Possible directions for future RAPA2 studies on probabilistic sources

Brigitte Vallée

CNRS and Université de Caen

December, 8, 2021

Two types of studies .....

for a source on the alphabet $\mathcal{A}$

associated with a numeration system on $\mathcal{I} = [0, 1]$

... or a sequence $\mathcal{P}_n$ of partitions

$$x \in \mathcal{I} \quad \longrightarrow \quad x = (a_1(x), a_2(x), \ldots, a_k(x), \ldots) \qquad a_i(x) \in \mathcal{A}$$

First study (Paper '20) : the average depth of a trie built on this source

with Valérie Berthé, Eda Cesaratto, Frédéric Paccaut, Martin Safe, Pablo Rotondo, Brigitte Vallée

Tools from analytic combinatorics

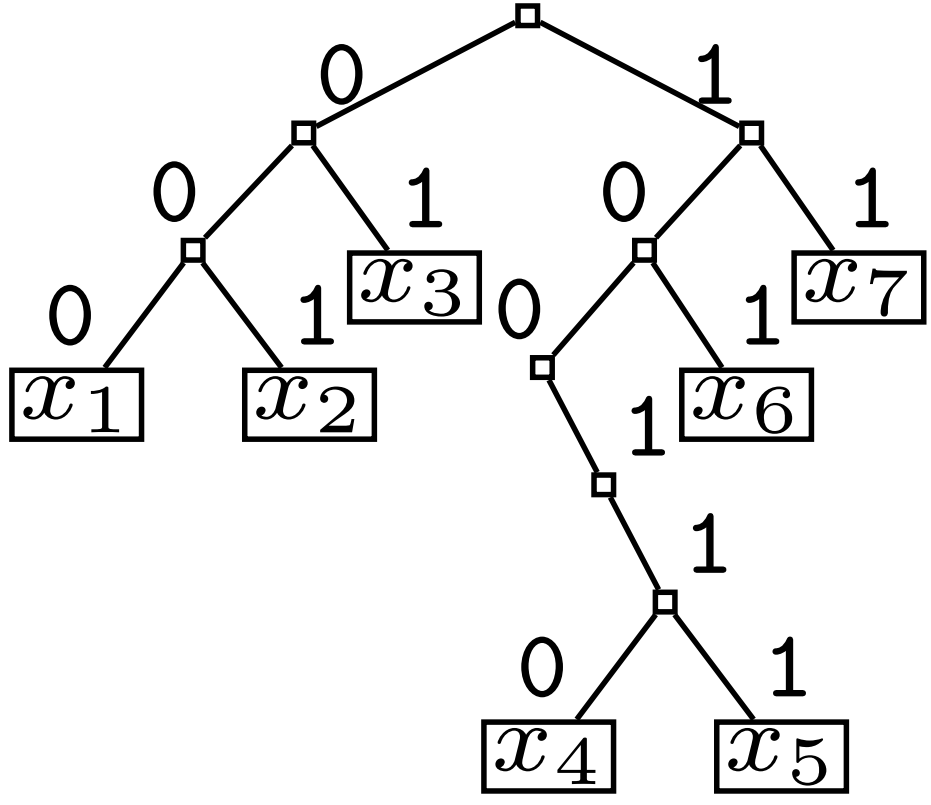
Second study (Paper '21): change of basis (between two sources)

with Valérie Berthé, Eda Cesaratto, Martin Safe, Pablo Rotondo

Tools from probability and ergodic theory

These studies focus on two particular sources with zero entropy.
I try to compare the two studies (for general sources)
and introduce natural sources with zero entropy

First study: the average depth of a trie (Paper '20)



Example of a trie built on seven (infinite) words on the alphabet $\{0, 1\}$

First study: the average depth of a trie

Main role in the analysis: fundamental intervals and their length

$$\mathcal{I}_w := \{x \mid (a_1(x), a_2(x), \ldots, a_k(x)) = w\}, \qquad I_w := |\mathcal{I}_w|,$$

the interval that gathers all the reals x with the same prefix $w \in \mathcal{A}^k$.

....together with the $\Lambda(s)$ series of the source of Dirichlet type

$$\Lambda_k(s) := \sum_{w \in \mathcal{A}^k} I_w^s \qquad \Lambda(s) := \sum_{w \in \mathcal{A}^\star} I_w^s \qquad (s \in \mathbb{C})$$

Result (Paper '20). The series $\Lambda(s)$ intervenes via its dominant singularity:

- its position : the smallest $s = s_0$ for which $\Lambda(s)$ is well defined
- the type of singularity (pole, etc...); for instance, a pole of order $k_0 \geq 1$

With "tameness", it provides an estimate of the average trie depth

$$\mathbb{E}[D_n] = \Theta(n^{s_0 - 1}) \cdot (\log n)^{\ell_0}$$

with $\ell_0 = k_0 - 1$ (s_0 non integer), $\ell_0 = k_0$ (s_0 integer)

The average depth of a trie. Instances of application of the result

Result. $\mathbb{E}[D_n] = \Theta(n^{s_0-1}) \cdot (\log n)^{\ell_0}$

with $\ell_0 = k_0 - 1$ (s_0 non integer), $\ell_0 = k_0$ (s_0 integer)

First instances : the binary source

$$\Lambda_k(s) = 2^k 2^{-ks} = 2^{k(1-s)} \qquad \Lambda(s) = \frac{1}{1-2^{1-s}}$$

and all the "good" sources –with positive entropy H– for which

$$s_0 = 1, \quad k_0 = 1, \quad H = -\lim_{k\to\infty} \frac{1}{k} \sum_{w\in\mathcal{A}^k} I_w \log I_w = -\frac{1}{k}\Lambda_k'(1)$$

For these instances. $\mathbb{E}[D_n] \sim \frac{1}{H} \log n$

Other instances given by our two favorite sources of zero entropy....

Sturm source: $s_0 = 3/2, \quad k_0 = 1 \qquad \mathbb{E}[D_n] = \Theta(n^{1/2})$

Stern-Brocot source: $s_0 = 1, \quad k_0 = 2 \qquad \mathbb{E}[D_n] = \Theta(\log^2 n)$

Question. Find (natural) sources (of zero entropy) with a given behaviour of their Λ series, and thus with a given behaviour of their average trie depth.

Second study : Change of basis and weight of sources

Main role in the analysis : Coincidence intervals and their lengths

$$\mathcal{I}_k(x) = \{y \in \mathcal{I} \mid \ y \text{ has the same expansion as } x \text{ until depth } k\}$$

$$I_k(x) := |\mathcal{I}_k(x)|$$

$x \mapsto I_k(x)$ has value I_w on the interval $\mathcal{I}_w$ $(w \in \mathcal{A}^k)$

Main interest in the random variable $-\log I_k(x)$

informally speaking: related with the growth rate of $I_k(x)$

Two (or three?) notions of weight : a.e, in measure, on average

Definition. The source has a weight $f(k)$ [a.e, in measure, on average]

$$\iff \quad -\log I_k(x) \sim_{k\to\infty} f(k) \quad \text{[a.e, in measure, on average]}$$

A weight $f(k)$ in measure

$$\forall \epsilon > 0, \lim_{k\to\infty} \left| \{x \in \mathcal{I} \mid \left| \frac{-\log I_k(x)}{f(k)} - 1 \right| > \epsilon\} \right| = 0$$

A weight $f(k)$ on average –or in L^1– $\quad \mathbb{E}[-\log I_k(x)] \sim_{k\to\infty} f(k)$

Three notions of weights

For the sequence of random variables $\log I_k(x)$

$$\text{Convergence a.e} \Longrightarrow \text{Convergence in measure}$$
$$\Uparrow$$
$$\text{Convergence in } L^1$$

The third notion (average weight) not studied in the Paper '21.
Yet, it is well adapted for relating the two studies,
and perhaps easier to deal with for natural sources.

Why ? $x \mapsto I_k(x)$ is a random variable with value I_w on the interval $\mathcal{I}_w$
The following expectations are thus expressed with the Λ series

$$\mathbb{E}[I_k^s(x)] = \sum_{w\in\mathcal{A}^k} I_w^s \cdot I_w = \Lambda_k(s+1), \qquad \mathbb{E}[\log I_k(x)] = \sum_{w\in\mathcal{A}^k} (\log I_w)\cdot I_w = \Lambda_k'(1)$$

The source has an average weight $f(k) \Longleftrightarrow \Lambda_k'(1) \sim_{k\to\infty} f(k)$
For a good source of positive entropy $\Lambda_k'(1) \sim_{k\to\infty} H \cdot k$
The average weight gives an extension of Shannon entropy for any source,
useful in particular for sources with zero entropy

The average weight, again

Questions:

- Find natural sources with a given average weight $f(k)$.
- Is it possible to directly relate (at least in some easy cases...)
  - the average weight $f(k)$
  - the abscissa of convergence of the Λ series

Another interpretation of the average weight:
A measure of the division speed of the sequence $(\mathcal{P}_n)$ of partitions

Begin with the binary partition:
At time k, the partition $\mathcal{B}_k$ is formed with 2^k intervals of length 2^{-k}

Change the time scale and define $\mathcal{P}_k = \mathcal{B}_{\lfloor g(k) \rfloor}$
The partition $\mathcal{P}_k$ is formed with $2^{\lfloor g(k) \rfloor}$ intervals of length $2^{-\lfloor g(k) \rfloor}$.

Its Λ series are $\Lambda_k(s) = 2^{\lfloor g(k) \rfloor (1-s)}$ $\Lambda(s) = \sum_{k \geq 0} 2^{\lfloor g(k) \rfloor (1-s)}$

Building a source with a given average trie depth

$$\mathbb{E}[D_n] = \Theta(n^{1/a}) \cdot (\log n)^b \qquad (a \in \mathbb{R}^+,\ b \text{ integer})$$

Take the function $g(k)$ defined by $2^{\lfloor g(k) \rfloor} = \lfloor k^a \rfloor, \quad a \in \mathbb{R}^+.$
Forget the integer parts (!!?), $g(k) \sim (a/\log 2) \log k$.
Average weight$= g(k) \log 2 = a \log k$

$\Lambda(s) = \zeta(-a(1-s))$ $\qquad s_0 = 1 + (1/a)$
$\mathbb{E}[D_n] = \Theta(n^{1/a}) \log n$ $\qquad$ or $\qquad \mathbb{E}[D_n] = \Theta(n^{1/a})$
(depending if $1/a$ is integer or not)

More generally, take the function $g(k)$ defined by
$2^{\lfloor g(k) \rfloor} = \lfloor (\log n)^b \cdot k^a \rfloor, \quad a \in \mathbb{R}^+, \quad b$ integer
Forget the integer parts (!!?) $g(k) \sim (1/\log 2)(a \log k + b \log\log k)$
Average weight $= g(k) \log 2 = a \log k + b \log\log k$

$\Lambda(s) = \zeta^{(b)}(-a(1-s)),$ $\qquad s_0 = 1 + (1/a), \quad k_0 = b + 1$
$\mathbb{E}[D_n] = \Theta(n^{1/a}) \cdot (\log n)^b$ $\qquad$ or $\quad \mathbb{E}[D_n] = \Theta(n^{1/a}) \cdot (\log n)^{b+1}$
(depending if $1/a$ is integer).

A natural class of sources
Variable Length Markov Chains (VLMC).

- The VLMC are the simplest sources
  where the dependency from the past may be unbounded;
- A good compromise between
  - the simple definition of this model
  - the expressivity of the model (used for instance in musicology)
- The depth of suffix tries has already been studied
  on some subclasses of VLMC's, with probabilistic methods
  (see the works of Frédéric P. with his co-authors)
- A future work? Perform the two previous studies in this model
  - Analyze the average trie depth on this class
  - Determine the (average) weight of these sources

An example of a VLMC source of intermittent type

$\mathcal{R}_k := \{$the prefix ends with a sequence of exactly $k \geq 0$ occurrences of $0\}$

$$\Pr[0 \mid \mathcal{R}_0] = \Pr[1 \mid \mathcal{R}_0] = \frac{1}{2},$$

and for $k \geq 1$, with a function $a(k) \to 1$ for $k \to \infty$

$$\Pr[0 \mid \mathcal{R}_k] = a(k) \quad \text{and thus} \quad \Pr[1 \mid \mathcal{R}_k] = 1 - a(k)$$

Some facts:

- The probability of emitting a symbol only depends on the longest run of zeroes that has been just emitted
- When there are many zeroes that have been just emitted, the probability of emitting '0' is large (and emitting '1' is small)
- When the symbol 1 is (finally) emitted, the last prefix belongs to $\mathcal{R}_0$, and the source "restarts" as at the beginning

With the fundamental decomposition $\{0,1\}^\star = (0^\star 1)^\star \cdot 0^\star$,
any finite word emitted by the source is written as

$$(00.....1) \quad (00...1) \quad ... \quad (00...1) \quad (0.....0)$$

An example of a VLMC source of intermittent type

$$(00.....1) \quad (00...1) \quad ... \quad (00...1) \quad (0.....0)$$

The source is completely defined by the two families of probabilities

$p_k = \pi_{0^{k-1}1}$ for $k \geq 1$, $\quad q_k = \pi_{0^k}$ for $k \geq 0$ $\quad$ with $\quad q_k + p_k = q_{k-1}$

$$q_k = \pi_{0^k} = \prod_{i=0}^{k-1} \Pr[0 \mid \mathcal{R}_i] = \frac{1}{2} \prod_{i=1}^{k-1} a(i)$$

The Λ series of the source is written as a product

$$\left[\frac{1}{1 - \sum_{k \geq 1} p_k^s} \right] \cdot \left[\sum_{k \geq 0} q_k^s \right]$$

The red term is a quasi-inverse because the red sequences are i.i.d.