $w = 3, 4$. One can of course get these improved estimates also for w larger than 4 (as long as $K_w(\mathcal{A}) < \infty$) but we do not need those higher order moments here.

Proposition 15. *Let μ be uniformly strong mixing and assume that $K_w(\mathcal{A}) < \infty$ and $\mu(A) \le e^{-w}\,\forall\, A \in \mathcal{A}$ for some $w > 4$. Also assume that ψ decays at least polynomially with power $> 8 + \frac{24}{w-4}$. Then there exists a constant C_8 so that for all n*

$$M_4(\mathcal{A}^n) \le C_8 n^2.$$

Proof. With $\mathcal{B} = \mathcal{A}^n$, $\mathcal{C} = T^{-\Delta-n}\mathcal{A}^n$ we get (by Lemma 12) $H(\mathcal{B} \vee \mathcal{C}) = H(\mathcal{B}) + H(\mathcal{C}) + \mathcal{O}(\psi(\Delta)n^{2\beta} + n^{1-(\beta-1)w})$ and with Minkowski's inequality (on L^4 spaces)

$$\begin{aligned} M_4^{\frac{1}{4}}(\mathcal{B} \vee \mathcal{C}) &= \left(\sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)\left(\log\frac{1}{\mu(B\cap C)} - H(\mathcal{B}\vee\mathcal{C})\right)^4\right)^{\frac{1}{4}} \\ &\le E_4^{\frac{1}{4}}(\mathcal{B},\mathcal{C}) + \mathcal{O}(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w}) + F_4^{\frac{1}{4}}(\mathcal{B},\mathcal{C}), \end{aligned}$$

where by Lemma 11 (with $a = 4$)

$$\begin{aligned} F_4(\mathcal{B},\mathcal{C}) &= \sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)\log^4\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right) \\ &= \mathcal{O}(\psi(\Delta)n^{5\beta} + n^{4\beta-(\beta-1)w}) \end{aligned}$$

and

$$\begin{aligned} E_4(\mathcal{B},\mathcal{C}) &= \sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)(J_{\mathcal{B}}(B) + J_{\mathcal{C}}(C))^4 \\ &= M_4(\mathcal{B}) + M_4(\mathcal{C}) + \sum_{B,C}\mu(B\cap C)(4J_{\mathcal{B}}(B)^3 J_{\mathcal{C}}(C) + 6J_{\mathcal{B}}(B)^2 J_{\mathcal{C}}(C)^2 \\ &\quad + 4J_{\mathcal{B}}(B)J_{\mathcal{C}}(C)^3). \end{aligned}$$

We look individually at the terms in the bracket:

$$\begin{aligned} \left|\sum_{B\in\mathcal{B}, C\in\mathcal{C}} \mu(B\cap C)J_{\mathcal{B}}(B)^3 J_{\mathcal{C}}(C)\right| &= \left|\sum_{B,C}(\mu(B)\mu(C) + \rho(B,C))J_{\mathcal{B}}(B)^3 J_{\mathcal{C}}(C)\right| \\ &\le \sum_{B,C}|\rho(B,C)|\cdot|J_{\mathcal{B}}(B)|^3|J_{\mathcal{C}}(C)| \\ &\le \psi(\Delta)M_3(\mathcal{B})\sigma(\mathcal{C}) \end{aligned}$$

because $J_{\mathcal{B}}$ and $J_{\mathcal{C}}$ have zero average and by Schwarz's inequality. In the same way we get

$$\left|\sum_{B\in\mathcal{B},C\in\mathcal{C}}\mu(B\cap C)J_{\mathcal{B}}(B)J_{\mathcal{C}}(C)^3\right|\le\psi(\Delta)\sigma(\mathcal{B})M_3(\mathcal{C}).$$

Moreover,

$$\begin{aligned}\sum_{B\in\mathcal{B},C\in\mathcal{C}}\mu(B\cap C)J_{\mathcal{B}}(B)^2J_{\mathcal{C}}(C)^2&=\sum_{B,C}(\mu(B)\mu(C)+\rho(B,C))J_{\mathcal{B}}(B)^2J_{\mathcal{C}}(C)^2\\&=\sigma^2(\mathcal{B})\sigma^2(\mathcal{C})+G(\mathcal{B},\mathcal{C}),\end{aligned}$$

where

$$|G(\mathcal{B},\mathcal{C})|=\left|\sum_{B\in\mathcal{B},C\in\mathcal{C}}\rho(B,C)J_{\mathcal{B}}(B)^2J_{\mathcal{C}}(C)^2\right|\le\psi(\Delta)\sigma^2(\mathcal{B})\sigma^2(\mathcal{C}).$$

Thus

$$\begin{aligned}E_4(\mathcal{B},\mathcal{C})=M_4(\mathcal{B})+M_4(\mathcal{C})+(6+\psi(\Delta))\sigma^2(\mathcal{B})\sigma^2(\mathcal{C})\\+\psi(\Delta)(M_3(\mathcal{B})\sigma(\mathcal{C})+\sigma(\mathcal{B})M_3(\mathcal{C})).\end{aligned}$$

As $\sigma^2(\mathcal{B})=\sigma^2(\mathcal{C})=\sigma_n^2\le c_1n$ (Proposition 14) and since by assumption $\psi(\Delta)=\mathcal{O}(\Delta^{-p})$ where $p>8+\frac{24}{w-4}$ we can choose $\beta=1+\frac{2}{w-4}$, $\delta=\frac{1}{p}(4+\frac{12}{w-4})$ and put $\Delta=[n^\delta]$. This implies $\Delta<\sqrt{n}$ (as $\delta<\frac{1}{2}$) and $\psi(\Delta)n^{6\beta}+n^{4\beta-(\beta-1)w}=\mathcal{O}(n^2)$. Using the *a priori* estimates $M_3(\mathcal{A}^n)\le K_3(\mathcal{A}^n)\le C_2n^3$ we obtain in particular that $\psi(\Delta)(M_3(\mathcal{B})\sigma(\mathcal{C})+\sigma(\mathcal{B})M_3(\mathcal{C}))=\mathcal{O}(n^2)$ and therefore

$$M_4^{\frac{1}{4}}(\mathcal{B}\vee\mathcal{C})=\sqrt[4]{M_4(\mathcal{C})+M_4(\mathcal{B})+c_2n^2}+\mathcal{O}(\psi(\Delta))n^{2\beta}+n^{\beta-(\beta-1)w}),$$

where the error term on the right-hand side is $\mathcal{O}(n^{-3})$. To fill in the gap of length Δ we use Lemma 10 and the estimate on K_4 (Corollary 9):

$$\left|M_4^{\frac{1}{4}}(\mathcal{A}^{2n+\Delta})-M_4^{\frac{1}{4}}(\mathcal{B}\vee\mathcal{C})\right|\le M_4^{\frac{1}{4}}(\mathcal{A}^{2n+\Delta}\,|\,\mathcal{B}\vee\mathcal{C})\le K_4^{\frac{1}{4}}(\mathcal{A}^\Delta)\le c_3\Delta.$$

Hence

$$M_4^{\frac{1}{4}}(\mathcal{A}^{n'})\le\sqrt[4]{2M_4(\mathcal{A}^n)+c_2n^2}+c_3\Delta\le\sqrt[4]{2M_4(\mathcal{A}^n)+c_4n^2}$$

(as $\Delta\le\sqrt{n}$), and by induction $M_4(\mathcal{A}^k)\le C_8k^2$ (with $C_8\ge c_4/2$). □

A Hölder estimate can be used to estimate the third absolute moments of J_n as follows:

Corollary 16. *Under the assumptions of Proposition* 15 *there exists a constant* C_9 *so that for all* n

$$M_3(\mathcal{A}^n)\le C_9n^{\frac{3}{2}}.$$

4. Proof of Theorem 2 (CLT for Shannon–McMillan–Breiman)

As before $N(t)$ denotes the normal distribution with zero mean and variance one. We will first show the following result (in which nh has been replaced by H_n and $\sigma\sqrt{n}$ by σ_n).

Theorem 17. *Under the assumptions of Theorem 2 one has:*

(I) *The limit* $\sigma^2 = \lim_{n\to\infty} \frac{1}{n}(K_2(\mathcal{A}^n) - H_n^2)$ *exists (and is positive if* $|\mathcal{A}| = \infty$*).*
(II) *If* $\sigma > 0$ *then*

$$\mathbb{P}\left(\frac{I_n - H_n}{\sigma_n} \le t\right) = N(t) + \mathcal{O}\left(\frac{1}{n^\kappa}\right)$$

for all t *and all*

(i) $\kappa < \frac{1}{10} - \frac{3}{5}\frac{w}{(p+2)(w-2)+6}$ *if* ψ *decays polynomially with power* p,
(ii) $\kappa < \frac{1}{10}$ *if* ψ *decays hyper polynomially.*

Proof of Theorem 17. It is enough to prove the theorem with the partition $\mathcal{A}$ replaced by one of its joins $\mathcal{A}^k$ for some k. Since by Lemma 7 $\mu(A) \le e^{-w}\,\forall\, A \in \mathcal{A}^k$ for some $k \ge 1$ we therefore replace the original partition by $\mathcal{A}^k$ and will henceforth assume that $\mu(A) \le e^{-w}$ for all $A \in \mathcal{A}$.

Theorem 17 part (I) follows from Proposition 14. For the proof of part (II) let us assume that σ is positive. We will use Stein's method to prove the CLT in the form of the following proposition which is modeled after [37] (Röllins [32] has a version that does not require exchangeability):

Proposition 18. ([31]) *Let* (W, W') *be an exchangeable pair so that* $\mathbb{E}(W) = 0$ *and* $\mathrm{var}(W) = 1$ *and assume*

$$\mathbb{E}(W'|W) = (1-\lambda)W$$

for some $\lambda \in (0,1)$*. Then for all real* t:

$$|\mathbb{P}(W \le t) - N(t)| \le \frac{6}{\lambda}\sqrt{\mathrm{var}(\mathbb{E}((W'-W)^2\,|\,W))} + 6\sqrt{\frac{1}{\lambda}\mathbb{E}(|W'-W|^3)}.$$

We proceed in five steps: (A) We begin with a classical "big block-small block" argument and approximate $W_n = \frac{J_n}{\sigma_n}$ by a sum of random variables which are separated by gaps. In (B) we then replace those random variables by independent random variables. In (C) we define the interchangeable pair in the usual way and estimate the terms on the right-hand side of Proposition 18. In (D) and (E) we estimate the effects the steps (A) and (B) have on the distributions.

We approximate $W_n = \frac{J_n}{\sigma_n}$ (clearly $\mathbb{E}(W_n) = 0, \sigma(W_n) = 1$) by the random variable $\hat{W}_n = \frac{1}{\sqrt{r}}\sum_{j=0}^{r-1} W_m \circ T^{m'j}$ (that is $\hat{W}_n = \frac{1}{\sqrt{r}\sigma_m}\sum_{j=0}^{r-1} J_m \circ T^{m'j}$) where $m' = m + \Delta$ and $n = rm + (r-1)\Delta$. (For other values of n not of this form we get an additional error term of the order m'.)

(A) If we put $\hat{\mathcal{A}}^n = \bigvee_{j=0}^{r-1} T^{-m'j}\mathcal{A}^m$ then

$$\|\hat{W}_n - W_n\|_2 \le \frac{1}{\sigma_n}\|J_{\mathcal{A}^n} - J_{\hat{\mathcal{A}}^n}\|_2 + \frac{1}{\sigma_n}\left\|I_{\hat{\mathcal{A}}^n} - \sum_{j=0}^{r-1} I_m \circ T^{m'j}\right\|_2$$
$$+\frac{1}{\sigma_n}|H(\hat{\mathcal{A}}^n) - rH_m| + \left|\frac{1}{\sigma_n} - \frac{1}{\sqrt{r}\sigma_m}\right| \cdot \left\|\sum_{j=0}^{r-1} J_m \circ T^{m'j}\right\|_2.$$

We individually estimate the four terms on the right-hand side as follows:

(i) By Lemma 10 and Proposition 14

$$\|J_{\mathcal{A}^n} - J_{\hat{\mathcal{A}}^n}\|_2 = \sigma(\mathcal{A}^n \mid \hat{\mathcal{A}}^n) = \sigma\left(\bigvee_{j=1}^{r-1} T^{-m-m'j}\mathcal{A}^\Delta \middle| \hat{\mathcal{A}}^n\right)$$
$$= \sigma\left(\bigvee_{j=1}^{r-1} T^{-m-m'j}\mathcal{A}^\Delta\right) \le c_1 r\sqrt{\Delta}.$$

(ii) If $\mathcal{D}_k = \bigvee_{j=0}^{k-1} T^{-m'j}\mathcal{A}^m$ then $\mathcal{D}_{k+1} = \mathcal{D}_k \vee T^{-m'k}\mathcal{A}^m$, $k = 1, 2, \ldots, r$, and by Lemma 11 ($a = 2$)

$$\|I_{\mathcal{D}_{k+1}} - I_{\mathcal{D}_k} - I_m \circ T^{m'k}\|_2^2$$
$$= \sum_{D\in\mathcal{D}_k, A\in T^{-m'k}\mathcal{A}^m} \mu(D\cap A)\log^2\left(\frac{1}{\mu(D\cap A)} - \frac{1}{\mu(D)} - \frac{1}{\mu(A)}\right)$$
$$= \sum_{D\in\mathcal{D}_k, A\in T^{-m'k}\mathcal{A}^m} \mu(D\cap A)\log^2\left(1 + \frac{\rho(D,A)}{\mu(D)\mu(A)}\right)$$
$$\le c_2(\psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w})$$

for $k = 1, 2, \ldots, r$. Hence (as $\mathcal{D}_1 = \mathcal{A}^m$)

$$\left\|I_{\hat{\mathcal{A}}^n} - \sum_{j=0}^{r-1} I_m \circ T^{m'j}\right\|_2 \le \sum_{k=1}^{r} \|I_{\mathcal{D}_{k+1}} - I_{\mathcal{D}_k} - I_m \circ T^{m'k}\|_2$$
$$\le c_3 r\sqrt{\psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w}}.$$

(iii) $|H(\hat{\mathcal{A}}^n) - rH_m| \le c_4 r(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w})$ by Lemma 12.

(iv) Since by Proposition 14

$$\left|\frac{1}{\sigma_n} - \frac{1}{\sqrt{r}\sigma_m}\right| = \frac{|\sigma_n - \sqrt{r}\sigma_m|}{\sqrt{r}\sigma_n\sigma_m} \le c_5\frac{m^{-\eta}}{\sqrt{n}},$$

Lemma 10 and again Proposition 14

$$\left\|\sum_{j=0}^{r-1} J_m \circ T^{m'j}\right\|_2 = \sigma\left(\bigvee_{j=0}^{r-1} T^{-m'j}\mathcal{A}^m\right) \le r\sigma(\mathcal{A}^m) = \mathcal{O}(r\sqrt{m}),$$

we obtain that the fourth term is $\mathcal{O}(\sqrt{r}m^{-\eta})$, for any $\eta < \eta_0$. Therefore, if n is large enough,

$$\begin{aligned}\|\hat{W}_n - W_n\|_2 &\le c_6\bigg(\frac{r\sqrt{\Delta}}{\sqrt{n}} + \frac{r}{\sqrt{n}}\sqrt{\psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w}} \\ &\quad + \frac{r}{\sqrt{n}}(\psi(\Delta)n^{2\beta} + n^{\beta-(\beta-1)w}) + \frac{\sqrt{r}}{m^{\eta}}\bigg) \\ &\le c_7\bigg(\frac{r\Delta}{\sqrt{n}} + rn^{\frac{3}{2}\beta-\frac{1}{2}}\sqrt{\psi(\Delta)}(1 + n^{\frac{1}{2}\beta}\sqrt{\psi(\Delta)}) \\ &\quad + rn^{\beta-\frac{1}{2}-\frac{1}{2}(\beta-1)w} + \frac{\sqrt{r}}{m^{\eta}}\bigg)\end{aligned}$$

as $\sigma_n \sim \sqrt{n}$ and $\beta - 1 > 0$.

(B) Now let X_j for $j = 0, 1, \ldots, r-1$ be independent random variables that have the same distributions as $W_m \circ T^{m'j}$, $j = 0, 1, \ldots, r-1$. Put $D_{V_n}(t)$ for the distribution function of the random variable $V_n = \frac{1}{\sqrt{r}}\sum_{j=0}^{r-1} X_j$ and $D_{\hat{W}_n}(t)$ for the distribution function of $\hat{W}_n$. Since V_n and $\hat{W}_n$ assume the same values, the difference between the distributions is given by (with $\mathcal{D}_k = \bigvee_{j=0}^{k-1} T^{-m'j}\mathcal{A}^m$ as above):

$$\begin{aligned}\sup_t |D_{\hat{W}_n}(t) - D_{V_n}(t)| &\le \sum_{A_0\in\mathcal{A}^m} \cdots \sum_{A_{r-1}\in T^{-m'(r-1)}\mathcal{A}^m}\left|\mu\left(\bigcap_j A_j\right) - \prod_j \mu(A_j)\right| \\ &\le \sum_{k=0}^{r-1}\sum_{D\in\mathcal{D}_k}\sum_{A\in T^{-m'k}\mathcal{A}^m} |\mu(D\cap A) - \mu(D)\mu(A)| \\ &= \sum_{k=0}^{r-1}\sum_{D\in\mathcal{D}_k}\sum_{A\in T^{-m'k}\mathcal{A}^m} |\rho(D, A)| \\ &\le c_8 r\psi(\Delta)\end{aligned}$$

by the mixing property if we assume n is large enough.

(C) In order to apply Proposition 18 let us now define an interchangeable pair in the usual way by setting $V_n' = V_n - \frac{1}{\sqrt{r}}X_Y + \frac{1}{\sqrt{r}}X^*$ where $Y \in \{0, 1, \ldots, r-1\}$ is a randomly chosen index and X^* is a random variable which is independent of all other random variables and has the same distribution as the X_j. Since the random variables X_j for $j = 0, 1, \dots, r-1$, are i.i.d., the pair (V_n', V_n) is exchangeable. Moreover,

$$\mathbb{E}(V_n' \,|\, V_n) = \left(1 - \frac{1}{r}\right) V_n,$$

i.e. $\lambda = \frac{1}{r}$.

We now estimate the two terms on the right-hand side of Proposition 18 separately:

(i) The third moment term of Proposition 18 is estimated using Corollary 16:

$$\mathbb{E}(|V_n' - V_n|^3)^{\frac{1}{3}} = \frac{1}{\sqrt{r}\sigma_m}\mathbb{E}(|J_m \circ T^{m'Y} + J_m^*|^3)^{\frac{1}{3}} \le \frac{2}{\sqrt{r}\sigma_m} M_3(J_m^3)^{\frac{1}{3}} \le \frac{c_9}{\sqrt{r}}.$$

Hence

$$\sqrt{\frac{1}{\lambda}\mathbb{E}(|V_n' - V_n|^3)} = \sqrt{\frac{c_9^3 r}{r^{\frac{3}{2}}}} = \mathcal{O}(r^{-\frac{1}{4}}).$$

(ii) To estimate the variance term we follow Stein [37] and obtain

$$\operatorname{var}(\mathbb{E}((V_n' - V_n)^2 \,|\, V_n)) \le \frac{1}{r^2}\operatorname{var}((X_Y - X^*)^2 \,|\, X_0, X_1, \dots, X_{r-1}).$$

Since

$$\mathbb{E}(X_j^2 \,|\, V_n) = \frac{1}{r}\sum_i \mathbb{E}(X_i^2 \,|\, V_n) = \frac{1}{r}\sum_i X_i^2,$$

we get

$$\operatorname{var}(\mathbb{E}(X_Y^2 \,|\, V_n)) = \operatorname{var}\left(\frac{1}{r}\sum_i X_i^2\right) = \frac{1}{r^2}\operatorname{var}\left(\sum_i X_i^2\right)$$
$$= \frac{1}{r^2} r \operatorname{var}(X_0^2) = \frac{1}{r}\operatorname{var}(X_0^2).$$

Since X_0 has the same distribution as $\frac{1}{\sigma_m}J_m$ we have $\mathbb{E}(X_0) = 0$ and by Propositions 14 and 15

$$\operatorname{var}(X_0^2) = \operatorname{var}\left(\frac{1}{\sigma_m^2}J_m^2\right) = \frac{1}{\sigma_m^4}\sigma^2(J_m^2) \le \frac{1}{\sigma_m^4}M_4(\mathcal{A}^m) \le c_{10}.$$

Hence,

$$\frac{6}{\lambda}\sqrt{\operatorname{var}(\mathbb{E}((V_n' - V_n)^2 \,|\, V_n))} \le c_{11} r \sqrt{\frac{1}{r^3}} \le c_{11}\frac{1}{\sqrt{r}}.$$

Combining the estimates (i) and (ii) yields by Proposition 18

$$|\mathbb{P}(V_n \le t) - N(t)| \le c_{11}\frac{1}{\sqrt{r}} + \frac{6\sqrt{c_9}}{r^{\frac{1}{4}}} \le c_{12}\frac{1}{\sqrt[4]{r}}.$$

(D) Part (B) and (C) combined yield

$$|\mathbb{P}(\hat{W}_n \le t) - N(t)| \le |\mathbb{P}(V_n \le t) - N(t)| + \|D_{\hat{W}_n} - D_{V_n}\|_\infty \le c_{12}\frac{1}{\sqrt[4]{r}} + c_8 r\psi(\Delta).$$

Let us put $\epsilon = \|W_n - \hat{W}_n\|_2$ and $\epsilon' = \sup_t |\mathbb{P}(\hat{W}_n \le t) - N(t)|$. Then ($D_{W_n}$ is the distribution function of W_n) $N(t) \le \epsilon'$ for $t \le -|\log\epsilon'|$ and therefore $D_{\hat{W}_n}(t) \le 2\epsilon'$ for $t \le -|\log\epsilon'|$ and similarly $|1 - N(t)| \le \epsilon'$ and consequently $|1 - D_{\hat{W}_n}(t)| \le 2\epsilon'$ for all $t \ge |\log\epsilon'|$ we get

$$\|(D_{W_n} - D_{\hat{W}_n})\chi_{[-|\log\epsilon'|,|\log\epsilon'|]}\|_\infty \le 2|\log\epsilon'| \cdot \|W_n - \hat{W}_n\|_2 = 2|\log\epsilon'|\epsilon$$

and (since distribution functions are increasing)

$$\|D_{W_n} - D_{\hat{W}_n}\|_\infty \le 2|\log\epsilon'|\epsilon + 2\epsilon'.$$

(E) To optimise the bound

$$|\mathbb{P}(W_n \le t) - N(t)| \le |\mathbb{P}(\hat{W}_n \le t) - N(t)| + \|D_{W_n} - D_{\hat{W}_n}\|_\infty \le 2|\log\epsilon'|\epsilon + 3\epsilon'$$

we distinguish between the case when ψ decays (i) polynomially and (ii) hyper polynomially.

(i) Assume that ψ decays polynomially with power $p > 12$. Let $\delta, \alpha \in (0,1)$ and put $m = [n^\alpha]$, $\Delta = [m^\delta]$ (i.e. $\Delta \sim n^{\alpha\delta}$, $\psi(\Delta) = \mathcal{O}(n^{-\alpha\delta p})$. Then (assuming $n^{\frac{1}{2}\beta}\sqrt{\psi(\Delta)} = \mathcal{O}(1)$ which will be satisfied once we choose β and δ)

$$\|\hat{W}_n - W_n\|_2 \le c_{13}(n^{\frac{1}{2}-\alpha+\alpha\delta} + n^{\frac{1}{2}-\alpha+\frac{3}{2}\beta-\frac{1}{2}\alpha\delta p} + n^{\frac{1}{2}+\beta-\alpha-\frac{1}{2}(\beta-1)w} + n^{\frac{1}{2}-\frac{\alpha}{2}-\alpha\eta}).$$

The first three terms on the right-hand side are optimized by $\beta = \frac{w(p+2)}{(p+2)(w-2)+6}$ and $\alpha\delta = \frac{3\beta}{p+2}$. Then $\|\hat{W}_n - W_n\|_2 \le \epsilon$, $\epsilon = \mathcal{O}(n^x)$, where $x = \max(\frac{1}{2} - \alpha + \frac{3w}{(p+2)(w-2)+6}, \frac{1}{2} - \frac{\alpha}{2} - \alpha\eta)$. The fourth term is smaller than the first three since we can assume that $\eta > \frac{1}{3}$ as $w > 4$. The value of α is found by minimizing the error term $2\epsilon|\log\epsilon'| + 3\epsilon'$. Ignoring the logarithmic term we obtain $\alpha = \frac{3}{5} + \frac{12}{5}\frac{w}{(p+2)(w-2)+6}$ which implies

$$|\mathbb{P}(W_n \le t) - N(t)| \le c_{14}\frac{1}{n^\kappa},$$

for any $\kappa < \frac{1}{10} - \frac{3}{5}\frac{w}{(p+2)(w-2)+6}$. Note that $\alpha\eta > \kappa$ for all (possible) values of p and w.

(ii) If ψ decays faster than any power then we can choose $\delta > 0$ arbitrarily close to zero and obtain $\alpha < \frac{3}{5}$ which yields the estimate $|\mathbb{P}(W_n \le t) - N(t)| \le c_{15}\frac{1}{n^\kappa}$, for any $\kappa < \frac{1}{10}$.

This concludes the proof since $W_n = \frac{I_n - H_n}{\sigma_n}$. □

Proof of Theorem 2. We use Theorem 17 and have to make the following adjustments:

(i) To adjust for the difference between H_n and nh we use Lemma 13:

$$\mathbb{P}\left(\frac{I_n(x)-nh}{\sigma\sqrt{n}}\le t\right)=\mathbb{P}\left(\frac{I_n(x)-H_n}{\sigma\sqrt{n}}\le t+\mathcal{O}(n^{\frac{1}{2}-\gamma})\right)$$
$$=N(t)+\mathcal{O}(n^{-\kappa})+\mathcal{O}(n^{\frac{1}{2}-\gamma}).$$

Since p is large enough, γ can be chosen so that $\gamma-\frac{1}{2}>\kappa$.

(ii) By Proposition 14 $\frac{\sigma_n}{\sqrt{n}}=\sigma+\mathcal{O}(n^{-\eta})$ which yields

$$\mathbb{P}\left(\frac{I_n(x)-H_n}{\sigma\sqrt{n}}\le t\right)=\mathbb{P}\left(\frac{I_n(x)-H_n}{\sigma_n}\le t_n\right)=N(t_n)+\mathcal{O}(n^{-\kappa})$$
$$=N(t)+\mathcal{O}(n^{-\min(\eta,\kappa)}),$$

where $t_n=t\frac{\sigma\sqrt{n}}{\sigma_n}=t(1+\mathcal{O}(n^{-\eta}))$. This concludes the proof since η can be taken to be $>\kappa$. □

5. Proof of Theorem 5 (Weak Invariance Principle)

In order to prove the WIP for $I_n(x)=-\log\mu(A_n(x))$ denote by $W_{n,x}(t)$, $t\in[0,1]$, its interpolation

$$W_{n,x}(k/n)=\frac{I_k(x)-kh}{\sigma\sqrt{n}}$$

$x\in\Omega$ and linearly interpolated on each of the subintervals $[\frac{k}{n},\frac{k+1}{n}]$. In particular $W_{n,x}\in C_\infty([0,1])$ (with supremum norm). Denote by D_n the distribution of $W_{n,x}$ on $C_\infty([0,1])$, namely

$$D_n(H)=\mu(\{x\in\Omega\colon W_{n,x}\in H\})$$

where H is a Borel subset of $C_\infty([0,1])$. The WIP then asserts that the distribution D_n converges weakly to the Wiener measure, which means that $S_n=I_n-nh$ is for large n, and after a suitable normalization distributed approximately as the position at time $t=1$ of a particle in Brownian motion [3].

If we put $S_i=-\log\mu(A_i(x))-ih(\mu)$ then two conditions have to be verified ([3] Theorem 8.1), namely (A) The tightness condition: There exists a $\lambda>0$ so that for every $\varepsilon>0$ there exists an N_0 so that

$$\mathbb{P}\left(\max_{0\le i\le n}|S_i|>2\lambda\sqrt{n}\right)\le\frac{\varepsilon}{\lambda^2}\tag{3}$$

for all $n\ge N_0$.

(B) The finite-dimensional distributions of S_i converge to those of the Wiener measure.

(A) *Proof of tightness.* As before let $J_i = I_i - H_i$ and note that $ih - H_i = \mathcal{O}(i^{1-\gamma})$, $1-\gamma \in (\frac{2w}{p(w-1)}, 1)$, (Lemma 13) is easily absorbed by the term $\lambda\sqrt{n}$ as $1-\gamma < \frac{1}{2}$. In the usual way (cf. e.g. [3]) we get

$$\mathbb{P}\left(\max_{0\le i\le n} |J_i| > 2\lambda\sqrt{n}\right) \le \mathbb{P}(|J_n| > \lambda\sqrt{n}) + \sum_{i=0}^{n-1} \mu(E_i \cap \{|J_i - J_n| \ge \lambda\sqrt{n}\}),$$

where E_i is the set of points x such that $|J_i(x)| > 2\lambda\sqrt{n}$ and $|J_k(x)| \le 2\lambda\sqrt{n}$ for $k = 0, \ldots, i-1$. Note that E_i lies in the σ-algebra generated by $\mathcal{A}^i$. Clearly the sets E_i are pairwise disjoint. To estimate $\mu(E_i \cap \{|J_i - J_n| \ge \lambda\sqrt{n}\})$ let us first "open a gap" of length $\Delta < \frac{n}{2}$. Let $\tilde{\mathcal{A}}^n = \mathcal{A}^i \vee T^{-i-\Delta}\mathcal{A}^{n-i-\Delta}$ (if $i < \frac{n}{2}$ and $\tilde{\mathcal{A}}^n = \mathcal{A}^{i-\Delta} \vee T^{-i}\mathcal{A}^{n-\Delta}$ if $i \ge \frac{n}{2}$), denote by $\tilde{I}_n$ its information function and by $\tilde{H}_n = \mu(\tilde{I}_n)$ its entropy. Obviously $H_n \ge \tilde{H}_n$ and moreover $\mu(I_n - \tilde{I}_n) = H_n - \tilde{H}_n \le H_\Delta \le c_1\Delta$. Since by Lemma 10 and Corollary 9 (as $\mathcal{A}^n$ refines $\tilde{\mathcal{A}}^n$)

$$\sigma(I_n - \tilde{I}_n) = \sigma(\mathcal{A}^n|\tilde{\mathcal{A}}^n) \le \sqrt{K_2(\mathcal{A}^\Delta)} \le c_2\Delta$$

we obtain by Chebychev's inequality ($\tilde{J}_n = \tilde{I}_n - \tilde{H}_n$)

$$\mathbb{P}(|J_n - \tilde{J}_n| \ge \ell) \le \frac{\sigma^2(I_n - \tilde{I}_n)}{\ell^2} \le c_3\frac{\Delta^2}{\ell^2}. \tag{4}$$

By the uniform strong mixing property

$$\tilde{I}_n(B) = I_i(B) + I_{n-i-\Delta}(C) - \log\left(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)}\right)$$

for all $(B, C) \in \mathcal{A}^i \times T^{-i-\Delta}\mathcal{A}^{n-i-\Delta}$. If Y denotes the random variable on $\mathcal{A}^i \times T^{-i-\Delta}\mathcal{A}^{n-i-\Delta}$ whose values are $Y(B,C) = -\log(1 + \frac{\rho(B,C)}{\mu(B)\mu(C)})$ then by Lemma 11 ($a = 2$)

$$\sigma^2(Y) \le \|Y\|_{L^2}^2 \le C_4(\psi(\Delta)(n-\Delta)^{3\beta} + (n-\Delta)^{2\beta-(\beta-1)w})$$

for $\beta > 1$ arbitrary. By Chebychev's inequality this implies

$$\begin{aligned}&\mathbb{P}(|\tilde{J}_n - J_i - J_{n-i-\Delta}\circ T^{i+\Delta}| \ge \ell)\\ &\quad\le \frac{\sigma^2(Y)}{\ell^2} \le C_4\frac{\psi(\Delta)(n-\Delta)^{3\beta} + n^{2\beta-(\beta-1)w}}{\ell^2}.\end{aligned} \tag{5}$$

Then

$$\begin{aligned}&\mu(E_i \cap \{|J_n - J_i| \ge \lambda\sqrt{n}\})\\ &\quad\le \mu(E_i \cap \{|J_n - \tilde{J}_n| \ge \ell\}) + \mu(E_i \cap \{|\tilde{J}_n - J_i - J_{n-i-\Delta}\circ T^{i+\Delta}| \ge \ell\})\\ &\qquad+ \mu(E_i \cap \{|J_{n-i-\Delta}\circ T^{i+\Delta}| \ge \lambda\sqrt{n} - 2\ell\}).\end{aligned}$$

The last term on the right-hand side can be estimated using the mixing property (note that E_i is in the σ-algebra generated by $\mathcal{A}^i$, and $\{|J_{n-i-\Delta}| \ge \lambda\sqrt{n} - 2\ell\}$ is in the σ-algebra generated by $T^{-i-\Delta}\mathcal{A}^{n-i-\Delta}$)

$$\begin{aligned}
&\mu(E_i \cap \{|J_{n-i-\Delta} \circ T^{i+\Delta}| \geq \lambda\sqrt{n} - 2\ell\})\\
&\quad = \mu(E_i)\mathbb{P}(|J_{n-i-\Delta}| \geq \lambda\sqrt{n} - 2\ell) + \sum_{B \subset E_i} \sum_{C \subset T^{-i-\Delta}\{|J_{n-i-\Delta}| \geq \lambda\sqrt{n}-2\ell\}} \rho(B, C)\\
&\quad \leq \mu(E_i)\left(2N\left(\frac{\lambda\sqrt{n} - 2\ell}{\sigma_{n-i-\Delta}}\right) + C_0(n-i-\Delta)^{-\kappa}\right) + \psi(\Delta)
\end{aligned}$$

using Theorem 17 in the last step.

We finally obtain (as $\mathbb{P}(|J_n| > \lambda\sqrt{n}) \leq 2N(\lambda) + c_4 n^{-\kappa}$)

$$\begin{aligned}
&\mathbb{P}\left(\max_{0 \leq i \leq n} |J_i| > 2\lambda\sqrt{n}\right)\\
&\quad \leq 2N(\lambda) + c_4 n^{-\kappa} + \sum_i \mu(E_i \cap \{|J_n - \tilde{J}_n| \geq \ell\})\\
&\qquad + nC_4 \frac{\psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w}}{\ell^2}\\
&\qquad + \sum_i \mu(E_i)\left(2N\left(\frac{\lambda\sqrt{n} - 2\ell}{\sigma_{n-i-\Delta}}\right) + C_0(n-i-\Delta)^{-\kappa}\right) + n\psi(\Delta)\\
&\qquad \leq 2N(\lambda) + c_5 n^{-\kappa} + c_6 \frac{\Delta^2 + \psi(\Delta)n^{3\beta} + n^{2\beta-(\beta-1)w}}{\ell^2} + 2N\left(\frac{\lambda\sqrt{n} - 2\ell}{\sqrt{n}}\right)
\end{aligned}$$

(if $\Delta < \frac{n}{2}$ is small enough). If ψ decays at least polynomially with a power larger than $8 + \frac{24}{w-4}$ then we can put $\ell \sim n^{\alpha}, \Delta \sim n^{\alpha'}$ and choose $\alpha' < \alpha < \frac{1}{2}$ and $\beta > 1$ (e.g. $\beta = \frac{w}{w-2}$, $\alpha' < \frac{3\beta}{p}$) so that the terms on the right-hand side which do not involve the normal probability N decay polynomially in n. This proves the tightness condition (3), since for every $\varepsilon > 0$ one can find a $\lambda > 1$ so that the quadratic estimate holds for all n large enough.

(B) *Proof of the finite-dimensional distribution convergence.* For $t \in [0, 1]$ define the random variable

$$X_n(t, x) = \frac{1}{\sigma\sqrt{n}}(S_{[nt]}(x) + (nt - [nt])(S_{[nt]+1}(x) - S_{[nt]}(x)))$$

which interpolates $S_{[nt]}$. It is defined on Ω and has values in $C_\infty([0, 1])$.

We must show that the distribution of $(X_n(t, x), X_n(t, x) - X_n(s, x))$ converges to $(\mathcal{N}(0, t), \mathcal{N}(0, t-s))$ $(0 \leq s < t)$ as $n \to \infty$, where $\mathcal{N}(0, t)$ is the normal distribution with zero mean and variance t^2. To prove this as well as the convergence of higher finite-dimensional distributions it suffices to show that $X_n(t, x) - X_n(s, x)$ converges to $\mathcal{N}(0, t-s)$ ([3] Theorem 3.2). We obtain by Lemma 13

$$S_{[nt]} - S_{[ns]} = J_{[nt]} - J_{[ns]} + \mathcal{O}((nt)^{1-\gamma})$$

and by (4), (5) and Theorem 2

$$\begin{aligned}
\mathbb{P}\left(\frac{S_{[nt]}-S_{[ns]}}{\sigma\sqrt{n}}\geq\lambda\right)
&\leq \mathbb{P}(|J_{[nt]}-\tilde{J}_{[nt]}|\geq\ell)+\mathbb{P}(|\tilde{J}_{[nt]}-J_{[ns]}-J_{[nt]-[ns]-\Delta}\circ T^{[ns]+\Delta}|\geq\ell)\\
&\quad+\mathbb{P}(|J_{[nt]-[ns]-\Delta}|\geq\lambda\sigma\sqrt{n}-2\ell)+\mathcal{O}((nt)^{\frac{1}{2}-\gamma})\\
&\leq\frac{\sigma^2(I_n-\tilde{I}_n)}{\ell^2}+\frac{\sigma^2(Y)}{\ell^2}+N\left(\frac{\lambda\sigma\sqrt{n}-2\ell}{\sqrt{[nt]-[ns]-\Delta}}\right)\\
&\quad+\frac{C_0}{([nt]-[ns]-\Delta)^\kappa}+\frac{\mathcal{O}(1)}{(nt)^{1-\gamma}}\\
&\leq c_3\frac{\Delta^2}{\ell^2}+C_4\frac{\psi(\Delta)(nt)^{3\beta}+(nt)^{2\beta-(\beta-1)w}}{\ell^2}+\frac{c_7}{(n(t-s))^\kappa}+N\left(\frac{\lambda}{\sqrt{t-s}}\right),
\end{aligned}$$

assuming $\frac{1}{2}-\gamma\geq\kappa$ and $n(t-s)\gg\Delta$. Similarly to above we used a random variable Y on $\mathcal{A}^{[ns]}\times T^{-[ns]-\Delta}\mathcal{A}^{[nt]-[ns]-\Delta}$ given by $Y(B,C)=-\log(1+\frac{\rho(B,C)}{\mu(B)\mu(C)})$. Now let $\ell\sim n^\alpha,\Delta\sim n^{\alpha'}$ and $\alpha'<\alpha<\frac{1}{2}$ and $\beta>1$ so that the terms on the right hand side other than $N(\lambda/\sqrt{t-s})$ decay polynomially in n. Hence $S_{[nt]}-S_{[ns]}$ and therefore $X_n(t,x)-X_n(s,x)$ converges in distribution to $\mathcal{N}(0,\sqrt{t-s})$ as $n\to\infty$. □

Appendix (Markov Chains)

Here we compute the variance for the Markov measure on an infinite alphabet. As in Sec. 2.3 let Σ be the shiftspace over the alphabet $\mathbb{N}$ and μ the Markov measure generated by the probability vector $\mathbf{p}$ and stochastic matrix P. Then

$$\sigma_n^2=\frac{1}{2}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^n}\mu(\mathbf{x})\mu(\mathbf{y})\left(\log\frac{p_{x_1}}{p_{y_1}}+\sum_{j=1}^{n-1}\log\frac{P_{x_jx_{j+1}}}{P_{y_jy_{j+1}}}\right)^2=A_n+B_n+C_n+D_n,$$

where

$$A_n=\frac{1}{2}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^n}\mu(\mathbf{x})\mu(\mathbf{y})\log^2\frac{p_{x_1}}{p_{y_1}}=\frac{1}{2}\sum_{ij}p_ip_j\log^2\frac{p_i}{p_j}=\mathcal{O}(1)$$

and

$$\begin{aligned}
B_n&=\sum_{j=1}^{n-1}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^n}\mu(\mathbf{x})\mu(\mathbf{y})\log\frac{p_{x_1}}{p_{y_1}}\log\frac{P_{x_jx_{j+1}}}{P_{y_jy_{j+1}}}\\
&=\sum_{j=1}^{n-1}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^{j+1}}\mu(\mathbf{x})\mu(\mathbf{y})(\log p_{x_1}\log P_{x_jx_{j+1}}+\log p_{y_1}\log P_{y_jy_{j+1}}\\
&\quad-\log p_{x_1}\log P_{y_jy_{j+1}}-\log p_{y_1}\log P_{x_jx_{j+1}})\\
&=2\sum_{j=1}^{n-1}\sum_{\mathbf{x}\in\mathcal{A}^{j+1}}\mu(\mathbf{x})\log p_{x_1}\log P_{x_jx_{j+1}}+2(n-1)h\sum_i p_i\log p_i.
\end{aligned}$$

For the sum on the R.H.S. we use the Birkhoff ergodic theorem. Denote by χ_1 the function which is defined by $\chi_1(\mathbf{x}) = \log p_{x_1}$ and by χ_2 the function defined by $\chi_2(\mathbf{x}) = \log P_{x_1 x_2}$. Then we get for the first term on the R.H.S.

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n-1}\sum_{\mathbf{x}\in\mathcal{A}^{j+1}} \mu(\mathbf{x}) \log p_{x_1} \log P_{x_j x_{j+1}} &= \lim_{n\to\infty}\int_\Sigma \chi_1(x)\frac{1}{n}\sum_{j=1}^{n-1}\chi_2\circ\sigma^j(x)d\mu(x)\\ &= \int_\Sigma \chi_1(x)d\mu(x)\int_\Sigma \chi_2(x)d\mu(x)\\ &= h\sum_i p_i \log p_i\end{aligned}$$

since $\int \chi_2 d\mu = -h$. Hence we obtain that $\frac{1}{n}B_n \to 0$ as $n\to\infty$. The principal term is

$$D_n = \frac{1}{2}\sum_{j=1}^{n-1}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^n}\mu(\mathbf{x})\mu(\mathbf{y})\log^2\frac{P_{x_j x_{j+1}}}{P_{y_j y_{j+1}}} = \frac{n-1}{2}\sum_{ijk\ell} p_i P_{ij} p_k P_{k\ell}\log^2\frac{P_{ij}}{P_{k\ell}}.$$

Lastly we get the correction term

$$\begin{aligned}C_n &= \sum_{i\neq j}\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^n}\mu(\mathbf{x})\mu(\mathbf{y})\log\frac{P_{x_i x_{i+1}}}{P_{y_i y_{i+1}}}\log\frac{P_{x_j x_{j+1}}}{P_{y_j y_{j+1}}}\\ &= 2\sum_{k=1}^{n-1}(n-k)\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^{k+1}}\mu(\mathbf{x})\mu(\mathbf{y})\log\frac{P_{x_1 x_2}}{P_{y_1 y_2}}\log\frac{P_{x_k x_{k+1}}}{P_{y_k y_{k+1}}}\\ &= 2\sum_{k=1}^{n-1}(n-k)\sum_{\mathbf{x},\mathbf{y}\in\mathcal{A}^{k+1}}\mu(\mathbf{x})\mu(\mathbf{y})(\log P_{x_1 x_2}\log P_{x_k x_{k+1}}\\ &\quad + \log P_{y_1 y_2}\log P_{y_k y_{k+1}} - \log P_{x_1 x_2}\log P_{y_k y_{k+1}} - \log P_{y_1 y_2}\log P_{x_k x_{k+1}})\\ &= 4\sum_{k=1}^{n-1}(n-k)\left(\sum_{\mathbf{x}\in\mathcal{A}^{k+1}}\mu(\mathbf{x})\log P_{x_1 x_2}\log P_{x_k x_{k+1}} - h^2\right).\end{aligned}$$

Since $\sigma^2 = \lim_{n\to\infty}\frac{\sigma_n^2}{n}$ we finally obtain

$$\sigma^2 = \frac{1}{2}\sum_{ijk\ell} p_i P_{ij} p_k P_{k\ell}\log^2\frac{P_{ij}}{P_{k\ell}} + 4\sum_{k=1}^{\infty}\sum_{\mathbf{x}\in\mathcal{A}^{k+1}}\mu(\mathbf{x})(\log P_{x_1 x_2}\log P_{x_k x_{k+1}} - h^2),$$

where the infinite sum converges because the terms (correlations) decay exponentially fast.

Acknowledgements

This work was supported by a grant from the NSF (DMS-0301910).

I would like to thank my colleague Larry Goldstein for many conversations in which he explained Stein's method to me.

References

1. M. Abadi, Exponential approximation for hitting times in mixing stochastic processes, *Math. Phys. Electronic J.* **7** (2001).
2. S. Bernstein, Sur l'extension du théorème limite du calcul des probabilités aux sommes du quantités dépendantes, *Math. Ann.* **97** (1926) 1–59.
3. P. Billingsley, *Convergence in Probability Measures* (Wiley, 1968).
4. X. Bressaud, Subshifts on an infinite alphabet, *Ergod. Th. Dynam. Syst.* **19** (1999) 1175–1200.
5. M. Brin and A. Katok, On local entropy, in *Geometric Dynamics*, Lecture Notes in Mathematics, #1007 (Springer, 1983), pp. 30–38.
6. A. Broise, Transformations dilatantes de l'intervalle et théorèmes limites, Asterisque #238, 1996.
7. H. Bruin and S. Vaienti, Return times for unimodal maps, submitted to *Forum Math.*
8. L. Carleson, Two remarks on the basic theorem of information theory, *Math. Scand.* **6** (1958) 175–180.
9. N. Chernov, Limit theorems and Markov approximations for chaotic dynamical systems, *Probab. Th. Relat. Fields* **101** (1995) 321–362.
10. K. L. Chung, A note on the ergodic theorem of information theory, *Ann. Math. Statist.* **32** (1961) 612–614.
11. P. Collet, A. Galves and B. Schmitt, Fluctuations of repetition times for Gibbsian sources, *Nonlinearity* **12** (1999) 1225–1237.
12. P. Doukhan, *Mixing*: *Properties and Examples*, Lecture Notes in Statistics, Vol. 85 (Springer, 1995).
13. P. Ferrero, N. Haydn and S. Vaienti, Entropy fluctuations for parabolic maps, *Nonlinearity* **16** (2003) 1203–1218.
14. M. Field, I. Melbourne and A. Török, Decay of correlations, central limit theorems and approximations by brownian motion for compact Lie group extensions, *Ergod. Th. Dynam. Syst.* **23** (2003) 87–110.
15. A. Galves and B. Schmitt, Inequalities for hitting times in mixing dynamical systems, *Random Comput. Dynam.* **5** (1997) 337–348.
16. M. Gordin, The central limit theorem for stationary processes, *Sov. Math. Dok.* **10** (1969) 1174–1176.
17. N. Haydn and S. Vaienti, Fluctuations of the metric entropy for mixing measures, *Stoch. Dynam.* **4** (2004) 595–627.
18. I. A. Ibragimov, Some limit theorems for stationary processes, *Theory Probab. Appl.* **7** (1962) 349–382.
19. I. Kontoyiannis, Asymptotic recurrence and waiting times for stationary processes, *J. Theor. Probab.* **11** (1998) 795–811.
20. C. Liverani, Central Limit theorem for deterministic systems, in *Int. Congress on Dyn. Syst.*, Montevideo 1995 (Pitman, 1996), pp. 56–75.
21. R. Mañé, *Ergodic Theory and Differentiable Dynamics* (Springer, 1987).
22. S. Nagaev, Some limit theorems for stationary Markov chains, *Th. Probab. Appl.* **2** (1957) 378–406.
23. A. Nobel and A. Wyner, A recurrence theorem for dependent processes with applications to data compression, *IEEE Trans. Infor. Th.* **38** (1992) 1561–1564.
24. D. Ornstein and B. Weiss, Entropy and data compression schemes, *IEEE Trans. Infor. Th.* **39** (1993) 78–83.
25. D. Ornstein and B. Weiss, Entropy and recurrence rates for stationary random fields, *IEEE Trans. Inf. Th.* **48** (2002) 1694–97.

26. F. Paccaut, Propriétés statistiques de systèmes dynamiques non Markovian, Thèse (Doctorat) Dijon 2000.
27. F. Pène, Rates of convergence in the CLT for two-dimensional dispersive billiards, *Commun. Math. Phys.* **225** (2002) 91–119.
28. K. Petersen, *Ergodic Theory*, Cambridge Studies in Advanced Mathematics, #2 (Cambridge Univ. Press, 1983).
29. V. V. Petrov, On a relation between an estimate of the remainder in the central limit theorem and the law of the iterated logarithm, *Teor. Veroyatn. Primen.* **11** (1966) 514–518; English translation: *Theor. Probab. Appl.* **11** (1966) 454–458.
30. W. Philipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, AMS Memoirs, Vol. 2, No. 161 (Amer. Math. Soc., 1975).
31. Y. Rinott and V. Rotar, Normal approximations by Stein's method, *Decis. Econ. Fin.* **23** (2000) 15–29.
32. A. Röllins, A note on the exchangeability condition in Stein's method, *Statist. Probab. Lett.* **78** (2008) 1800–1806.
33. V. I. Romanovskii, *Discrete Markov Chains* (Wolters-Noordhoff, 1970).
34. M. Rosenblatt, A central limit theorem and a strong mixing condition, *Proc. Nat. Acad. Sci. USA* **42** (1956) 43–47.
35. M. Rosenblatt, *Markov Processes. Structure and Asymptotic Behavior* (Springer, 1971).
36. O. Sarig, Thermodynamic formalism for countable Markov shifts, *Ergod. Th. Dynam. Syst.* **19** (1999) 1565–1593
37. C. Stein, *Approximate Computation of Expectations*, IMS Lecture Notes #7 (World Scientific, 1986).
38. P. Walters, *An Introduction to Ergodic Theory* (Springer, 1981).
39. A. Yushkevich, On limit theorems connected with the concept of the entropy of Markov chains (in Russian), *Usp. Math. Nauk* **8** (1953) 177–180.