

**DISTRIBUTION OF
AVERAGE EXTERNAL DEPTH
FOR TRIES
IN DYNAMICAL SOURCES CONTEXT**

Logic, Computability and Randomness 2007

Eda Cesaratto

Facultad de Ingeniería

Universidad de Buenos Aires, Argentina

Brigitte Vallée

GREYC Dépt. d'Informatique

Université de Caen, France

Trie: tree structure used as a dictionary

Given

– an alphabet $\mathcal{M} = \{m_1, \dots, m_r\}$ possibly infinite

and

– a finite set X of words, each word of X is a sequence of symbols taken from \mathcal{M} ,

trie(X) is defined recursively by

– If $|X| = 0$, $\text{trie}(X) = \emptyset$

– If $|X| = 1$, $X = \{\mathbf{m}\}$, $\text{trie}(X)$ is a leaf labeled by \mathbf{m} .

– If $|X| \geq 2$,

$$\text{trie}(X) = \langle \text{trie}(X \setminus m_1), \dots, \text{trie}(X \setminus m_r) \rangle$$

where $X \setminus m$ means the subset of X consisting of strings that start with the symbol m stripped of their initial symbol m .

Example

$$X = \{m_1, m_2, m_3, m_4\}$$

$$m_1 = 11100\dots$$

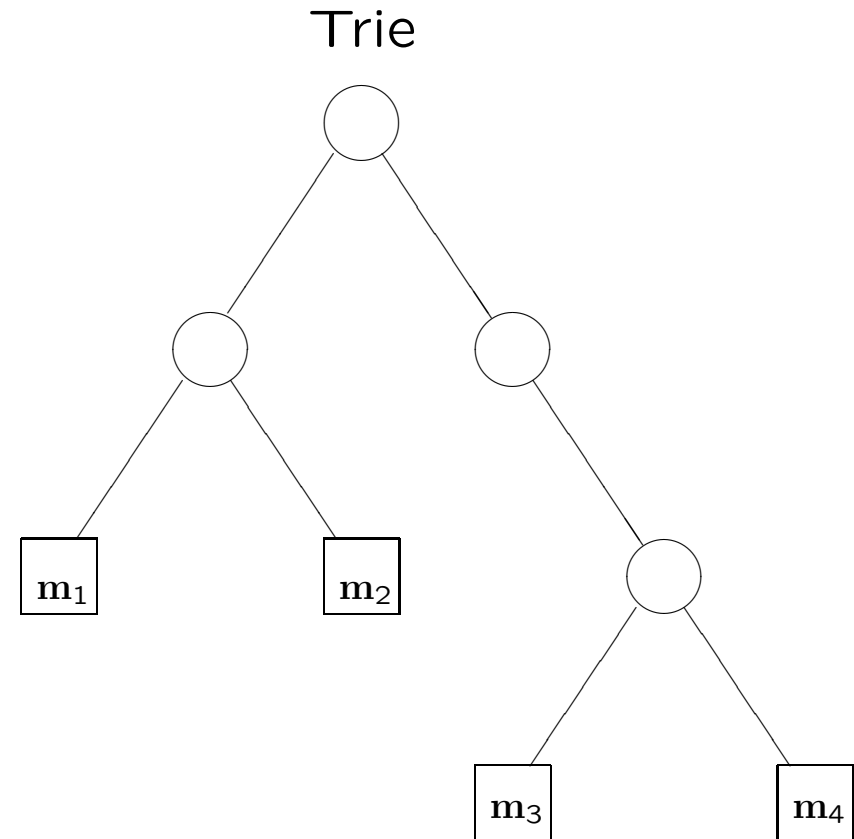
$$m_2 = 10111\dots$$

$$m_3 = 00110\dots$$

$$m_4 = 00001\dots$$

External depth of leaves:

$$D_n(1) = D_n(2) = 2 \text{ and } D_n(3) = D_n(4) = 3 \text{ (} n := |X| = 4 \text{)}$$



Parameters on tries

The complexity of many algorithms on strings can be expressed with various parameters on tries.

Parameter analyzed here: The **average external depth**

suppose that X is a set of n strings.

The **average external depth** D_n is the number of nodes of in a path from the root to a uniformly randomly selected leaf.

Probabilistic models on strings

Probabilistic sources:

Words are generated by an infinite sequence of random variables

$$M := \{M_k\}_{k=1}^{\infty}$$

defined over an alphabet $\mathcal{M} = \{m_1, \dots, m_r\}$ possibly infinite.

Classical examples of probabilistic sources

– **Memoryless:** the random variables $\{M_k\}_{k=1}^{\infty}$ are independently, identically distributed, that is, for a fixed $m_i \in \mathcal{M}$ we have

$$\mathbb{P}[M_k = m/M_{k-1} = n, \dots, M_1 = m_1] = \mathbb{P}[M_k = m] := p_m \text{ for all } k.$$

Unbiased memoryless model: $|\mathcal{M}| < \infty$ and $p_m = \frac{1}{|\mathcal{M}|}$ for all i .

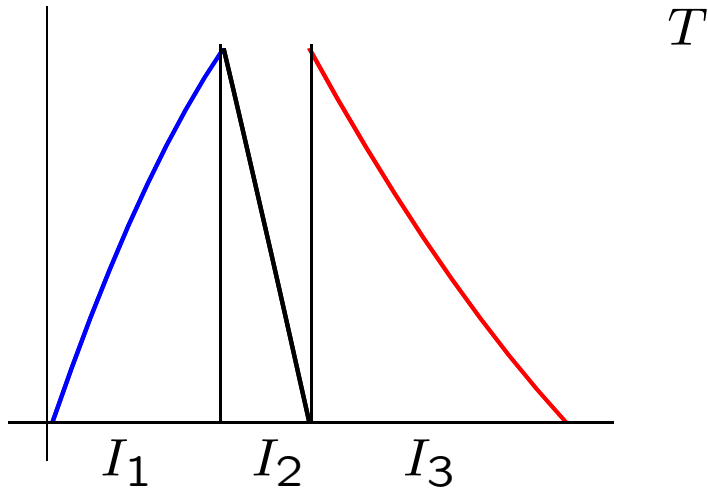
– **Bounded memory or Markovian sources:**

the sequence of random variables $\{M_k\}_{k=1}^{\infty}$ is a stationary Markov process with an unique, invariant measure.

Here we consider Markov processes of degree one

$$\mathbb{P}[M_k = m/M_{k-1} = n, \dots, M_1 = m_1] = \mathbb{P}[M_k = m/M_{k-1} = n] := p_{m,n} \text{ for all } k$$

Dynamical Sources [Vallée, 2001]



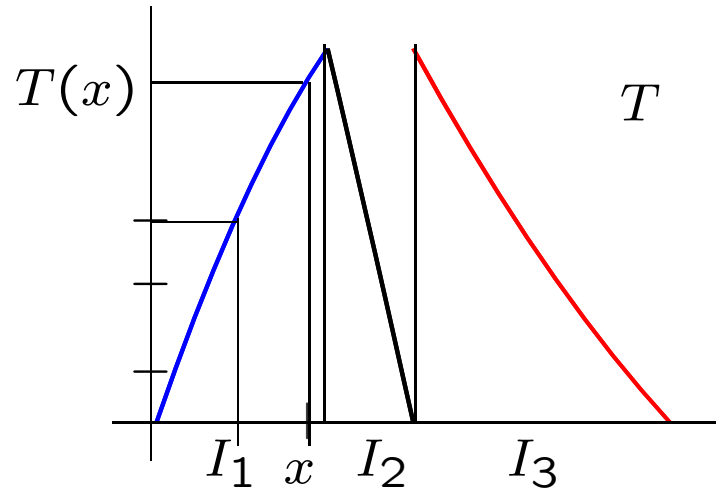
- An expansive map $T : [0, 1] \rightarrow [0, 1]$
- The alphabet \mathcal{M}
- A topologic partition $(I_m)_{m \in \mathcal{M}}$ of $[0, 1]$, that is, $[0, 1] = \cup_{m \in \mathcal{M}} \overline{I_m}$
- A codification function $\sigma : [0, 1] \rightarrow \mathcal{M}$, constant over each I_m ($\sigma|_{I_m} = m$)

Construction of the Source

$$x \rightarrow (x, Tx, T^2x, \dots, T^kx, \dots) \quad m_k(x) := \sigma(T^{k-1}x)$$

$$M_0 = \mathcal{U}[0, 1] \quad M_k := \sigma(T^{k-1}(M_0))$$

Schemes for representing reals numbers



$$x \in I_1$$

$$T(x) \in I_3$$

$$x \in I_{13}$$

$$\overline{I_{13}}$$

Inverse branches T : h_1, h_2, h_3

Fundamental intervals:

$$I_{\mathbf{m}} = I_{m_1, \dots, m_k} = h_{m_1} \circ h_{m_2} \dots \circ h_{m_k}([0, 1])$$

$$x = \bigcap_{k \in \mathcal{M}} I_{m_1(x), \dots, m_k(x)}$$

Hypothesis on the dynamical system

1. T is a bijection from I_m to $]0, 1[$, and denote by $\mathcal{H} := \{h_m, m \in \mathcal{M}\}$ the set of the inverse branches of T
2. T is strongly expanding:
there exist $0 < \alpha_m, \delta_m < 1$: $\alpha_m \leq |h'_m(x)| \leq \delta_m \forall x \in [0, 1]$.
3. Bounded Distortion:
 $\sup\left\{\frac{|h''_m(x)|}{|h'_m(x)|} : x \in I_m, m \in \mathcal{M}\right\} < \infty$.

Memoryless: linear branches

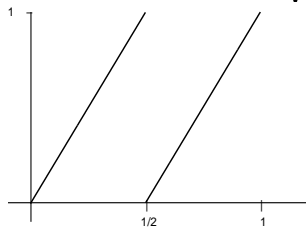
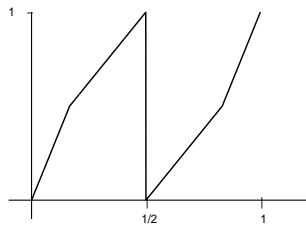


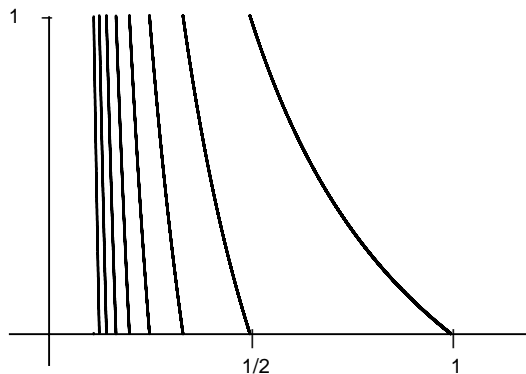
Figure: Binary expansion $T_2 = 2x - \lfloor 2x \rfloor$
 $\sigma(x) = \lfloor 2x \rfloor$

Bounded Memory or Markovian processes: piecewise linear branches



Infinite Memory: non zero curvature

Figure: Continued fractions



$$T_{CF}(x) = \begin{cases} \frac{1}{x} - \lfloor \frac{1}{x} \rfloor & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

$$\sigma(x) = \lfloor \frac{1}{x} \rfloor$$

$$x = \frac{1}{m_1 + \frac{1}{m_2 + \dots}}$$

The problem of this talk

Assumptions.

- The number of strings of X is fixed and equal to n .
- Strings of X are statistically independent.
- Symbols from the alphabet \mathcal{M} are generated by a dynamical source.

Our aim is to study the asymptotic distribution of the average external depth of $\text{trie}(X)$, denoted here by D_n , when $n \rightarrow \infty$,

i. e. $\lim_{n \rightarrow \infty} \mathbb{P}[D_n \leq x]$

Previous Results

[Jacquet, Regnier, Szpankowski, 1986-1991]

Finite alphabet, biased memoryless sources and markovian sources,

The average external depth of $\text{trie}(X)$ is asymptotically Gaussian,

$$\mathbb{E}[D_n] = \frac{1}{h} \log(n) + O(1)$$

where $h := -\lim_{n \rightarrow \infty} \sum_{|w|=n} |I_w| \log |I_w|$ is the entropy of the source,

$$\text{Var}[D_n] = c \log(n) + O(1).$$

Finite alphabet, unbiased memoryless sources,

The average external depth of $\text{trie}(X)$ is asymptotically double exponential.

Results on dynamical sources

Dynamical source, no restrictions on the alphabet

[Clément-Flajolet-Vallée, 2000]

$\mathbb{E}[D_n] = \frac{1}{h} \log(n) + O(1)$ where h is the entropy of the source

For the Gauss dynamical system, a more accurately expression of the expectation is possible because of the connection between continued fractions and the Riemann Zeta function.

Our result

Extra hypothesis on the source: Uniform non Integrability (UNI). This hypothesis essentially says that the branches of the dynamical systems are not affine and it entails infinite memory.

This hypothesis excludes memoryless and markovian sources

Theorem: *Consider a dynamical source which satisfies the UNI Condition. The average external depth of a trie built on n words produced by this dynamical source is asymptotically Gaussian,*

$$\mathbb{E}[D_n] = \frac{1}{h} \log(n) + O(1)$$

where h is the entropy of the source,

$$\text{Var}[D_n] = c \log(n) + O(1),$$

and a speed of convergence of order $1/\sqrt{\log n}$.

Main tools used in the proof

1. Moment generating functions with Hwang quasipowers theorem.
2. Extensions of the Ruelle transfer operator of the dynamical system.

Here the hypothesis UNI appears

3. Rice's Integrals.

Hwang quasi-powers theorem (1994)

Assume that the moment generating functions

$$G_n(w) := \mathbb{E}[\exp(wD_n)]$$

of a sequence of random variables D_n are analytic in $|w| < \delta$ for some $\delta > 0$, and satisfy there the expansion

$$\mathbb{E}[\exp(wD_n)] = \exp(\beta_n U(w) + V(w)) \left(1 + O\left(\frac{1}{\kappa_n}\right)\right),$$

for $\beta_n, \kappa_n \rightarrow \infty$ as $n \rightarrow \infty$, and $U(w), V(w)$ are analytic in $|w| \leq \delta$. Assume also that $U''(0) \neq 0$. Then

$$\mathbb{E}[D_n] = \beta_n U'(0) + V'(0) + O(\kappa_n^{-1})$$

$$\mathbb{V}[D_n] = \beta_n U''(0) + V''(0) + O(\kappa_n^{-1})$$

$$\mathbb{P}\left[\frac{D_n - \beta_n U'(0)}{\sqrt{\beta_n U''(0)}} < x\right] = \Phi(x) + O\left(\frac{1}{S_n}\right)$$

where $\Phi(x)$ is the standard normal distribution and $S_n = \min(\sqrt{\beta_n}, \kappa_n)$.

Probability and length of fundamental intervals

The moment generating function verifies

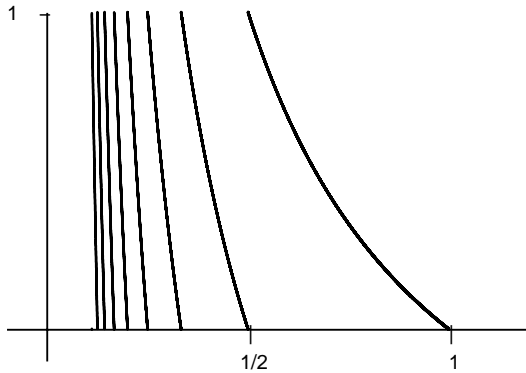
$$\mathbb{E}[\exp(wD_n)] = \exp w - \frac{1 - \exp w}{n} \sum_{k=1}^{\infty} \mathbb{P}[D_n \geq k + 1] \exp w$$

and

$$\mathbb{P}[D_n \geq k + 1] = \sum_{|\mathbf{m}|=k} |I_{\mathbf{m}}| [1 - (1 - |I_{\mathbf{m}}|)^{n-1}].$$

Transfer operators

\mathcal{H} is the set of inverse branches of the map T .



Density Transformer:

$$\mathbf{H}[f](x) = \sum_{h \in \mathcal{H}} |h'(x)| f \circ h(x),$$

Transfer operator [Ruelle, 1978]:

$$\mathbf{H}_s[f](x) = \sum_{h \in \mathcal{H}} |h'(x)|^s f \circ h(x).$$

The secant transfer operator

$$\bar{\mathbf{H}}_s[F](x) = \sum_{h \in \mathcal{H}} \left| \frac{h(x) - h(y)}{x - y} \right|^s F(h(x), h(y)).$$

The k -th iterate satisfies

$$\bar{\mathbf{H}}_s^k[F](x) = \sum_{h \in \mathcal{H}^k} \left| \frac{h(x) - h(y)}{x - y} \right|^s F(h(x), h(y)).$$

Alternative expression of the moment generating function

The moment generating function can be written as

$$\mathbb{E}[\exp(wD_n)] = \exp(w) - \frac{1 - \exp(w)}{n} \sum_{\ell=1}^{n-1} (-1)^{\ell-1} \binom{n-1}{\ell} f_w(\ell)$$

where the function $f_w(s)$ involves the quasi-inverse of the secant Ruelle operator $\bar{\mathbf{H}}_s$

$$f_w(s) := (I - \exp(w)\bar{\mathbf{H}}_{s+1})^{-1} \circ \bar{\mathbf{H}}_{s+1}[\mathbf{1}](\mathbf{0}, \mathbf{1})$$

Rice Method

Let $\varphi(s)$ be

- analytic on $[0, \infty[$,
 - meromorphic on the half plane defined by $\Re(s) \geq 1 - \alpha$ for some $\alpha > 0$
 - and of polynomial growth in a neighborhood of ∞
- then, for n large enough,

$$\sum_{\ell=0}^n (-1)^\ell \binom{n-1}{\ell} \varphi(\ell) =$$

$$-(-1)^n \sum_s \operatorname{Res} \left[\varphi(s-1) \frac{n!}{(s+1)s(s-1)\dots(s-n-1)} \right] + O(n^{-\alpha})$$

where the sum is extended to all poles s in $\Re(s) > 1 - \alpha$ and not on $[0, \infty[$.

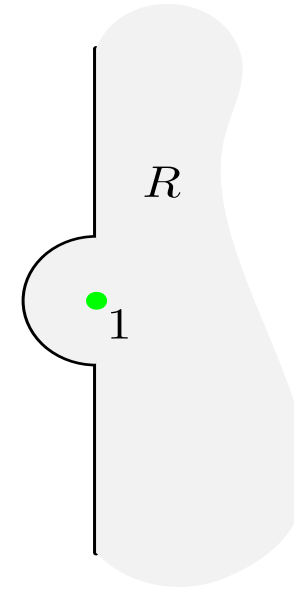
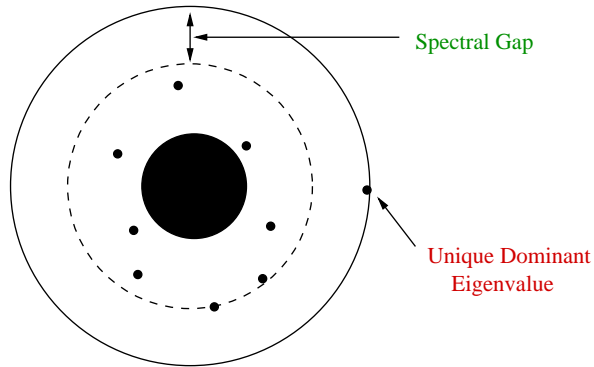
The main technical point: To prove that

$$f_w(s) := (I - \exp(w)\overline{\mathbf{H}}_s)^{-1} \circ \overline{\mathbf{H}}_s[1](0, 1)$$

verifies the hypothesis of the previous theorem.

Spectral properties of transfer operators

Near $s = 1$, the secant transfer operator has an unique dominant eigenvalue $\lambda(s)$ and a spectral gap

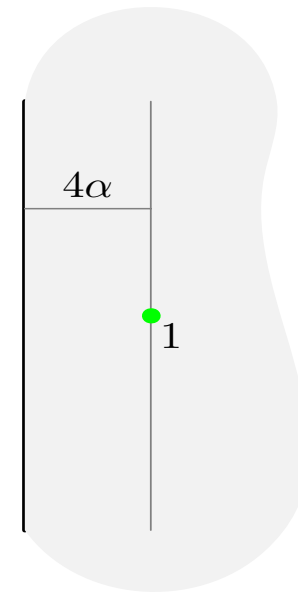


In an neighborhood of $s = 1$, the quasi-inverse $(I - \bar{\mathbf{H}}_s)^{-1}$ behaves as $\frac{1}{1-\lambda(s)}$.

So $f_w(s)$ behaves as $\frac{1}{1-\exp(w)\lambda(s)}$ for w near 0. Moreover, it is meromorphic on R with a simple pole at $s = \sigma(w)$.

Refined results about spectral properties

Under the hypothesis that the dynamical system satisfies the **UNI conditions**, extensions of previous results of Dolgopyat (1998) sharpened by Baladi-Vallée (2003) prove that $f_w(s)$ is under the hypothesis of the Rice method with an **unique simple pole** at $s = \sigma(w)$.



Domain of the $f_w(s)$
uniformly for w near 0.

End of the proof

– Rice method applies for $f_w(s)$.

The value of

$$\text{Res}(f_w(s), s = \sigma(w))$$

depends on the dominant spectral objects of the secant transfer operator.

– Perturbation Theory entails the analytic dependence on w .

A quasi-powers expression à la Hwang holds for the moment generating function of the random variables D_n .

– Hwang quasi-powers theorem entails the asymptotic Gaussian Law for D_n .

Conclusions

1. The average external depth of a trie built on n words produced by a dynamical source is asymptotically Gaussian with expectation and variance of the order of $\log n$.
2. This result extends previous results for sources with bounded memory, but the proof presented here does not apply for sources with bounded memory.
3. Is an unified proof possible?