Dynamical systems and average-case analysis of general tries

Brigitte Vallée
GREYC, Université de Caen, France

June 9, 1997

[summary by Julien Clément]

Abstract

The three major parameters of a *trie* (sometimes called *digital tree*), number of nodes, path length, and height, are analyzed precisely in a general context where words are emitted by a source associated to a dynamical system. The results can all be stated in terms of two intrinsic characteristics of the source: the entropy and the probability of letter coincidence. These characteristics themselves are linked in a natural way to spectral properties of a Ruelle operator associated to the dynamical system.

1. Probabilistic dynamical sources

1.1. **Definitions.** A dynamical source, in the context of information theory, is a mechanism which produces infinite words over an alphabet $\mathcal{M}$. Such a system is defined by four elements: (*i*) an alphabet $\mathcal{M}$ included in $\mathbb{N}$, (*ii*) a quasi partition of $\mathcal{I} =]0,1[$ with intervals $\mathcal{I}_m$, $m \in \mathcal{M}$, (*iii*) a mapping $\sigma : \mathcal{I} \to \mathcal{M}$ which is constant over each $\mathcal{I}_m$ and equal to m and finally (*iv*) a mapping $T : \mathcal{I} \to \mathcal{I}$ which satisfies two properties: the restriction of T to $\mathcal{I}_m$ is a real analytic bijection from $\mathcal{I}_m$ to $\mathcal{I}$; the mapping T is expansive, i.e. $|T'(x)| > 1$ on $\mathcal{I}$. The words emitted by the source are produced by iterating T and coded thanks to σ. The word $M(x)$ of $\mathcal{M}^\infty$ (an infinite sequence of symbols), where $x \in \mathcal{I}$, is formed with the symbols

$$M(x) := (\sigma(x), \sigma(T(x)), \sigma(T^2(x)), \ldots, \sigma(T^k(x)), \ldots).$$

Each letter of the alphabet is associated to a distinct branch of T or equivalently to a distinct inverse branch of T denoted by h_m. The bijection $h_m : \mathcal{I} \to \mathcal{I}_m$ coincides with the inverse of the restriction of T to $\mathcal{I}_m$.

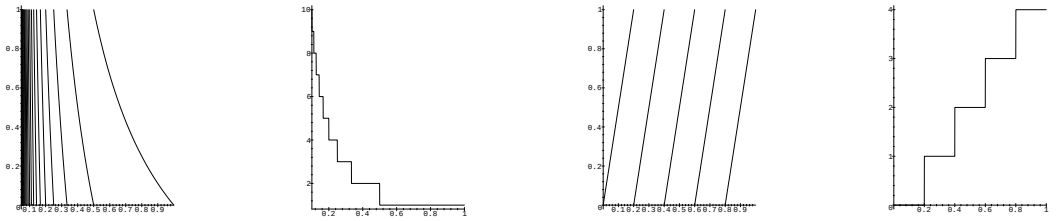

FIGURE 1. Graphical representation of the mappings T and σ for a source based on the continued fraction expansion (left) and for a memoryless source based on the 5-ary expansion of numbers (right).

1.2. **Examples.** A lot of probabilistic dynamical sources can be described in such a framework, including all *memoryless sources* where letters of the alphabet can be emitted with probabilities $\{p_i\}$ independently of previous letters. This gives a mapping T where branches are affine. In particular this encompasses the b-ary expansion model of numbers. *Markov sources* take into account a finite past for producing words and thus generalize memoryless sources. Finally, in a model where the source is based upon the *continued fraction* expansion of numbers, the alphabet is $\mathbb{N}$ and the probability for a character to be emitted depends on all the previous history.

1.3. **Fundamental intervals, entropy, coincidence probability.** Numbers which share the same prefix expansion $m_1, \ldots, m_k$ form an interval called *fundamental interval* of depth k which is exactly, with the preceding notations,

$$\mathcal{I}_{m_1,\ldots,m_k} = h_{m_1} \circ \cdots \circ h_{m_k}(\mathcal{I}).$$

In a general context, the interval $\mathcal{I}$ is endowed with a continuous density f (so that F denotes the associated distribution). Then the word $M(x)$ is produced according to the expansion process (using T and σ). In this context, the measure u_h of a fundamental interval $\mathcal{I}_h$ associated to $h = h_{m_1} \circ \cdots \circ h_{m_k}$ is

$$u_h := |F(h(0)) - F(h(1))|,$$

and plays a crucial role since it is the probability that a word begins with a certain prefix. During the analysis, two quantities relative to the source appear naturally. The *entropy* $h(\mathcal{S}, F)$ and the *coincidence probability* $c(\mathcal{S}, F)$ are defined as the limits

$$h(\mathcal{S}, F) := \lim_{k\to\infty} \frac{-1}{k} \sum_{|h|=k} u_h \log u_h, \qquad c(\mathcal{S}, F) := \lim_{k\to\infty} \left[\sum_{|h|=k} u_h \log u_h \right]^{1/k}.$$

It is interesting to note that these limits exist and are independent of the distribution F.

2. Tries associated to a general source

Let $\mathcal{M} \in \mathbb{N}$ be a set of elements called digits, and $\mathcal{M}^\infty$ the set of all infinite sequences built over $\mathcal{M}$. For any word produced by the dynamical source $M(x) = (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$ (with $x \in \mathcal{I}$), the head and tail functions are defined by

$$\text{head}(M(x)) = \sigma(x), \quad \text{tail}(M(x)) = M(Tx).$$

Any finite set of infinite words produced by the same source can be written as $M(X) = \{M(x) | x \in X\}$, and one associates to X a trie, Trie(X), defined by the following recursive rules

- (R1) If $X = \{x\}$ has cardinality equal to 1, then Trie(X) consists of a *single leaf node* that contains $M(x)$.
- (R2) If X has cardinality at least 2, then Trie(X) is *an internal node* represented generically by 'o' to which are attached ℓ subtrees, where $\ell = \text{card}\{\sigma(x) | x \in X\}$ is the number of different head symbols in $M(X)$. Let $b_1 < \cdots < b_\ell$ be these head symbols; Trie(X) is defined by

  $$\text{Trie}(X) = \langle o, \text{Trie}(X_1), \ldots, \text{Trie}(X_\ell) \rangle, \quad \text{where} \quad X_j = \{Tx \mid \sigma(x) = b_j, \quad x \in X\}.$$

  The trie Trie(X_j) collects all the suffixes of words that begin with b_j.

2.1. **Parameters.** The main parameters of a trie are *size* (number of internal nodes), *height* and *external length path*, which is the sum of all links from the root to each leaf.

The model considers here infinite strings emitted independently by the same dynamical source. Rather than considering a fixed number n of strings, a Poisson model of rate n is used, where the number of strings N is also a random variable which strongly concentrates around n. The strong property of independence of this particular model makes the analysis easier and gives access to the expectations of parameters. The expectations of height, size and external path length under a Poisson model of rate n and distribution F over $\mathcal{I}$ are respectively

$$D(n) = \sum_k [1 - \prod_{|h|=k} (1+nu_h)e^{-nu_h}], \quad S(n) = \sum_h [1-(1+nu_h)e^{-nu_h}], \quad P(n) = \sum_h nu_h[1-e^{-nu_h}].$$

2.2. **Asymptotics.** These quantities are easily recognized as *harmonic sums* of the form $F(x) = \sum_{k\in K} \lambda_k f(\mu_k x)$ (excepted for the height which needs a small calculation step before). The best tool to analyze the asymptotics of such harmonic sums is the *Mellin transform*, which leads to locating poles of the associated *Dirichlet series* $\Lambda(s) = \sum_{k\in K} \lambda_k \mu_k^s$. Here, the key quantity for the analysis of size and path length is the *series of fundamental intervals*,

$$\Lambda(F,s) = \sum_h u_h^s = \sum_h |F(h(0)) - F(h(1))|^s,$$

considered for complex values of s. For a general source, it is not always easy (or possible) to locate precisely the singularities. In this case a Tauberian theorem, under some constraints, can be used to extract the asymptotic expansion.

3. Generalized Ruelle operators

3.1. **Presentation.** The generalized Ruelle operators are derived from the original Ruelle operators, and involve secants of the inverse branches $H(u,v) := |(h(u)-h(v))/(u-v)|$. The generalized Ruelle operators are defined by

$$\mathbf{G}_s[F](u,v) = \sum_{|h|=1} \tilde{H}(u,v)^s F(h(u),h(s)),$$

where $\tilde{H}$ is the analytic extension of H and s is a complex parameter. Here the sum is taken over branches of depth 1. If we define the secant L of the distribution F,

$$L(x,y) = \left| \frac{F(x)-F(y)}{x-y} \right|,$$

then the Dirichlet series can be expressed as

$$\Lambda(F,s) = \sum_h u_h^s = \sum_h |F(h(0)) - F(h(1))|^s = (I - \mathbf{G}_s)^{-1}[L^s](0,1).$$

3.2. **Singularities of the quasi inverse $(I - \mathbf{G}_s)^{-1}$.** Singularities of $(I - \mathbf{G}_s)^{-1}$ are of special interest because these are also singularities of $\Lambda(F,s)$. These singularities arise for values of s where $\mathbf{G}_s$ has an eigenvalue equal to 1. In particular, there is always a pole at $s = 1$ (easy to prove from the previous form of $\Lambda(F,s)$). One can derive from strong properties of the operator $\mathbf{G}_s$ that there are three different cases, called periodic, quasi-periodic and aperiodic, depending on the precise nature of the eigenvalues of $\mathbf{G}_s$ on the line $\Re(s) = 1$. The operator $\mathbf{G}_s$ has also special properties at $s = 1$ and $s = 2$ since the entropy of the source is $h(\mathcal{S}) = -\lambda'(1)$ (the derivative of the dominant eigenvalue at $s = 1$), while the coincidence probability is $c(\mathcal{S}) = \lambda(2)$.

4. Average-case analysis of general tries

4.1. **Analysis of height.** The analysis of height is based on estimates of the individual probabilities $\pi_k(n) = \prod_{|h|=k}(1 + n u_h)e^{-n u_h}$ followed by a Mellin analysis. This leads to the asymptotic expansion

$$D(n) = \frac{2}{|\log c(\mathcal{S})|} \log n + P_F(\log n) + \gamma + A_F + o(1)$$

4.2. **Analysis of size and path length.** The operator $(I - \mathbf{G}_s)^{-1}$ has a simple pole at $s = 1$, and thus gives the main term of the asymptotic expansion. For a general source, a Tauberian theorem can be applied to estimate the contribution of others poles. Finally one has

$$P(n) = \frac{1}{h(\mathcal{S})} n \log n + o(n \log n), \qquad S(n) = \frac{1}{h(\mathcal{S})} n + o(n).$$

5. Some important particular cases

5.1. **Bernoulli sources.** The Bernoulli source considers a finite alphabet $\mathcal{M} = \{1, \ldots, r\}$ with probability of emission $\{p_1, \ldots, p_r\}$ (with $p_1 + \cdots + p_r = 1$). In this case the entropy and the coincidence probability have classical expressions

$$H = -\sum_{i=1}^{r} p_i \log p_i, \qquad c = \sum_{i=1}^{r} p_i^2.$$

5.2. **Continued fraction source.** The continued fraction expansion of numbers can be considered as a dynamical source over the infinite alphabet $\mathcal{M} = \mathbb{N}$. The operator of Ruelle is then called the Ruelle-Mayer operator and is defined by

$$\mathcal{G}_s[f](z) = \sum_{m \geq 1} \frac{1}{(m+z)^s} f\left(\frac{1}{m+z}\right).$$

The entropy of the source is linked to the so-called Lévy's constant which plays a central rôle in the the analysis of the Euclidean algorithm. The coincidence probability is a constant which intervenes in two-dimensional generalizations of the Euclidean algorithm. These constants are

$$\lambda'(1) = \frac{\pi^2}{6 \log 2}, \quad \lambda(2) \sim 0.1994.$$

References

[1] Brigitte Vallée. – Dynamical systems and average-case analysis of general tries. – Les cahiers du GREYC, 1997.

[2] Flajolet (Philippe) and Vallée (Brigitte). – *Continued Fraction Algorithms, Functional Operators, and Structure Constants.* – Research Report n° 2931, Institut National de Recherche en Informatique et en Automatique, July 1996. 33 pages. Invited lecture at the 7th Fibonacci Conference, Graz, July 1996; to appear in *Theoretical Computer Science.*

[3] Hervé Daudé (Philippe Flajolet) and Vallée (Brigitte). – *An Average-case Analysis of the Gaussian Algorithm for Lattice Reduction.* – Research Report n° 2798, Institut National de Recherche en Informatique et en Automatique, February 1996. To appear in *Combinatorics, Probability and Computing*, 1997.

[4] Jacquet (Philippe) and Szpankowski (Wojciech). – Analysis of digital tries whith Markovian dependency. *IEEE Transactions on Information Theory*, vol. 37, n° 5, September 1991, pp. 1470–1475.

[5] Vallée (Brigitte). – Opérateurs de Ruelle-Mayer généralisés et analyse en moyenne des algorithmes d'Euclide et de Gauss. *Acta Arithmetica*, vol. LXXXI, n° 2, 1997, pp. 101–144.