Matching strings in encoded sequences

Adriana Coutinho[1] & Rodrigo Lambert[2] & Jérôme Rousseau[3]

[1] Universidade de São Paulo (Brazil)
[2] Universidade Federal de Uberlândia (Brazil)
[3] Faculdade de Ciências da Universidade do Porto / Universidade Federal da Bahia (Portugal /Brazil)

adrianacoutinho@usp.br, rodrigolambert@ufu.br, jerome.rousseau@ufba.br

IME-USP

Abstract

We investigate the length of the longest common substring for encoded sequences and its asymptotic behaviour. The main result is a strong law of large numbers for a re-scaled version of this quantity, which presents an explicit relation with the Rényi entropy of the source. We apply this result to the zero-inflated contamination model and the stochastic scrabble. In the case of dynamical systems, this problem is equivalent to the shortest distance between two observed orbits and its limiting relationship with the correlation dimension of the pushforward measure.

Introduction

Let χ (respectively $\tilde{\chi}$) be an alphabet, $\Omega = \chi^{\mathbb{N}}$ (respectively $\tilde{\Omega} = \tilde{\chi}^{\mathbb{N}}$) the space of all sequences with symbols in χ (respectively $\tilde{\chi}$).

Definition 0.1. *Let $f : \Omega \to \tilde{\Omega}$ be a code. Given two sequences $x, y \in \Omega$, we define the n-length of the longest common substring for the encoded pair $(f(x), f(y))$ by*

$$M_n^f(x,y) = \max\left\{k \,:\, f(x)_i^{i+k-1} = f(y)_j^{j+k-1} \text{ for some } 0 \le i,j \le n-k\right\},$$

where $f(x)_i^{i+k-1}$ and $f(y)_j^{j+k-1}$ denote the substrings of length k beginning in $f(x)_i$ and $f(y)_j$ respectively.

Definition 0.2. *Consider the dynamical system $(\Omega, \mathbb{P}, \sigma)$. We say that it is α-mixing if there exists a function $\alpha : \mathbb{N} \to \mathbb{R}$ where $\alpha(g)$ converges to zero when g goes to infinity and such that*

$$\sup_{A\in\mathcal{F}_0^n \,;\, B\in\mathcal{F}_0^m} \left|\mathbb{P}\left(A \cap \sigma^{-g-n}B\right) - \mathbb{P}(A)\mathbb{P}(B)\right| \le \alpha(g)\,,$$

for all $m, n \in \mathbb{N}$.

We say that the system is ψ-mixing if there exists a function $\psi : \mathbb{N} \to \mathbb{R}$ where $\psi(g)$ converges to zero when g goes to infinity and such that

$$\sup_{A\in\mathcal{F}_0^n \,;\, B\in\mathcal{F}_0^m} \left|\frac{\mathbb{P}\left(A \cap \sigma^{-g-n}B\right) - \mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)\mathbb{P}(B)}\right| \le \psi(g),$$

for all $m, n \in \mathbb{N}$. In the cases that $\alpha(g)$ or $\psi(g)$ decreases exponentially fast to zero, we say that the system has an exponential decay.

The Rényi entropy of the pushforward measure is defined by

$$H_2(f_*\mathbb{P}) = -\lim_{k\to\infty}\frac{1}{k}\log\sum_{C_k} f_*\mathbb{P}(C_k)^2,$$

where the sums are taken over all k-cylinders.

Principal Result

Theorem 0.3. *Consider $f : \Omega \to \tilde{\Omega}$ a code such that $\underline{H}_2(f_*\mathbb{P}) > 0$. For $\mathbb{P}\otimes\mathbb{P}$-almost every $(x, y) \in \Omega \times \Omega$,*

$$\overline{\lim_{n\to\infty}}\frac{M_n^f(x,y)}{\log n} \le \frac{2}{\underline{H}_2(f_*\mathbb{P})}. \tag{1}$$

Moreover, if

(i) *the system $(\Omega, \mathbb{P}, \sigma)$ is α-mixing with an exponential decay (or ψ-mixing with $\psi(g) = g^{-a}$ for some $a > 0$);*

(ii) *$C_n \in \tilde{\mathcal{F}}_0^n$ implies $f^{-1}C_n \in \mathcal{F}_0^{h(n)}$, where $h(n) = o(n^\gamma)$, for some $\gamma > 0$,*

then, for $\mathbb{P}\otimes\mathbb{P}$-almost every $(x, y) \in \Omega \times \Omega$,

$$\underline{\lim_{n\to\infty}}\frac{M_n^f(x,y)}{\log n} \ge \frac{2}{\overline{H}_2(f_*\mathbb{P})}. \tag{2}$$

Therefore, if the Rényi entropy exists, we get for $\mathbb{P}\otimes\mathbb{P}$-almost every $(x, y) \in \Omega \times \Omega$,

$$\lim_{n\to\infty}\frac{M_n^f(x,y)}{\log n} = \frac{2}{H_2(f_*\mathbb{P})}. \tag{3}$$

Aplication

Suppose that each letter $a \in \chi$ is associated to a weight $v(a) \in \mathbb{N}^*$. Denote the score of a string z_0^{m-1} by $V(z_0^{m-1}) = \sum_{j=0}^{m-1} v(z_j)$. If x and y are two realizations of the χ-valued stochastic processes $(X_n)_n$ and $(Y_n)_n$,

$$V_n(x,y) = \max_{0\le i,j\le n-m}\left\{V(z_0^{m-1}) : \exists\, 1 \le m \le n \text{ s. t. } z_0^{m-1} = x_i^{i+m-1} = y_j^{j+m-1}\right\}$$

is the n^{th} highest-scoring matching substring [1]. For two copies independently generated by the same Markov source $\mathbb{P}$ with positive transition probabilities $[p_{ij}]$, the authors stated that:

$$\lim_{n\to\infty}\frac{V_n}{\log n} = \frac{2}{-\log p} \qquad \mathbb{P}\otimes\mathbb{P} - \text{a.s.}, \tag{4}$$

where $p \in (0,1)$ is the largest eigenvalue of $P = \left[p_{ij}^2\right]$.

One can observe that this result (4) can be obtained as particular case of Theorem 0.3. Indeed, inspired by [1], we can construct a specific code f such that

$$\begin{aligned} f\,:\ & \chi^{\mathbb{N}} \to \chi^{\mathbb{N}} \\ & x_0^\infty \mapsto \underbrace{x_0x_0\cdots x_0}_{v(x_0)}\underbrace{x_1x_1\cdots x_1}_{v(x_1)}\cdots\underbrace{x_nx_n\cdots x_n}_{v(x_n)}\cdots \end{aligned} \tag{5}$$

With this particular code, we get that $M_n^f(x,y) = V_n(x,y)$ and thus to get (4) we need to compute $H_2(f_*\mathbb{P})$ and check that conditions (i) and (ii) are satisfied. We recall that if (X_n) is a Markov chain in $\chi = \{1, 2, \cdots, d\}$, we can see $f(X_n)$ as a Markov Chain in $\tilde{\chi}$, which is a $(\sum_{i\in\chi} v(i))$-sized alphabet, given by

$$\tilde{\chi} = \left\{1_1, 1_2, \cdots, 1_{v(1)}, 2_1, 2_2, \cdots, 2_{v(2)}, \cdots, d_1, d_2, \cdots, d_{v(d)}\right\}.$$

In this context, we will consider that $f : \chi^{\mathbb{N}} \to \tilde{\chi}^{\mathbb{N}}$. Then, if $Q = [Q_{ij}]$, $1 \le i, j \le d$ is the transition matrix for (X_n) we get that the transition matrix Q^* for the chain $(f(X_n))$ on $\tilde{\chi}$ is related to Q. Notice that

- If $v_{min} = \min_{i\in\chi}\{v(i)\}$ is the minimum weight, we get for any cylinder C_n, $f^{-1}C_n \in \mathcal{F}_0^{\left\lfloor\frac{n}{v_{min}}\right\rfloor}$, and since $n/v_{min} = o(n^{1+\epsilon})$ for all $\epsilon > 0$, condition (ii) of Theorem 0.3 is then satisfied.
- An irreducible and aperiodic positive recurrent Markov chain is an α-mixing process with exponential decay of correlation which implies condition (i).
- The Rényi entropy of its stationary measure μ is given by $H_2(\mu) = -\log p$, where p is the largest positive eigenvalue of the matrix $\left[(Q^*)_{ij}^2\right]$, $1 \le i, j \le (\sum_{i\in\chi} v(i))$. In general, $f_*\mathbb{P}$ is not stationary. However, we prove that $H_2(\mu) = H_2(f_*\mathbb{P})$.

Finally, we can combine it with equation (3) in Theorem 0.3 to conclude the same result as in (4).

References

[1] R. Arratia, P. Morris and M. Waterman, *Stochastic scrabble: Large deviations for sequences with scores,* J. Appl. Prob. **25** no 1 (1988), 106-119.

[2] R. Arratia and M. Waterman, *An Erdös-Rényi Law with Shifts,* Adv. Math. **55** (1985), 13-23.

[3] V. Barros, L. Liao and J. Rousseau, *On the shortest distance between orbits and the longest common substring problem,* Adv. Math., **334** (2019), 311-339.

[4] N. Garcia and L. Moreira, *Stochastically perturbed chains of variable memory,* J. Stat. Phys **159**, no 5 (2015), 1107-1126.

[5] N. Haydn and S. Vaienti, *The Rényi entropy function and the large deviation of short return times,* Ergodic Theory Dynam. Systems **30** (2010), no. 1, 159-179.

Acknowledgements

This work is partially supported by CAPES, CNPQ and FAPESB.