Matching of observations of dynamical systems, with applications to sequence matching

Théophile Caby

Matching of observations of dynamical systems, with applications to sequence matching

Théophile Caby,
CMUP, Departamento de Matemática, Faculdade de Ciências, Universidade do Porto,
Rua do Campo Alegre s/n, 4169007 Porto, Portugal.
caby.theo@gmail.com

Abstract

We study the statistical distribution of the closest encounter between generic smooth observations computed along different trajectories of a rapidly mixing dynamical system. At the limit of large trajectories, we obtain a distribution of Gumbel type that depends on both the length of the trajectories and on the Generalized Dimensions of the image measure. It is also modulated by an Extremal Index, that informs on the tendency of nearby observations to diverge along with the evolution of the dynamics. We give a formula for this quantity for a class of chaotic maps of the interval and regular observations. We present diverse numerical applications illustrating the theory and discuss the implications of these results for the study of physical systems. Finally, we discuss the connection between this problem and the problem of the longest matching block common to different sequences of symbols. In particular, we obtain a distributional result for strongly mixing processes.

1 Introduction

Certain real-world systems, such as climate, take place in high-dimensional spaces and exhibit chaotic and multi-scaled properties. To study such complex dynamics, physicists often have access to only a limited number of observable quantities through the measurement process. The latter can be modeled by computing an observation function along a typical trajectory of the system. Understanding the geometric and statistical properties of such observations, and their relationship to the properties of the original underlying system is a problem of great interest in physics, that has been instigated only recently. The study of the recurrence properties of observations have been initiated by Rousseau and Saussol in [33, 32], in which asymptotic and distributional results were obtained for both hitting times and return times of observations. In a recent paper [12], this problem was studied from the point of view of Extreme Value Theory (EVT). This approach turned out to provide information on the local geometry of the image measure, which, for a good choice of observation, can characterize the local fractal structure of the original underlying attractor. In this paper, we pursue the statistical analysis of observed dynamical systems by studying the statistics of the shortest distance between several observed trajectories. Closely related problems have gained interest in recent years. The case of real, unobserved trajectories was considered in [19] and [11], using EVT techniques, while asymptotic results for the shortest distance between two orbits were obtained in [7] and then generalized to multiple orbits [8] and finally to observed orbits [14].

Yet another motivation to study this problem is its deep relationship with a seemingly distinct one; the length of the longest matching block common to different sequences of symbols drawn from the same probability distribution. This old problem has been initiated by Waterman and Arratia, who brought a plethora of results in the i.i.d. case [5, 3], most of which are presented in the reference book [35]. Several authors have extended these results, giving for example distributional results in the i.i.d case [4, 31]. In many applications, however, the sequences cannot be modeled as i.i.d. sequences. For example, in biological applications, genes are specific sequences encoding information, and DNA brands do not constitute independent

sequences of nucleotides. When it comes to written text, a complex dependence structure can arise from specific sequences of letters, such as words, and higher-order syntactic and narrative structures. Recently, Barros, Liao and Rousseau adopted a dynamical system point of view to give the asymptotic behavior of the length of the longest sub-sequence common to different $\alpha-$mixing sequences [7, 8]. This problem is different from the present one, because the sub-sequences may be present at different locations of the different strings of symbols, but we will also follow a dynamical system approach to derive our results.

The paper is organized as follows: In the first section, we present the problem and derive our main result concerning the convergence of the statistics of observation matching to a Gumbel distribution. In the following sections, we discuss the parameters of the limit law, since these quantities provide relevant dynamical information on the system and can be estimated numerically. We first focus on the generalized dimensions of the image measure by emphasizing their central role in the statistical properties of observations. We also study their relations with the generalized dimension spectrum of the original measure. In the third section, we derive a formula for the Extremal Index associated with this problem for a class of chaotic maps of the interval and perform a numerical study of this index for higher dimensional systems. In the last part, we present some applications of our results to sequence matching problems. In particular, we obtain distributional results for the length of the longest sequence of symbols common to independent strings of symbols drawn from the same strongly mixing probability measure.

2 The general approach

Let us consider the dynamical system $(\mathscr{M}, T, \mu)$, where $\mathscr{M}$ denotes the phase space and $T : \mathscr{M} \to \mathscr{M}$ is a discrete transformation [1] that leaves the probability measure μ invariant. In order to model the process of measurement, we consider a C^1 function $f : \mathscr{M} \to \mathscr{J}$, which we refer to as the *observation*. Both the phase space $\mathscr{M}$ and the observation space $\mathscr{J}$ are compact metric spaces endowed with two distances that we will both call d to simplify notations. For physical applications, we take $\mathscr{J} \subset \mathbb{R}^m$, as observational data usually consists of a collection of real numbers that can be arranged into vectors. For applications to the problem of sequence matching, we will take $\mathscr{M}$ to be the space of all infinite sequences of symbols of a given alphabet $\mathscr{A}$. Because we are interested in the statistical properties of observations, we need a measure that is supported in the observational space.

Definition 1 *We call the push-forward, or image of the measure μ by the function f, the measure μ_f defined by*

$$\mu_f(A) = \mu(f^{-1}(A)),$$

for all $A \subset \mathscr{J}$ such that $f^{-1}(A)$ is $\mu-$measurable.

A more detailed presentation of this object is available in [33] and a discussion of its properties can be found in [12].

Definition 2 *We call the generalized dimension of order $q \neq 1$ of the image measure μ_f the following quantity (if it exists):*

$$D_q^f = \lim_{r \to 0} \frac{\log \int_{\mathscr{J}} \mu_f(B(y,r))^{q-1} d\mu_f(y)}{(q-1)\log r}. \tag{1}$$

$B(y,r)$ denotes a ball centered at $y \in \mathscr{J}$ of radius r.

The information dimension of μ_f is defined as

$$D_1^f = \lim_{q \to 1} D_q^f. \tag{2}$$

We write $D_q = D_q^{Id}$, the generalized dimension of order q of the original measure.

[1] it could be a discretized version of a flow

We will place ourselves in physical situations where the limits defining the previous quantities exist. Now that we have introduced the important objects of the theory, we go forward and consider the following process:

$$Y_i = -\log \max_{j=2,\dots,q} d(f(T^i x_1), f(T^i x_j)),$$

$(x_1, ..., x_q) \in \mathscr{M}^q$ being a starting point drawn from the product measure μ_q with support in $\mathscr{M}^q$. To follow the usual procedure of Extreme Value Theory, we consider a sequence of thresholds $u_n(s)$, where $s \in \mathbb{R}$, such that:

$$\mu_q(Y_0 > u_n(s)) \sim \frac{e^{-s}}{n}. \tag{3}$$

Since the q trajectories are independent, we also have:

$$\begin{aligned} \mu_q(Y_0 > u_n(s)) &= \int_{\mathscr{J}} \mu_f(B(y, e^{-u_n}))^{q-1} d\mu_f(y) \\ &\sim e^{-u_n D_q^f (q-1)}, \end{aligned} \tag{4}$$

from definition 2. To satisfy both scalings 3 and 4, we take

$$u_n(s) = \frac{\log n}{D_q^f(q-1)} + \frac{s}{D_q^f(q-1)}.$$

Now, for a given threshold u_n, the quantity $\mu_q(Y_0 > u_n)$ gives the probability of having all the observations contained in the same small region of the observational space; a ball of radius e^{-u_n} centered at $f(x_1)$. Equivalently, it gives the probability that the product dynamics has entered the following target set:

$$S_n^q = \{(s_1, ..., s_q) \in \mathscr{M}^q, \max_{j=2,\dots,q} d(f(s_1), f(s_j)) < e^{-u_n}\}.$$

Following the ideas of [23], studying the behavior of the maximum of the process (Y_i) over a trajectory of size n:

$$M_{n,q}(x_1, ..., x_q) = \max\{Y_0, \dots, Y_{n-1}\},$$

and in particular its cumulative distribution:

$$F_n(u_n) = \mu_q(\{(x_1, ..., x_q) \in \mathscr{M}^q \text{ s.t. } M_{n,q}(x_1, ..., x_q) \leq u_n\}),$$

is equivalent to studying the Hitting Time Statistics of the product dynamics in the set S_n^q. Indeed, $F_n(u_n)$ gives the probability that the dynamics has not entered the set S_n^q after n iterations of the dynamics. We can now apply results from EVT, in particular, the spectral theory developed by Keller and Liverani [30, 29], to obtain the convergence of $F_n(u_n)$ to its limit law.

Proposition 1 *For a large class of exponentially-mixing systems and* regular *observations, there exists* $0 < \theta_q^f \leq 1$ *such that:*

$$|F_n(u_n(s)) - \exp(-\theta_q^f e^{-s})| \underset{n\to\infty}{\to} 0. \tag{5}$$

The term θ_q^f is called the Extremal Index (EI) and quantifies the tendency of the process (Y_i) to form clusters of high values. To be applicable, the spectral theory requires that the couple system/observation satisfies the so-called REPFO property [30, 29], which is verified for rapidly mixing systems for which the measure of the nested target sets S_n^q goes to zero in a regular fashion. More detailed presentations of the theory and its domain of application can be found in various publications [19, 11, 30, 29, 16]. The theory

is proven to be particularly adapted to expanding maps of the interval [18, 16] and certain well-behaved 2-dimensional systems [6].

More classical tools can also be used to prove the convergence to the limit law, in particular under the following conditions, that are particularly adapted to processes generated by dynamical systems.

Definition 3 *We say that the condition $Д_1(u_n)$ is satisfied for the process $Y_0, Y_1, ...$ if there exist a function $\gamma(n,t)$ such that for every $l, t, n \in \mathbb{N}$,*

$$|\mu_q(A_n \cap B_{t,l,n}) - \mu_q(A_n)\mu_q(B_{0,l,n})| \leq \gamma(n,t), \tag{6}$$

where $A_n = \{Y_0 > u_n, Y_1 \leq u_n\}$, $B_{t,l,n} = \bigcap_{i=t}^{t+l-1} T^{-i}(A_n^c)$, and the function $\gamma(n,t)$ is such that it is decreasing in t for each n and such that there exists a sequence $(t_n)_n \in \mathbb{N}$ satisfying $t_n = o(n)$ and $n\gamma(n,t_n) \underset{n\to\infty}{\to} 0$.

Definition 4 *We say that $Д'_1(u_n)$ holds if there exist a sequence $(k_n)_n$ such that:*

1. $k_n \underset{n\to\infty}{\to} \infty$.
2. $k_n t_n = o(n)$, *where $(t_n)_n$ is the sequence in definition 3.*
3. $\lim_{n\to\infty} n \sum_{j=2}^{\lfloor \frac{n}{k_n} \rfloor - 1} \mu_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) = 0$.

Under these two conditions, the result of Proposition 1 holds [18]. We stress that these conditions depend both on the application T and on the observation f. $Д_1(u_n)$ is expected to hold for rapidly mixing systems and regular observations. In particular, we show in the annex that, at least in the context of symbolic dynamics and if $f = Id$, strong exponential mixing implies $Д_1(u_n)$. Condition $Д'_1(u_n)$ concerns the clustering structure of the process Y_i. More particularly, it controls the probabilities of short returns to the target set S_n^q. It is not our focus to give more appropriate conditions of convergence to the limit law, since these can be hard to check in dimension more than one, or sometimes two[2], and even more so when a non-trivial observation f is introduced. We will however provide numerical evidence of the convergence to the extreme value law. Let us now discuss the values of the different parameters of the limit law, that can acquire a physical interpretation.

3 The Generalized Dimensions of the image measure D_q^f

3.1 On the relation between the Generalized dimensions of the image measure and the one of the original invariant measure

We have seen in the preceding section that the quantity D_q^f appears as a parameter of the limit law and therefore modulates the synchronization properties of observations. In fact, these quantities play a central role in different aspects of the statistical properties of observations, and in particular their recurrence times. It is well known that both return and hitting times of certain chaotic systems in small balls (in fact, re-scaled versions of these quantities) have large deviations that are governed by the spectrum of generalized dimensions D_q of the invariant measure [11, 15]. These kinds of large deviations relations are known to hold for real trajectories, but similar results are also expected to apply to the recurrence times of observations for such systems. This matter will be investigated more in detail in a future publication. For now, let us focus on the properties of D_q^f, and in particular on their relation to the generalized dimensions D_q of the original system. In [27], Hunt and Kaloshin give results concerning the effect of typical projections on the generalized dimensions for $1 \leq q \leq 2$. In this range, they show that if $\mathscr{M}$ is a compact subset of $\mathbb{R}^n$ and $\mathscr{J} = \mathbb{R}^m$, and if the generalized dimension of order q, D_q of the invariant measure exists, then:

[2]for simple systems such as automorphisms of the torus [22] or systems admitting a product structure [21]

$$D_q^f = \min(D_q, m), \tag{7}$$

for a prevalent set of C^1 observables. See [28] for a review of prevalence, which is a notion of genericity for infinite-dimensional spaces. For $q > 2$, no such result holds and the behavior of D_q^f in this range is not yet completely understood. Under the light of Hunt and Kaloshin's result, it is possible to access the correlation dimension D_2 of a physical system using a generic observation if the rank is large enough (larger than the correlation dimension of the original attractor). This quantity can be obtained by fitting the empirical distribution of $M_{n,q}$ and extracting the desired parameter, as we will do in the following subsection. Such EVT-based methods of computation of fractal dimensions is in use in climate studies, in particular for the computation of the local dimension, which can be used as a tool to characterize certain climatic patterns [17, 10].

Different kinds of large rank observations can be used by physicists to recover information on the original system. A first approach is to record simultaneously the value of a scalar quantity at different locations of a spatially extended system. These measurements can be arranged into a vector and constitute a so-called gridded observation in $\mathbb{R}^m$. Instead of recording the same quantity at different positions, one can also record different independent observables (temperature, position, speed, pressure, ...) at a given time. Yet another possibility is to consider delay coordinates observables used in embedding techniques [34]. In this context, it is well known that if one considers enough delay coordinates (larger than $\lceil 2D_0 \rceil$), the dynamics of the observation settles on an object (the so-called reconstructed attractor) that is a smooth deformation of the original attractor, which preserves the dimensions [34]. With our approach, only $m \geq D_2$ delay coordinates are required to access the correlation dimension D_2, although the reconstructed attractor is now likely to have a different fine structure from the original one.

3.2 Numerical extraction of D_q^f

Let us now investigate the values of D_q^f for $q > 2$ from a numerical perspective. This procedure will also allow us to experimentally intuit the convergence of the distribution of $M_{n,q}$ to its limit law. Let us consider a system for which the explicit values of D_q are available; the motion on a Sierpinski gasket given by the following Iterated Function System on the unit square $\mathscr{M} = [0,1]^2$:

$$\begin{cases} T_1(x,y) = (x/2, (y+1)/2),\ p_1 = 1/4, \\ T_2(x,y) = ((x+1)/2, (y+1)/2),\ p_2 = 1/4, \\ T_3(x,y) = (x/2, y/2),\ p_3 = 1/2. \end{cases} \tag{8}$$

At each iteration, the application T_i is applied with probability p_i. The associated generalized dimensions spectrum is given, for $q \neq 1$, by [11]:

$$D_q = \frac{\log_2(p_1^q + p_2^q + p_3^q)}{1-q}. \tag{9}$$

In figure (1), we compare the numerical estimates of D_q^f for different observations f and the theoretical values of D_q given by equation (9). These estimates are obtained by evaluating the scale parameter of the empirical maximum distribution of the process (Y_i) over blocks of size 5.10^4, using the maximum likelihood estimator provided by the Matlab function gevfit [25]. The results are averaged over 10 runs, using each time different randomly selected trajectories of length 2.10^8. The error bars represent the standard deviations of the results over these 10 runs.
The functions f_1, f_2 are diffeomorphisms, which are known to preserve the generalized dimensions [27]. Indeed, for these two functions, good agreement is found, so that the two curves are hardly distinguishable visually in the picture. These results suggest that this method of computation of D_q can be completed and even improved by introducing a diffeomorphism computed along the orbits of the system, which may,

if well chosen, speed up the convergence of the method and provide better estimates. Function f_3 is a very oscillatory function, which gives a point in the observational space many antecedents, having the effect to alter significantly the fine structure of the image measure. We do not know whether the disagreement with the D_q spectrum is due to the method not being at convergence, or if is a sign that the spectrum is not preserved under the action of f_3. However, the small disagreement for $q = 2$ seems to indicate that the method may not be at convergence, since the correlation dimension is preserved by typical observations. f_4 is not a diffeomorphism either, but has a more simple structure. For this function, the generalized dimensions seem to be preserved. f_5 is a degenerate function yielding values close to 1.

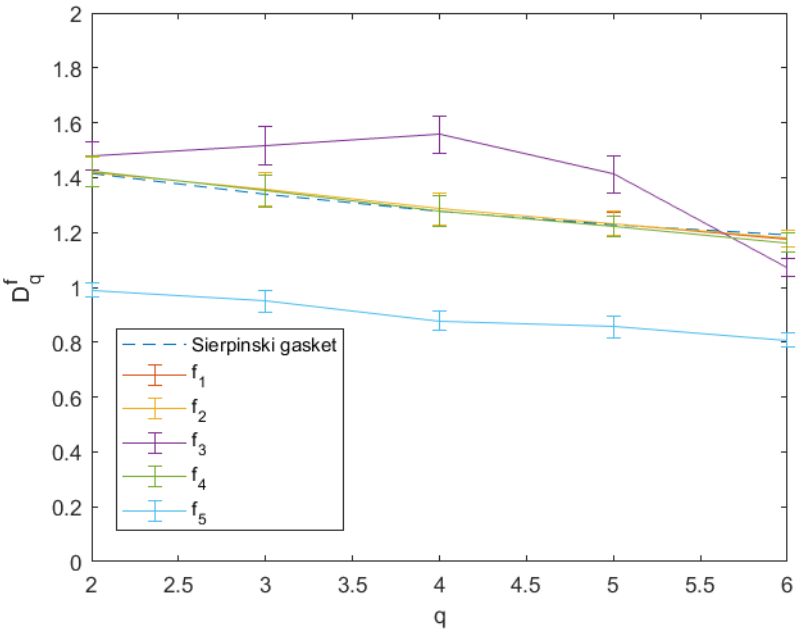


Figure 1: Numerical estimates of D_q^f for different observations: $f_1 = Id$, $f_2(x, y) = (2x + y, 2y)$, $f_3(x, y) = (\sin(\frac{1}{x}), \cos(\frac{1}{y}))$, $f_4(x, y) = ((x - 0.5)^2, 2y)$ and $f_5 = (1, y^2 + x)$. In dashed lines is the D_q spectrum of the underlying system. Estimates are computed as described in the text.

In [12], we showed that for the two-dimensional baker's map, which has a non trivial D_q spectrum, a typical linear uni-dimensional projection gives $D_q^f = 1$ for all q. Overall, this result, along with our numerical computations, suggests that Hunt and Kaloshin's results may extend to $q > 2$ for a certain class of measures and certain smooth observations. We hope to provide their characterization on future investigations.

4 The Extremal Index θ_q^f

4.1 An explicit formula for expanding maps of the interval

When considering real trajectories (i.e. when $f = Id$), the Extremal Index θ_q, and more specifically the quantity

$$h_q = \frac{\log(1 - \theta_q)}{1 - q},$$

encodes the hyperbolic properties of the system (see [11] for a detailed review). In particular, h_q as a function of q is constant for maps with constant Jacobian and is close to the metric entropy of the system

(its Lyapunov exponent in dimension 1). When an observation f is introduced, the use of the Extremal Index to quantify the rate at which nearby trajectories diverge becomes less relevant, in particular because two nearby points in observational space may have antecedents far away in the actual phase space of the system. Let us investigate this matter with more detail. Keller and Liverani [30] provide a general formula for the Extremal Index of time series originated by dynamical systems. Applied to the present situation and if the limits defining the different quantities exist, we have that

$$\theta_q^f = 1 - \sum_{k=0}^{\infty} p_{k,q}, \tag{10}$$

where

$$p_{0,q} = \lim_{n\to\infty} \frac{\mu_q(S_n^q \cap T^{-1}S_n^q)}{\mu_q(S_n^q)} \tag{11}$$

and for $k \geq 1$,

$$p_{k,q} = \lim_{n\to\infty} \frac{\mu_q(S_n^q \cap \bigcap_{i=1}^{k} T^{-i}(S_n^q)^c \cap T^{-k-1}S_n^q)}{\mu_q(S_n^q)}. \tag{12}$$

In this general set up, obtaining a formula for θ_q^f is challenging, so let us place ourselves in the more simple case of expanding maps of the unit interval $I = [0,1]$. We define the following sets for a given $x \in I$:

$$A_0(x) = \{y \in I \text{ such that } f(y) = f(x) \text{ and } f(Ty) = f(Tx)\}$$

and

$$A_k(x) = \{y \in I \text{ such that } f(y) = f(x), f(T^i y) \neq f(T^i x), \text{ for } i = 1,..,k \text{ and } f(T^{k+1}(y)) = f(T^{k+1}(x))\}.$$

Proposition 2 *Let T be an expanding map of the unit interval $I = [0,1]$ which is C^1 by part and admitting an absolutely continuous invariant measure $d\mu(x) = h(x)dx$. Let $f : I \to J \subset \mathbb{R}$ be C^1 by part, finite to one and such that $f' \neq 0$ on I. Suppose moreover that the couple (T, f) satisfies the conditions of Proposition 1, that*

$$\mu(\{x \in I, A_0(x) = \{x\}\}) = 1 \tag{13}$$

and that, for all $k \geq 1$,

$$\mu(\{x \in I, A_k(x) = \emptyset\}) = 1. \tag{14}$$

Then:

$$\theta_q^f = 1 - \frac{\int_I \frac{h(x)^q}{\max(|f'(x)|,|(f\circ T)'(x)|)^{q-1}} dx}{\int_I \sum_{(y_1,\dots y_{q-1})\in(f^{-1}\{f(x)\})^{q-1}} \prod_{i=1}^{q-1} \frac{h(y_i)}{|f'(y_i)|} h(x)dx}. \tag{15}$$

Proof. We write the proof for $q = 2$, the cases $q > 2$ can be obtained in a similar fashion. We start from formula (10) and evaluate both the numerators and the denominators defining the $p_{k,2}$ terms. Let us start by the denominator, for the case $k = 0$. Following the lines of the proof in [16] (where the case $f = Id$ is treated), and making use of the mean value theorem, we get:

$$\begin{aligned} \mu_2(S_n^2) &\sim \int_I \sum_{y\in f^{-1}\{f(x)\}} \mu(B(y, \frac{e^{-u_n}}{|f'(y)|}))d\mu(x) \\ &\sim 2e^{-u_n} \int_I \sum_{y\in f^{-1}\{f(x)\}} \frac{h(y)}{|f'(y)|} h(x)dx. \end{aligned} \tag{16}$$

On the other hand, still for the case $k=0$, we get for the numerator:

$$\begin{aligned}\mu_2(S_n^2 \cap T^{-1}S_n^2) &\sim \int_I \sum_{y\in A_0(x)} \mu(\{z\in I, z\in B(y,\frac{e^{-u_n}}{|f'(y)|}) \cap Tz \in B(Ty,\frac{e^{-u_n}}{|f'(Ty)|}\})d\mu(x)\\ &\sim \int_I \sum_{y\in A_0(x)} \mu(\{z\in I, |z-y|\le \frac{e^{-u_n}}{|f'(y)|} \cap T'(y)|y-z| \le \frac{e^{-u_n}}{|f'(Ty)|}\})h(x)dx.\\ &= \int_I \sum_{y\in A_0(x)} \mu(\{z\in I, |z-y|\le \min(\frac{e^{-u_n}}{|f'(y)|}, \frac{e^{-u_n}}{|T'(y)f'(Ty)|})\})h(x)dx.\\ &\sim 2e^{-u_n}\int_I \sum_{y\in A_0(x)} \frac{h(y)h(x)}{\max(|f'(y)|,|(f\circ T)'(y)|)}dx.\end{aligned} \tag{17}$$

By a similar reasoning, we get that for $k\ge 1$,

$$\mu_2(S_n^2 \cap \bigcap_{i=1}^{k} T^{-i}(S_n^2)^c \cap T^{-k-1}S_n^2) \sim 2e^{-u_n}\int_I \sum_{y\in A_k(x)} \frac{h(x)h(y)}{\max(|f'(x)|,|(f(T^{k+1}(y))'|)}dx. \tag{18}$$

Finally, combining eqs. (10),(2), (17) and (18), we obtain

$$\theta_2^f = 1 - \sum_{k=0}^{+\infty} \frac{\int_I \sum_{y\in A_k(x)} \frac{h(x)h(y)}{\max(|f'(x)|,|(f\circ T^{k+1})'(y)|)}dx}{\int_I \sum_{y\in f^{-1}\{f(x)\}} \frac{h'(y)}{|f'(y)|}h(x)dx}. \tag{19}$$

This formula is still difficult to handle, but under condition (14), $p_{k,2}=0$ for $k>0$, and if condition (13) holds, we obtain

$$\begin{aligned}\theta_2^f &= 1-p_{0,2}\\ &= 1 - \frac{\int_I \frac{h(x)^2}{\max(|f'(x)|,|(f\circ T)'(x)|)}dx}{\int_I \sum_{y\in f^{-1}\{f(x)\}} \frac{h(y)h(x)}{|f'(y)|}dx}.\end{aligned} \tag{20}$$

We can generalize this result for $q\ge 2$ to obtain the desired result. ■

Remark 1 *For a given map T, assumptions (13) and (14) should be satisfied for a generic observation f. The cases where these assumptions are not satisfied are when T and f share some particular symmetries and similarities in their structures. For example, $\mu(A_0(x)=\{x\})\neq 1$ if both the graphs of T and f are symmetric with respect to the straight line of equation $x=1/2$.*

Example 1 *Let us take $Tx = 2x \mod 1$ and*

$$f(x) = \begin{cases} 2x & \text{if } 0\le x\le 1/2\\ 3/2-x & \text{if } 1/2<x\le 1.\end{cases}$$

T is strongly mixing and the couple (T,f) satisfies conditions (13) and (14), so that (T,f) should satisfy the conditions of existence of the limit law, $\mathcal{A}_1(u_n)$ and $\mathcal{A}'_1(u_n)$. It constitutes a good test for our results, since computations can be worked out quite easily. Applying formula (25), we get

$$\theta_q^f = 1-p_{0,q} = 1-\frac{2+2^{2-q}}{1+3^q}.$$

This result is confirmed by numerical experiments (see figure (2)). We used the estimator $\hat{\theta}_5$ introduced in [13], which consists in evaluating the 5 first $p_{k,q}$ terms appearing in formula (12). To do so, we compute

Birkhoff sums for both the numerator and the denominator defining the $p_{k,q}$ terms. It requires fixing a high threshold u, that we take here equal to the 0.99999-quantile of the empirical Y_i distribution. As expected, we find that all the $p_{k,q}$ are 0 or very close to 0 for $k \geq 1$. The results are averaged over 10 runs, with trajectories of length 2.10^7. The error bars in figure (2) represent the standard deviations of the results over these 10 runs. In this example, the exact limit distribution can be computed explicitly; since the image measure is absolutely continuous with a density that does not vanish and that admits no singularities, $D_q^f = 1$ for all q.

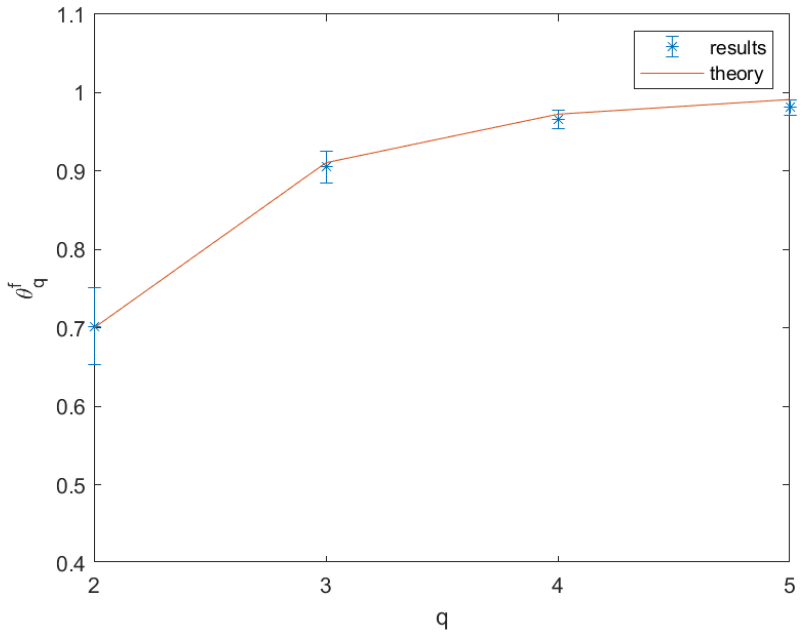


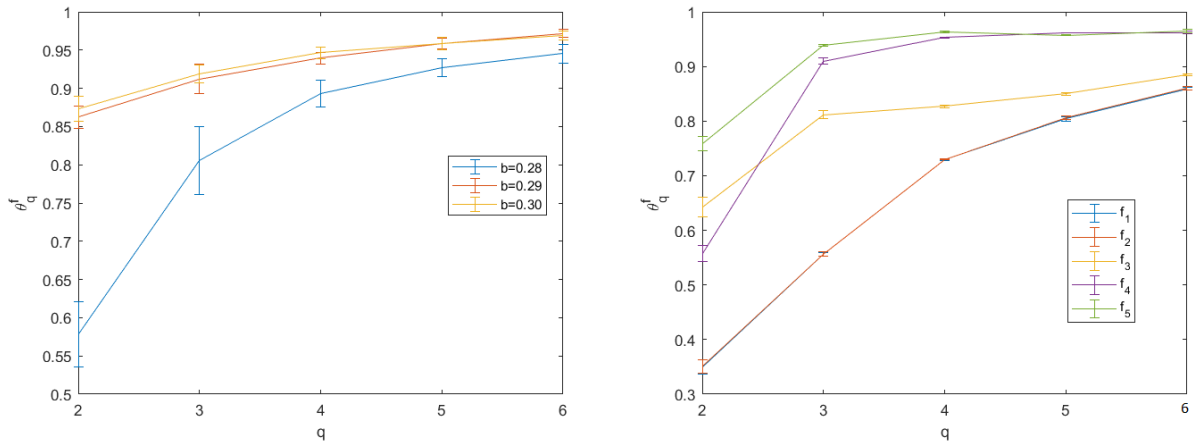
Figure 2: Comparison between theory and computation for the θ_q^f spectrum of the system in Example 1. Details of the computation can be found in the text.

4.2 Numerical estimation of θ_q^f in higher dimensional systems

A general formula for higher-dimensional system is out of scope, but we expect that with conditions of 'non compatibility' between the dynamics and the observation analogue to conditions (13) and (14), all the $p_{k,q}$ terms are 0 for $k \geq 1$. The aim of this section is to show that this hypothesis is corroborated by numerical experiments.

For the uni-dimensional case, the presence of the derivative of the observation in formula (25) renders the interpretation of θ_q^f less apparent than in the case $f = Id$. However, we point out two facts :

- For a given observation f, the larger the values of $|T'|$ over phase space, the larger the values of θ_q^f, so this index can still quantify the hyperbolic properties of T.
- For a given map T, the more the points in the observational space have antecedents by f, the larger is the denominator in equation (25), and the larger is θ_q^f. Oscillatory observations yield higher values for the extremal index.



(a) (b)

Figure 3: Left: Estimates for the θ_q^f spectrum computed for a Hénon system with different parameters b and for the observation $f(x,y) = \frac{x+y}{2}$. Right: Estimates for the θ_q^f spectrum computed for the Hénon system (b=0.3) and different observations : $f_1 = Id$, $f_2(x,y) = (100x + y, 100y)$, $f_3(x,y) = (x, 100y)$, $f_4(x,y) = (x^2, y^2)$, $f_5(x,y) = (\sin(1/x), \cos(1/y))$. For both figures, we used the estimate $\hat{\theta}_5$ introduced in [13], with trajectories of length 10^6 and a threshold value equal to the 0.999 quantile of the empirical Y_i distribution. The error bars represent the standard deviation of the results over 10 runs.

We expect analogous properties to hold for higher dimensional systems. To test this statement, we compare in figure (3a) the estimates of θ_q^f for the 2-dimensional Hénon system, defined by $T(x,y) = (1 - ax^2 + y, bx)$, with $a = 1.4$ and different values of b such that the system admits a strange attractor [26]. We consider the observation $f(x,y) = \frac{x+y}{2}$. The determinant of the Jacobian is given by b. We find indeed that for this fixed choice of observation, the more the original system tends to separate trajectories (the higher is parameter b), the higher are the values of θ_q^f, even for uni-dimensional projections. The estimates $\hat{p}_{k,q}$ of the $p_{k,q}$ terms, for $k > 0$ are all null or close to 0 for all the observations that we considered, as conjectured earlier.

In figure (3b), we plot the estimates of the extremal index for 2-dimensional Hénon system (using the usual parameters $a = 1.4$, $b = 0.3$) and different observations. We observe that for one-to-one observations, (f_1, f_2 and f_3), the θ_q^f spectrum remains relatively low, although the form of the Jacobian can impact significantly the values of θ_q^f. When the observation ceases to be one-to-one, the whole spectrum of extremal indices increases significantly (see the curve for f_4). This effect is even more important for the very oscillatory function f_5. For analogous reasons, we expect that for high dimensional systems, observations that perform a large drop of dimensionality tend to yield higher values for the θ_q^f spectrum.

5 Application to Sequence Matching

In this section, we discuss the connection between the present problem and sequence matching problems. Let

$$\begin{cases} X^1 = X_1^1 X_2^1 \dots X_n^1, \\ \vdots \\ X^q = X_1^q X_2^q \dots X_n^q \end{cases} \tag{21}$$

be q sequences of symbols of length n, drawn from the finite alphabet $\mathscr{A}$ with the same probability distribution $\mathbb{P}$. We will denote $\bar{X}_i = (X_i^1, X_i^2, ..., X_i^q)$. We suppose that the sequences have a good dependence structure that we will describe later. We are interested in deriving the limit distribution of the length of the longest matching block for the q sequences; the following random variable:

$$\Xi_{n,q}(X^1, ..., X^q) = \max_{l=0,...,n} \{X_{i+k}^1 = X_{i+k}^2 = ... = X_{i+k}^q \text{ for } k = 0, ..., l \text{ and } 1 \leq i \leq n-l\}. \tag{22}$$

To make the connection between the previous sections, let us now consider, as in [7], the discrete symbolic dynamical system $(\mathscr{A}^{\mathbb{N}}, \sigma, \mathbb{P})$, where σ is the right-sided shift and $\mathbb{P}$ is the probability measure associated to the process. We consider the symbolic distance in $\mathscr{A}^{\mathbb{N}}$ defined by:

$$d(x^1, x^2) = \exp(-\inf\{i \geq 0, \sigma^i x^1 \neq \sigma^i x^2\}). \tag{23}$$

For our purpose, we take $f = Id$. In this symbolic dynamics, the quantity D_q (if it exists) identifies with a well-known quantity that we now define.

Definition 5 *We call the Rényi entropy of order q of $\mathbb{P}$, the following quantity (if the limit exists):*

$$H_q = \lim_{k \to \infty} \frac{\log \sum_{C_k} \mathbb{P}(C_k)^q}{(1-q)k}, \tag{24}$$

where $C_k(x) = \{y \in \mathscr{A}^{\mathbb{N}} : \sigma^i x = \sigma^i y \text{ for all } 0 \leq i \leq k\}$ is the cylinder of length k containing $x \in \mathscr{A}^{\mathbb{N}}$.

To see that D_q identifies with H_q in this context, it is enough to start from definition (2), take $f = Id$ and use the symbolic distance, allowing to replace balls by cylinders.

The Dynamical Extremal Index $\theta_q = \theta_q^{Id}$ becomes in this set up (if it exists, and from equation (10)):

$$\begin{aligned} \theta_q &= 1 - p_{0,q} \\ &= \lim_{k \to \infty} \mathbb{P}(\sigma^{k+1} x^1 = \sigma^{k+1} x^2 = ... = \sigma^{k+1} x^q | \sigma^i x^1 = \sigma^i x^2 = ... = \sigma^i x^q \text{ for } 0 \leq i \leq k). \end{aligned} \tag{25}$$

Indeed one sees easily that only the $p_{0,q}$ in definition (10) is non-zero in this situation (we provide a more detailed argument in the annex).

The quantity

$$\begin{aligned} Y_i &= -\log(\max_{s=2,...,q} d(x_i^1, x_i^s)) \\ &= \inf_{j \geq 0} \{\sigma^j x_i^1 \neq \sigma^j x_i^s, \text{for some } s = 2, ..., q\} \end{aligned} \tag{26}$$

is the length of the longest matching sub-sequence starting from the i^{th} symbol of the different sequences. Now, the quantity

$$M_{n,q}(x^1, ..., x^q) = \max_{i=0,n-1} Y_i \tag{27}$$

is equal to

$$\max_{l \in \mathbb{N}} \{x_{i+k}^1 = x_{i+k}^2 = ... = x_{i+k}^q \text{ for } k = 1, ..., l \text{ and } 0 \leq i \leq n-1\}.$$

This object is closely related to the quantity $\Xi_{n,q}$ we are interested in. Since we work with different sequences of symbols, and Y_i is a variable defined in the product space, we will state our results with respect to

the product measure $\mathbb{P}_q$. We prove our results under the hypothesis that the process has certain mixing properties, that we now recall.

Definition 6 *The process $(\mathscr{A}^{\mathbb{N}}, \sigma, \mathbb{P})$ is said to be $\alpha-$mixing if there exists $\alpha(n) \to 0$ such that*

$$\sup_{A,B \subset \mathscr{A}^{\mathbb{N}}} |\mathbb{P}(A \cap \sigma^{-n}B) - \mathbb{P}(A)\mathbb{P}(\sigma^{-n}B)| \leq \alpha(n). \tag{28}$$

One could obtain a distributional result analogue to Proposition 1, by proving that conditions $Д_1(u_n)$ and $Д'_1(u_n)$ are satisfied. With this approach, we get the following result, whose detailed proof can be found in the annex:

Result 1 *If the sequences are α-mixing with $\alpha(n) < \beta e^{-\kappa n}$ for some $\beta \in \mathbb{R}^+$ and some $\kappa > H_q(q-1)$, and the limits defining θ_q and H_q exist and are different from 0, then*

$$|\mathbb{P}_q(\Xi_n^q \leq u_n(s)) - \exp(-\theta_q \exp(-s))| \underset{n\to\infty}{\to} 0,$$

with $u_n(s) = \lfloor \frac{\log n + s}{H_q(q-1)} \rfloor$.

Remark 2 *We took $f = id$, to ensure a clustering structure that satisfies the different conditions of existence of the limit law, in particular condition $Д'_1(u_n)$. We could also consider, as in the first section of the paper, a non-trivial f. In the context of sequence matching, f is called the encoding function (or encoder) and can model different treatments of the original source of information [14]. The clustering structure is however in this case too complex to yield such a general result.*

It is in fact possible to obtain a more general result than Result 1, under much weaker conditions. The latter is based on results by Abadi and Saussol concerning the Hitting Time Statistics of symbolic dynamics in cylinders [2]. This idea originates from a discussion with Jérôme Rousseau to whom the author is thankful.

Theorem 1 *If $\mathbb{P}$ is $\alpha-$ mixing, and if the limits defining θ_q and H_q exist and are different from 0, then*

$$|\mathbb{P}_q(\Xi_{n,q} \leq u_n(s)) - \exp(-\theta_q \exp(-s))| \underset{n\to\infty}{\to} 0,$$

with $u_n = u_n(s) = \lfloor \frac{\log n + s}{H_q(q-1)} \rfloor$.

Proof. Let us consider the process (Z_i) defined by

$$Z_i = \begin{cases} 1 \text{ if } X_i^1 = X_i^2 = ... = X_i^q, \\ 0 \text{ otherwise.} \end{cases} \tag{29}$$

The problem of finding the largest common substring to $X^1, ..., X^q$ is now equivalent to find the longest succession of ones in the process (Z_i). Let us consider the dynamical system $(\mathscr{B}, \tilde{\mathbb{P}}, \sigma)$, where $\mathscr{B} = \{0,1\}^{\mathbb{N}}$, z a point in $\mathscr{B}$ and $\tilde{\mathbb{P}}$ the probability measure defined by

$$\begin{aligned} \tilde{\mathbb{P}}(z_i = 1) &= \mathbb{P}_q(x_i^1 = ... = x_i^q) \\ &= \sum_{a \in \mathscr{A}} \mathbb{P}(x_i^1 = a)^q. \end{aligned} \tag{30}$$

Let us denote I_k the cylinder constituted of all sequences having their first k symbols equal to 1, and denote

$$\tau_{I_k}(z) = \inf\{j \geq 1 : \sigma^j z \in I_k\},$$

the first hitting time of the point z in the set I_k. We notice that

$$\mathbb{P}_q(M_{n,q} < u_n) = \tilde{\mathbb{P}}(\tau_{I_{u_n}} > n). \tag{31}$$

Since $\mathbb{P}$ is $\alpha - mixing$, so is $\tilde{\mathbb{P}}$, by theorem 5.1 in [9]. We are then in the set up of Theorem 1 in [2]. In particular, Hypothesis 1 of this theorem is satisfied, from Example 2 in [2]. Therefore:

$$\sup_{t\in\mathbb{R}^+} |\tilde{\mathbb{P}}(\lambda(I_{u_n})\tilde{\mathbb{P}}(I_{u_n})\tau_{I_{u_n}} > t) - \exp(-t)| \underset{n\to\infty}{\to} 0, \tag{32}$$

where, from [1]:

$$\begin{aligned}\lambda(I_{u_n}) &= 1 - \lim_{k\to\infty} \frac{\tilde{\mathbb{P}}(I_{k+1})}{\tilde{\mathbb{P}}(I_k)} \\ &= 1 - \lim_{k\to\infty} \frac{\sum_{C_{k+1}} \mathbb{P}(C_{k+1})^q}{\sum_{C_k} \mathbb{P}(C_k)^q} \\ &= \theta_q.\end{aligned} \tag{33}$$

Notice now that we have from equation (24):

$$\begin{aligned}H_q &= \lim_{k\to\infty} \frac{1}{(1-q)k} \log \sum_{C_k} \mathbb{P}(C_k)^q \\ &= \lim_{k\to\infty} \frac{\log \tilde{\mathbb{P}}(I_k)}{(1-q)k},\end{aligned} \tag{34}$$

so that

$$\tilde{\mathbb{P}}(I_{u_n}) \underset{n\to\infty}{\sim} e^{-(q-1)H_q u_n}. \tag{35}$$

If we put $t = e^{-s}$, equation (32) writes, after rearranging a bit:

$$\sup_{s\in\mathbb{R}} |\tilde{\mathbb{P}}(\tau_{I_{u_n}} > e^{-s+(q-1)H_q u_n}) - \exp(-\theta_q e^{-s})| \underset{n\to\infty}{\to} 0. \tag{36}$$

keeping in mind that $u_n = \lfloor \frac{\log n + s}{H_q(q-1)} \rfloor$, we get

$$\sup_{s\in\mathbb{R}} |\tilde{\mathbb{P}}(\tau_{I_{u_n}} > n) - \exp(-\theta_q e^{-s})| \underset{n\to\infty}{\to} 0. \tag{37}$$

Using now equation (31), we obtain that for all $s \in \mathbb{R}$:

$$|\mathbb{P}_q(M_{n,q} > u_n) - \exp(-\theta_q e^{-s})| \underset{n\to\infty}{\to} 0. \tag{38}$$

Now that we have a distributional result for the variable $M_{n,q}$, we can get one for $\Xi_{n,q}$, which is a slightly different object. In fact we have that

$$\begin{aligned}\mathbb{P}_q(\Xi_{n,q} \le u_n) &= \mathbb{P}_q(\Xi_{n,q} \le u_n \cap M_{u_n,q}(\sigma^{n-u_n}x^1, ..., \sigma^{n-u_n}x^q) \le u_n) \\ &+ \mathbb{P}_q(\Xi_{n,q} \le u_n \cap M_{u_n,q}(\sigma^{n-u_n}x^1, ..., \sigma^{n-u_n}x^q) > u_n).\end{aligned} \tag{39}$$

The second term is bounded above by the term $\mathbb{P}_q(M_{u_n,q}(\sigma^{n-u_n}x^1, ..., \sigma^{n-u_n}x^q) > u_n)$, which, by invariance of the measure by σ equals $\mathbb{P}_q(M_{u_n,q}(x^1, ...x^n) > u_n)$, which is clearly vanishing to 0 as $n \to \infty$, from (38).

The first term in (39) is exactly equal to $\mathbb{P}_q(M_{n,q}(\overline{x}) \le u_n)$. Therefore:

$$|\mathbb{P}_q(\Xi_{n,q} \le u_n) - \mathbb{P}_q(M_{n,q} \le u_n)| \underset{n\to\infty}{\to} 0. \tag{40}$$

We have that for all $s \in \mathbb{R}$:

$$|\mathbb{P}_q(\Xi_{n,q} \leq u_n) - \exp(-\theta_q e^{-s})| \leq |\mathbb{P}_q(\Xi_{n,q} \leq u_n) - \mathbb{P}_q(M_{n,q} \leq u_n)| + |\mathbb{P}_q(M_{n,q} \leq u_n) - \exp(-\theta_q e^{-s})|, \quad (41)$$

which, by relations (38) and (40) goes to 0. ■

6 Acknowledgement

The author was partially supported by CMUP, which is financed by national funds through FCT – Fundação para a Ciência e Tecnologia, I.P., under the project with reference UIDB/00144/2020. The author would like to thank Jorge M. Freitas, Jérôme Rousseau, Benoît Saussol and Sandro Vaienti for the fruitful discussions we had concerning this work and the anonymous referee for its constructive comments.

7 Annex (proof of Result 1, via EVT)

We first show that both conditions Д$_1(u_n)$ and Д'$_1(u_n)$ are satisfied, so we have an EVL for the random variable $M_{n,q}$. Then we show that $\Xi_{n,q}$ and $M_{n,q}$ have the same asymptotic distribution. Let us first take care of condition Д$'_1(u_n)$.
We observe that if $Y_0 = k \in \mathbb{N}^*$, then $Y_j = k - j$ for $1 \leq j \leq k$. Therefore, if $Y_0 > u_n$, then $Y_1 > Y_j > u_n - j$ for $2 \leq j < u_n$, so that all the probabilities in the sum in point 3 of definition 4 are 0 for $2 \leq j < u_n$, that is

$$\lim_{n\to\infty} n \sum_{j=2}^{u_n-1} \mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) = 0. \quad (42)$$

Let $0 < \varepsilon_2 < \varepsilon_1 < 1$ and $C_1 = 1 - \varepsilon_1$. We define $r_n = \lfloor C_1 u_n \rfloor$ and $\lambda_n = \lfloor n^{\varepsilon_2} \rfloor$. We take j such that $u_n \leq j \leq \lambda_n$. We observe that $\{Y_j > u_n\} \subset \{Y_{j+r_n} > u_n - r_n\}$, so that

$$\mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) \leq \mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_{j+r_n} > u_n - r_n). \quad (43)$$

Notice that $\{Y_0 > u_n \cap Y_1 \leq u_n\} = \{Y_0 = u_n + 1\}$, and this event depends only on the realizations of $\bar{X}_1, ..., \bar{X}_{u_n+2}$, whereas $\{Y_{j+r_n} > u_n - r_n\}$ depends only on the realizations of $\bar{X}_{j+r_n}, \bar{X}_{j+r_n+1}, ...$, which puts a gap of length $j + r_n - u_n - 2$. We now use the fact that the sequences are $\alpha-$mixing, which implies that the $q-$fold Cartesian product of the sequences is (α_q)-mixing, with $\alpha_q(n) \leq q\alpha(n)$ (see theorem 5.1 in [9]). We have

$$\begin{aligned}
\mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) &\leq \alpha_q(j + r_n - u_n - 2) + \mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n)\mathbb{P}_q(Y_{j+r_n} > u_n - r_n) \\
&\leq q\alpha(j + r_n - u_n - 2) + \mathbb{P}_q(Y_0 > u_n)\mathbb{P}_q(Y_{j+r_n} > u_n - r_n) \\
&\leq q\beta e^{-\kappa(j+r_n-u_n-2)} + \mathbb{P}_q(Y_0 > u_n)\mathbb{P}_q(Y_0 > u_n - r_n).
\end{aligned} \quad (44)$$

To get the last inequality, we used the invariance of the measure. Notice that $j \geq u_n$, so that $e^{-\kappa(j+r_n-u_n-2)} \leq e^{-\kappa(r_n-2)}$. We also have from relation (17) that $\mathbb{P}_q(Y_0 > u_n) \sim e^{-u_n\tau_q}$, where $\tau_q = H_q(q-1)$, so that there exists $C_2 > 1$ such that

$$\mathbb{P}_q(Y_0 > u_n) < C_2 e^{-u_n\tau_q}.$$

We then have:

$$\mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) \leq q\beta e^{-\kappa(r_n-2)} + C_2^2 e^{-(2u_n-r_n)\tau_q}. \quad (45)$$

Then we can write

$$
\begin{aligned}
n\sum_{j=u_n}^{\lambda_n}\mathbb{P}_q(Y_0>u_n\cap Y_1\leq u_n\cap Y_j>u_n) &\leq \sum_{j=u_n}^{\lambda_n}\left[nq\beta e^{-\kappa(r_n-2)}+nC_2^2e^{-(2u_n-r_n)\tau_q}\right]\\
&\leq (\lambda_n-u_n)nq\beta e^{-\kappa(r_n-2)}+(\lambda_n-u_n)nC_2^2e^{-(2u_n-r_n)\tau_q}\\
&\leq \lambda_n nq\beta e^{-\kappa(r_n-2)}+\lambda_n nC_2^2e^{-(2u_n-r_n)\tau_q}\\
&\leq (q\beta e^{2\kappa})n\lambda_n e^{-\kappa r_n}+C_2^2n\lambda_n e^{-2(u_n-r_n)\tau_q}\\
&\leq (q\beta e^{2\kappa})n\lambda_n e^{-\kappa\lfloor C_1u_n\rfloor}+C_2^2n\lambda_n e^{-2(u_n-\lfloor C_1u_n\rfloor)\tau_q}\\
&\leq (q\beta e^{2\kappa})n\lambda_n e^{-\kappa(C_1u_n-1)}+C_2^2n\lambda_n e^{-(2-C_1)u_n\tau_q}\\
&\leq (q\beta e^{3\kappa})n\lambda_n e^{-\kappa C_1\lfloor\frac{\log n+s}{\tau_q}\rfloor}+C_2^2n\lambda_n e^{-(2-C_1)\lfloor\frac{\log n+s}{\tau_q}\rfloor\tau_q}\\
&\leq (q\beta e^{3\kappa})n\lambda_n e^{-\kappa C_1(\frac{\log n+s}{\tau_q}-1)}+C_2^2n\lambda_n e^{-(2-C_1)(\frac{\log n+s}{\tau_q}-1)\tau_q}\\
&\leq C_3n\lambda_n e^{-\kappa C_1\frac{\log n}{\tau_q}}+C_4n\lambda_n e^{-(2-C_1)\log n},
\end{aligned}
\tag{46}
$$

with

$$C_3=q\beta e^{3\kappa}e^{-\kappa C_1(\frac{s}{\tau_q}-1)}$$

and

$$C_4=C_2^2e^{(C_1-2)(s-\tau_q)}.$$

For the first term, we have

$$
\begin{aligned}
C_3n\lambda_n e^{-\kappa C_1\frac{\log n}{\tau_q}} &= C_3n\lfloor n^{\varepsilon_2}\rfloor e^{-\kappa C_1\frac{\log n}{\tau_q}}\\
&\leq C_3nn^{\varepsilon_2}e^{-\kappa C_1\frac{\log n}{\tau_q}}\\
&\leq C_3n^{1+\varepsilon_2-\frac{\kappa C_1}{\tau_q}}\\
&\leq C_3n^{1+\varepsilon_2-\frac{\kappa(1-\varepsilon_1)}{\tau_q}}.
\end{aligned}
\tag{47}
$$

Since $\kappa>\tau_q$, we can always chose $\varepsilon_1,\varepsilon_2$ and $\varepsilon_3>0$ such that

$$\epsilon_3>\frac{(\varepsilon_1+\varepsilon_2)\tau_q}{1-\varepsilon_1} \tag{48}$$

and

$$\kappa>\tau_q+\varepsilon_3. \tag{49}$$

We then have

$$1+\varepsilon_2-\frac{\kappa(1-\varepsilon_1)}{\tau_q}<\varepsilon_1+\varepsilon_2-\frac{\varepsilon_3(1-\varepsilon_1)}{\tau_q}<0. \tag{50}$$

and so by relation (47):

$$C_3n\lambda_n e^{-\kappa C_1\frac{\log n}{\tau_q}}\underset{n\to\infty}{\to}0. \tag{51}$$

Let us now come to the second term in relation (46):

$$
\begin{aligned}
C_4n\lambda_n e^{-(2-C_1)\log n} &= C_4n\lfloor n^{\varepsilon_2}\rfloor e^{-(1+\varepsilon_1)\log n}\\
&\leq C_4n^{\varepsilon_2-\varepsilon_1}.
\end{aligned}
\tag{52}
$$

And since $\varepsilon_1 > \epsilon_2$:

$$C_4 n \lambda_n e^{-(2-C_1)\log n} \underset{n\to\infty}{\to} 0. \tag{53}$$

Combining relations (51), (53) and (46), we have that:

$$\lim_{n\to\infty} n \sum_{j=u_n}^{\lambda_n} \mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) = 0. \tag{54}$$

Combining equation (42) and (54), we get that

$$\lim_{n\to\infty} n \sum_{j=2}^{\lambda_n} \mathbb{P}_q(Y_0 > u_n \cap Y_1 \leq u_n \cap Y_j > u_n) = 0. \tag{55}$$

Taking $k_n = \frac{n}{\lambda_n} = n^{1-\varepsilon_2}$, we have that points 1 and 3 of condition $Д_1'(u_n)$ are satisfied. To satisfy point 2, we take $t_n = \lfloor n^{\varepsilon_4} \rfloor$, with $\varepsilon_4 < \varepsilon_2$. $Д_1'(u_n)$ is then satisfied.

Let us now come to condition $Д_1(u_n)$. Define the event $\Omega_n = \{Y_0 < \lfloor t_n/2 \rfloor\}$. We have that

$$\begin{aligned} |\mathbb{P}_q(A_n \cap B_{t_n,l,n}) - \mathbb{P}_q(A_n)\mathbb{P}_q(B_{0,l,n})| \leq\ & |\mathbb{P}_q(A_n \cap \Omega_n \cap B_{t_n,l,n}) - \mathbb{P}_q(A_n \cap \Omega_n)\mathbb{P}_q(B_{0,l,n})| \\ & + |\mathbb{P}_q(A_n \cap \Omega_n^c \cap B_{t_n,l,n}) - \mathbb{P}_q(A_n \cap \Omega_n^c)\mathbb{P}_q(B_{0,l,n})|. \end{aligned} \tag{56}$$

We have just introduced a gap of size $t_n - \lfloor t_n/2 \rfloor - 1$ in the first term of the right hand side of the previous inequation. Indeed, for n large enough, the event $A_n \cap \Omega_n$ depends only on the realizations of $\bar{X}_1, \bar{X}_2, ..., \bar{X}_{\lfloor t_n/2 \rfloor}$, while $B_{t_n,l,n}$ depends on the realizations of $\bar{X}_{t_n}, ..., \bar{X}_{t_n+l}$. We can then bound this term, using again theorem 5.1 in [9] :

$$\begin{aligned} |\mathbb{P}_q(A_n \cap \Omega_n \cap B_{t_n,l,n}) - \mathbb{P}_q(A_n \cap \Omega_n)\mathbb{P}(B_{0,l,n})| &\leq q\alpha(t_n - \lfloor t_n/2 \rfloor - 1) \\ &\leq q\beta e^{-\kappa(t_n - \lfloor t_n/2 \rfloor - 1)} \\ &\leq q\beta e^{-\kappa(t_n/2-1)}. \end{aligned} \tag{57}$$

For the second term, we can write

$$\begin{aligned} |\mathbb{P}_q(A_n \cap \Omega_n^c \cap B_{t,l,n}) - \mathbb{P}_q(A_n \cap \Omega_n^c)\mathbb{P}_q(B_{0,l,n})| &\leq |\mathbb{P}_q(A_n \cap \Omega_n^c \cap B_{t,l,n})| + |\mathbb{P}_q(A_n \cap \Omega_n^c)\mathbb{P}_q(B_{0,l,n})| \\ &\leq 2\mathbb{P}_q(\Omega_n^c) \sim 2e^{-\lfloor t_n/2 \rfloor \tau_q} \\ &\leq C_5 e^{-\lfloor t_n/2 \rfloor \tau_q} \\ &\leq C_5 e^{-(t_n/2-1)\tau_q}, \end{aligned} \tag{58}$$

for some $C_5 > 2$.

Let us now take

$$\gamma(n, t_n) = q\beta e^{-\kappa(t_n/2-1)} + C_5 e^{-(t_n/2-1)\tau_q}.$$

Combining expressions (56), (57), (58), we get

$$|\mathbb{P}_q(A_n \cap B_{t,l,n}) - \mathbb{P}_q(A_n)\mathbb{P}_q(B_{0,l,n})| \leq \gamma(n, t_n). \tag{59}$$

Let us recall that from condition $Д_1'(u_n)$, $t_n = \lfloor n^{\varepsilon_4} \rfloor = o(n)$. γ is clearly decreasing and we check easily that

$$n\gamma(n, t_n) \underset{n\to\infty}{\to} 0.$$

Condition $Д_1(u_n)$ is then satisfied. We can now apply corollary 4.1.7 in [18] to get that

$$\mathbb{P}_q(M_{n,q} \leq u_n) - \exp(-\theta_q \exp(-s)) \underset{n\to\infty}{\to} 0. \tag{60}$$

We conclude by using the same arguments as in the proof of Theorem 1, showing that $M_{n,q}$ and $\Xi_{n,q}$ have the same limit distribution.

References

[1] M. Abadi, B. Saussol, Almost sure convergence of the clustering factor in α -mixing processes, Stochast. Dyn. 16(03) (2016), 166-176.

[2] M. Abadi, B. Saussol, Hitting and returning to rare events for all alpha-mixing processes, Stochastic Process. Appl. 121 (2010), 314-323.

[3] R. Arratia, L. Gordon, M. S. Waterman, An Extreme Value Theory for Sequence Matching, Ann. Statist. 14 (1986), 971-993.

[4] R. Arratia, L. Gordon, M. S. Waterman, The Erdos-Renyi Law in Distribution, for Coin Tossing and Sequence Matching, Ann. Statist. 18(2) (1990), 539-570.

[5] R. Arratia, M. Waterman, An Erdos-Reyni law with shifts, Adv. Math. 55 (1985), 13-23.

[6] J. Atnip, N. Haydn, S. Vaienti, Extreme Value Theory with Spectral Techniques: application to a simple attractor (2020), under revision, https://arxiv.org/abs/2002.10863

[7] V. Barros, L. Liao, J. Rousseau, On the shortest distance between orbits and the longest common substring problem, Adv. Math. 344 (2019), 311-339.

[8] V. Barros, J. Rousseau, Shortest distance between multiple orbits and generalized fractal dimensions, Ann. Henri Poincaré 22(6) (2021), 1853-1885.

[9] R. C. Bradley, Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions, Probab. Surveys 2 (2005), 107-144.

[10] R. Caballero, D. Faranda, G. Messori, A dynamical systems approach to studying midlatitude weather extremes, Geophys. Res. Lett. 44(7) (2017), 3346-3354.

[11] T. Caby, D. Faranda, G. Mantica, S. Vaienti, P. Yiou, Generalized dimensions, large deviations and the distribution of rare events, Phys. D 400 (2019), 132-143.

[12] T. Caby, D. Faranda, S. Vaienti, P. Yiou, Extreme value distributions of observation recurrences, Nonlinearity 34 (2021), 118.

[13] T. Caby, D. Faranda, S. Vaienti, P. Yiou, On the computation of the extremal index for time series, J. Stat. Phys. 179 (2020), 1666–1697.

[14] A. Coutinho, R. Lambert, J. Rousseau, Matching strings in encoded sequences, Bernoulli 26(3) (2020), 2021-2050.

[15] A. Coutinho, J. Rousseau, B. Saussol, Large deviation for return times, Nonlinearity 31(11) (2018), 5162-5179.

[16] D. Faranda, H. Ghoudi, P. Guiraud, S. Vaienti, Extreme value theory for synchronization of coupled map lattices, Nonlinearity 31 (2018), 26–58.

[17] D. Faranda, G. Messori, P. Yiou, Dynamical proxies of North Atlantic predictability and extremes, Sci. rep. 7 (2017), 41-78.

[18] D. Faranda, A. C. Moreira Freitas, J. M. Milhazes Freitas, M. Holland, T. Kuna, V. Lucarini, M. Nicol, M. Todd, S. Vaienti, Extremes and Recurrence in Dynamical Systems, Wiley, New York (2016).

[19] D. Faranda, S. Vaienti, Correlation dimension and phase space contraction via extreme value theory, Chaos 28 (2018), 041103.

[20] A. C. M. Freitas, J. M. Freitas, On the link between dependence and independence in extreme value theory for dynamical systems, Statist. Probab. Lett. 78(9) (2008), 1088–1093.

[21] A. C. M. Freitas, J. M. Freitas, J. V. Soares, Rare events for product fractal sets, J. Phys. A: Math. Theor. 54 (2021), 345202.

[22] M. Carvalho, A. C. M. Freitas, J. M. Freitas, M. Holland, M. Nicol, Extremal dichotomy for uniformly hyperbolic systems, Dyn. Syst. 30(4) (2015), 383–403.

[23] A. C. M. Freitas, J. Freitas, M. Todd, Hitting Time Statistics and Extreme Value Theory, Probab. Theory Related Fields 147(3)(2010), 675-710.

[24] A. C. M. Freitas, J. M. Freitas, M. Todd, Speed of convergence for laws of rare events and escape rates, Stochastic Process. Appl. 125(4) (2015), 1653–1687.

[25] https://www.mathworks.com/help/stats/gevfit.html

[26] M. Hénon, A two-dimensional mapping with a strange attractor, Commun. Math. Phys. 50(1) (1976), 69-77.

[27] B. R. Hunt, V. Kaloshin, How projections affect the dimension spectrum of fractal measures, Nonlinearity 10 (1997), 10-31.

[28] B. R. Hunt, T. Sauer, J. A. Yorke, Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces, Bull. Amer. Math. Soc. 27(2) (1992), 217–238.

[29] G. Keller, Rare events, exponential hitting times and extremal indices via spectral perturbation, Dyn. Syst. 27(1) (2012), 11–27.

[30] G. Keller, C. Liverani, Rare events, escape rates and quasistationarity: some exact formulae J. Stat. Phys. 135 (2009), 519–534.

[31] C. Neuhauser, A Phase Transition for the Distribution of Matching Blocks, Combinatorics, Probability and Computing 5 (1996), 139-159.

[32] J. Rousseau, Hitting time statistics for observations of dynamical systems, Nonlinearity 27 (2014), 23-77.

[33] J. Rousseau, B. Saussol, Poincaré recurrence for observations, Trans. AMS 362(11) (2010), 5845–5859.

[34] F. Takens, Detecting strange attractors in turbulence. In: Rand D., Young LS. (eds) Dynamical Systems and Turbulence, Warwick 1980. Lecture Notes in Mathematics, vol 898. Springer, Berlin, Heidelberg.

[35] M. Waterman, Introduction to Computational Biology, Maps, Sequences and Genomes, Chapman and Hall/CRC, New York (1995).