



HAL
open science

Entropy for dynamical sources

Julien Clément, Loïck Lhote, Brigitte Vallée

► **To cite this version:**

Julien Clément, Loïck Lhote, Brigitte Vallée. Entropy for dynamical sources. Proceedings of IWAP 2008, Jul 2009, Compiègne, France. hal-01082045

HAL Id: hal-01082045

<https://hal.science/hal-01082045v1>

Submitted on 12 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Entropy for dynamical sources

Julien Clément, Loïck Lhote, and Brigitte Vallée

Laboratoire GREYC, CNRS, Université de Caen and ENSICAEN
Campus II, Boulevard du Maréchal Juin, BP 5186, 14032 Caen Cedex, France
e-mail: {julien.clement, loick.lhote, brigitte.vallee}@info.unicaen.fr

Abstract. A quite general model of source that comes from dynamical systems theory is introduced. Within this model, basic problems of algorithmic information theory contexts are analysed. The main tool for the analysis is a (generalized) transfer operator, which can be viewed as a “generating” operator for “fundamental” probabilities. Its dominant spectral objects are related to important parameters of the source, such as the entropy rate, and play a central rôle in all the results.

Keywords. Entropy, dynamical systems, transfer operators.

1 Dynamical sources.

In information theory contexts, data items are (infinite) words that are produced by a common mechanism, called a source. Real-life sources are often complex objects. We introduce here a general framework of sources related to dynamical systems theory which can be viewed as a generalization of numeration systems. This model goes beyond the cases of memoryless and Markov chains, it can describe nonmarkovian processes, where the dependency on past history is unbounded, and as such, it attains a high level of generality.

A probabilistic dynamical source is defined by five elements:

- (a) an alphabet Σ , finite or denumerable,
- (b) a topological partition of $\mathcal{I} := [0, 1]$ with disjoint open intervals $(\mathcal{I}_m)_{m \in \Sigma}$,
- (c) an encoding mapping σ which is constant and equal to m on each \mathcal{I}_m ,
- (d) a shift mapping T whose restriction to \mathcal{I}_m is a bijection of class \mathcal{C}^2 from \mathcal{I}_m to $\mathcal{J}_m := T(\mathcal{I}_m)$. The local inverse of $T|_{\mathcal{I}_m}$ is denoted by h_m ,
- (e) a real density f on interval \mathcal{I} , and its associated distribution function F .

The dynamical system $\mathcal{S} := (\mathcal{I}, T)$, defined by the first four elements, associates to a real number $x \in \mathcal{I}$ the word $M(x) := (\sigma x, \sigma T x, \sigma T^2 x, \dots)$ (Fig. 1). Then, the choice of the density f (or the distribution F) on the interval \mathcal{I} induces a probabilistic setting on the set $M(\mathcal{I})$ of source words.

Correlations. Such sources may possess a high degree of correlations, due to the *geometry* of the branches and also to the *shape* of branches.

The *geometry* of the branches is defined by the respective positions of intervals \mathcal{I}_m with respect to intervals \mathcal{J}_ℓ and allows to describe the set of symbols possibly emitted after a given symbol. The system is *complete* if any interval \mathcal{J}_m equals \mathcal{I} . It is *markovian*, if any interval \mathcal{J}_m is an union of intervals \mathcal{I}_ℓ . More generally, the following property is essential:

(P1) [*Topologically mixing*] For any pair of symbols (b, e) , there exists $n_0 \geq 1$ such that, for any $n \geq n_0$, there is a word of length n which begins with symbol b and finishes with symbol e .

The *shape* of the branches h_m , related to various properties of derivatives h'_m , has also a great influence on correlations between symbols:

(P2) [*Contraction*] There exists $\delta < 1$ such that, for any $m \in \Sigma$, for any $x \in \mathcal{I}$, one has : $|h'_m(x)| \leq \delta < 1$.

(P3) [*Strong contraction*] There exist $x_0 \in \mathcal{I}$ and two open disks D_S and D_L of same center x_0 and respective radii $R_S < R_L$, with $]0, 1[\subset D_L$, such that any inverse branch h_m admits an analytic expansion to D_L and strictly maps D_L inside D_S (i.e., $h_m(\overline{D_L}) \subset \overline{D_S}$).

(P4) [*Bounded distortion*] $\exists K, \forall m \in \Sigma, \forall x \in \mathcal{I}, |h''_m(x)| \leq K |h'_m(x)|$.

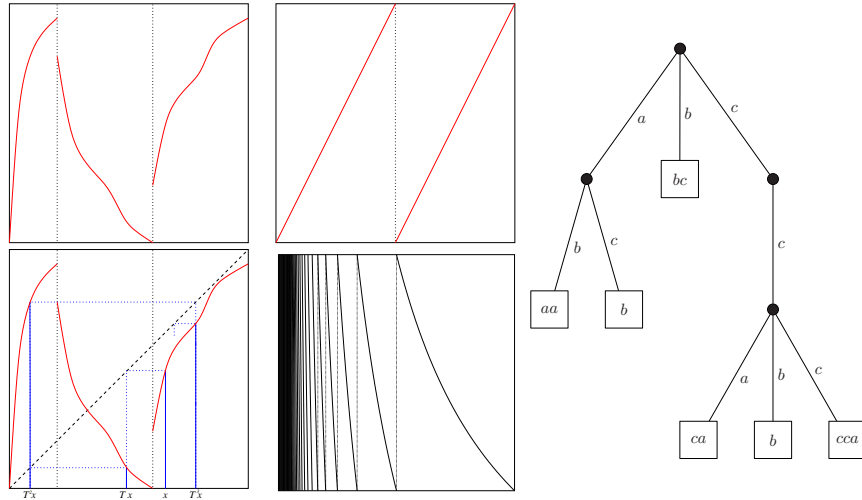


Fig. 1. On the left: A dynamical system, with a partition coded with $\Sigma = \{a, b, c\}$ and a word produced by a dynamical system $M(x) = (c, b, a, c \dots)$. In the middle: two classical sources resp. related to the binary expansion, or the continued fraction expansion. On the right, an example of a trie on alphabet $\Sigma = \{a, b, c\}$ built on the set of words $\{abaa, acb, ccaca, cbbb, cccca\}$.

Classical sources. The classical sources (memoryless sources, or Markov chains) are related to particular dynamical sources; a memoryless source corresponds to a complete system, with affine branches, and a uniform initial density. A Markov chain corresponds to a markovian system, with affine branches and an initial density which is piecewise constant on each \mathcal{I}_m .

Dirichlet series of fundamental measures. We consider in the sequel a complete system. The branches of T^k , and also its inverse branches, are then indexed by Σ^k , and, for any $w = m_1 \dots m_k \in \Sigma^k$, the mapping $h_w := h_{m_1} \circ h_{m_2} \circ \dots \circ h_{m_k}$ is a \mathcal{C}^2 bijection from \mathcal{I} onto $\mathcal{I}_w = h_w(\mathcal{I})$. All the words that begin with the same prefix w correspond to real numbers x that belong to the same interval \mathcal{I}_w . The probability p_w that a word begins with prefix w is called the *fundamental probability* of w .

The Dirichlet series of fundamental probabilities, defined (for complex s) as

$$\Lambda_k(\mathcal{S}_F, s) := \sum_{w \in \Sigma^k} p_w^s, \quad \Lambda(\mathcal{S}_F, s) := \sum_{w \in \Sigma^*} p_w^s \quad (1)$$

will play a fundamental rôle in the source statistics.

Generation of p_w^s . To each prefix w , one associates an operator, denoted by $\mathbf{G}_{s,[w]}$, defined with the secant H_w relative to branch h_w ,

$$H_w(x, y) := \left| \frac{h_w(x) - h_w(y)}{x - y} \right|,$$

which acts on functions L of two variables $(x, y) \in \mathcal{I} \times \mathcal{I}$ as

$$\mathbf{G}_{s,[w]}[L](x, y) := H_w^s(x, y) L(h_w(x), h_w(y)).$$

If now the function L is the secant of the distribution F , the relation

$$p_w^s = |F(h_w(0)) - F(h_w(1))|^s = \mathbf{G}_{s,[w]}[L^s](0, 1)$$

explains why the operator $\mathbf{G}_{s,[w]}$ is called the generating operator of w .

Generating operators. The generating operator relative to a collection \mathcal{L} of words is defined as the sum of all the generating operators relative to the words of \mathcal{L} . For two prefixes w, w' , the relation $p_{w.w'} = p_w p_{w'}$ is no longer true when the source has some memory, and is replaced by the following composition property

$$\mathbf{G}_{s,[w.w']} = \mathbf{G}_{s,[w']} \circ \mathbf{G}_{s,[w]}, \quad (2)$$

so that unions and cartesian products of collections of words translate into sums and compositions of the generating operators. The generating operator

$$\mathbf{G}_s := \sum_{m \in \Sigma} \mathbf{G}_{s,[m]}$$

related to the alphabet Σ plays a fundamental rôle here. By Eq. (2), the k -th iterate \mathbf{G}_s^k generates all the words of length k , and the quasi-inverse $(I - \mathbf{G}_s)^{-1}$ generates all the finite words. Then, alternative expressions for the Dirichlet series defined in Eq. (1) hold:

$$\Lambda_k(\mathcal{S}_F, s) = \mathbf{G}_s^k[L^s](0, 1), \quad \Lambda(\mathcal{S}_F, s) = (I - \mathbf{G}_s)^{-1}[L^s](0, 1). \quad (3)$$

Asymptotic properties of powers \mathbf{G}_s^k of the transfer operator will be needed and they are closely related to *dominant spectral properties* of operator \mathbf{G}_s . In the same vein, we are interested by the location of poles of the quasi-inverse $(I - \mathbf{G}_s)^{-1}$, that is, values s for which 1 is a spectral value of \mathbf{G}_s . Under quite general hypotheses on the dynamical system, the following holds:

Decomposable sources. *A dynamical system \mathcal{S} is said to be decomposable with respect to a Banach space \mathcal{L} iff the transfer operator \mathbf{G}_s , when acting on \mathcal{L} , possesses for real $s > s_0$ (with $s_0 < 1$) a unique dominant eigenvalue $\lambda(s)$ (real positive) separated from the remainder of the spectrum by a spectral gap. A source \mathcal{S}_F formed with a decomposable system \mathcal{S} (with respect to \mathcal{L}) and a function F whose secant belongs to \mathcal{L} is said to be decomposable.*

Properties of decomposable sources. *For such sources, one has:*

(a) *When s is near the real axis, $\Lambda_k(\mathcal{S}_F, s)$ asymptotically behaves as the k -th power of $\lambda(s)$. There exists $\mu < 1$, such that, when s is near a real s_0*

$$\Lambda_k(\mathcal{S}_F, s) = a(s) \cdot \lambda(s)^k \cdot [1 + O(\mu^k)], \quad \text{with a uniform } O.$$

(b) The function $s \mapsto \Lambda(\mathcal{S}_F, s)$ is analytic in the plane $\Re(s) > 1$, and it has a simple pole at $s = 1$, with a residue equal to $-\lambda'(1)$.

Sufficient conditions for decomposability. A markovian dynamical system on a finite alphabet satisfying (P1) and (P2) is decomposable, with respect to the space $\mathcal{C}_p^1(\mathcal{I})$ of piecewise functions of class \mathcal{C}^1 . A complete dynamical system on a (possibly infinite) alphabet satisfying (P2), (P4) is decomposable, with respect to the space $\mathcal{C}^1(\mathcal{I})$ of functions of class \mathcal{C}^1 . A complete dynamical system on a (possibly infinite) alphabet satisfying (P3), (P4) is decomposable, with respect to the space of analytic functions on D_S .

There are two kinds of “exceptional” sources, both related to exceptional properties of the pressure $\Lambda(s) := \log \lambda(s)$. The first one arises when the pressure $\Lambda(s)$ is periodic and the second when the pressure $\Lambda(s)$ is affine. They intervene as exceptions respectively for Result C and Result B. However we prove that an exceptional source is conjugated either to a memoryless source or to a Markov chain. Hence a “complex” source cannot be exceptional.

2 Main statistics for decomposable sources.

Entropy. The entropy rate $h(\mathcal{S}_F)$ of the source \mathcal{S}_F is defined as the limit, if it exists, of a quantity that involves probabilities p_w ,

$$h(\mathcal{S}_F) := \lim_{k \rightarrow \infty} \frac{-1}{k} \sum_{w \in \Sigma^k} p_w \log p_w = \lim_{k \rightarrow \infty} \frac{-1}{k} \frac{d}{ds} \Lambda_k(\mathcal{S}_F, s)|_{s=1}.$$

Result A. The entropy $h(\mathcal{S}_F)$ of a decomposable source \mathcal{S}_F equals $-\lambda'(1)$. It only depends on the mechanism \mathcal{S} .

Distribution of prefixes with fixed length k . The distribution of the fundamental probabilities of the words of Σ^k is studied via the random variable ℓ_k defined as a step function, constant on any fundamental interval \mathcal{I}_w of depth k and equal to p_w . We use the moment generating function $\mathbb{E}[\exp(s \log \ell_k)] = \mathbb{E}[\ell_k^s]$ which is closely related to $\Lambda_k(\mathcal{S}_F, s)$:

$$\mathbb{E}[\ell_k^s] = \sum_{w \in \Sigma^k} p_w^s p_w = \sum_{w \in \Sigma^k} p_w^{1+s} = \Lambda_k(\mathcal{S}_F, 1 + s).$$

Then, when \mathcal{S}_F is decomposable, the moment generating function behaves nearly like the k -th power of function $\lambda(1 + s)$. The central limit theorem of probability theory asserts that exact large powers induce Gaussian laws in the asymptotic limit. Here, we shall use an extension of the central limit theorem to “quasi-powers” due to Hwang (1998), valid provided that $\Lambda''(1)$ is not zero. We know that this is the case for the decomposable sources that do not behave like an unbiased memoryless source [see Vallée (2001)].

Result B. The distribution of probabilities of words of Σ^k follows asymptotically a log-normal law, for all not affine sources. This proves a strong “equipartition property”, in the flavour of the Shannon–MacMillan–Breiman Theorem. The dominant behaviours of the mean and the variance involve the

derivatives of $\Lambda(s) := \log \lambda(s)$ at $s = 1$: the mean is asymptotic to $k\Lambda'(1) = k\lambda'(1)$ and the variance is asymptotic to $k\Lambda''(1) = k[\lambda''(1) - \lambda'(1)^2]$.

Average shape of a trie. In pattern matching problems, an essential data structure is a digital tree, called trie, which plays the rôle of a dictionary. The internal nodes are used for directing the search, and the leaves contain the words of the dictionary. There are as many leaves as words in the dictionary (see Fig. 1). Trie analysis aims at describing the average shape of a trie (size S , external path length P); it is extensively done in the case of classical sources (see Szpankowski (2004)). In our context, a trie is built on a set Y of n words or equivalently on a set X of reals $\{x_1, x_2, \dots, x_n\}$ for which $Y = M(X)$. Such a trie is completely defined by the internal nodes which *actually* exist. These nodes are labelled by prefixes w for which the cardinality N_w of $\mathcal{I}_w \cap X$ is at least two (which means that at least two words share a common prefix). The probability that node n_w exists is just $\mathbb{P}[N_w \geq 2]$ while the contribution of node n_w to the external path length is $\mathbb{E}[N_w \mid N_w \geq 2]$. If we wish the contributions of set X to two disjoint intervals \mathcal{I}_w and $\mathcal{I}_{w'}$ be independent, we are led to first work in a Poisson model: first, one draws the cardinality N of the set X with respect to a Poisson law of parameter z , then one independently draws the N reals with respect to density f . In this case, the random variable N_w itself follows a Poisson law of parameter $p_w z$, and the two expectations admit the following expressions

$$\mathbb{E}[P_{(z)}] = \sum_{w \in \Sigma^*} p_w z (1 - e^{-p_w z}), \quad \mathbb{E}[S_{(z)}] = \sum_{w \in \Sigma^*} 1 - e^{-p_w z} (1 + p_w z),$$

that are harmonic sums of the form $A(z) = \sum_w \lambda_w g(\mu_w z)$. The convenient tool is the Mellin transform which transforms an harmonic sum into a product and gives access to the asymptotic behaviour of $\mathbb{E}[P_{(z)}]$ and $\mathbb{E}[S_{(z)}]$ through the Dirichlet series $\Lambda(\mathcal{S}_F, s)$, and particularly its singular behaviour near $\Re s = 1$, and its residue $h(\mathcal{S})$. Then, Depoissonization techniques allow to return to the Bernoulli model (where the number n of words is fixed). One obtains:

Result C. *The shape parameters of a trie [size, external length path] built on n words [independently drawn from the source] admit as asymptotic mean (for $n \rightarrow \infty$) the following quantities which involve the entropy*

$$\mathbb{E}[S_n] \sim \frac{n}{h(\mathcal{S})}, \quad \mathbb{E}[P_n] \sim \frac{n \log n}{h(\mathcal{S})}.$$

For sources that behaves like meoryless sources, the main asymptotic term of $\mathbb{E}[S_n]$ involves a periodic function of $\log n$ of small amplitude.

Computing the spectral constants. All the previous results involve the dominant eigenvalue $\lambda(s)$. Except for simpler sources, this function does not admit a closed form, and it is important to obtain proven numerical values for the first two derivatives of the pressure function at $s = 1$.

For sources with strong contraction (P3), the transfer operator acts on holomorphic functions, and can be viewed as an infinite matrix \mathbf{M}_s , whose coefficient of index (i, j) is the coefficient of $(z - x_0)^i$ in $\mathbf{G}_s[(Z - x_0)^j, (T - x_0)^j](z, z)$.

In particular, the truncated matrix $\mathbf{M}_{(n),s}$ where the indices i, j are less than n , is an approximation of \mathbf{G}_s (in a precise sense), and its dominant eigenvalue $\lambda_n(s)$ provides an approximate for $\lambda(s)$.

Result D. *Let $\lambda(s)$ be the dominant eigenvalue of a strongly contracting decomposable source. Then, there exist an integer n_0 , and a constant K (which depend on s), such that for all $n \geq n_0$, the matrix $\mathbf{M}_{(n),s}$ admits a unique simple dominant eigenvalue $\lambda_{(n)}(s)$ that satisfies*

$$|\lambda(s) - \lambda_{(n)}(s)| \leq K \left(\frac{R_S}{R_L} \right)^n .$$

For instance, in the case of the Euclidean dynamical source (for which the entropy $h(\mathcal{S})$ equals $\pi^2/(6 \log 2) \sim 2.37$), this result allows to compute with a guaranteed precision the value $\lambda''(1) \sim 9.08037$.

Bibliographic notes. Dynamical sources have been introduced by Vallée (2001), who proved the first two results of Section 2. Some hypotheses for proving decomposability are made precise in Chazal, Maume, Vallée (2004), and a general decomposability criterion is proven in Chazal and Maume (2004). Tries analysis is performed by Clément, Flajolet and Vallée (2001). The method for approximating $\lambda(s)$ is due to Daudé, Flajolet and Vallée (1997). This method was validated later by Lhote (2004) who proved Result D. The continued fraction expansion motivates the introduction of dynamical sources. For a specific treatment of this case, see Flajolet and Vallée (2000), where a relation between the shape of tries and the Riemann hypothesis is exhibited...

References

- Chazal, F., Maume-Deschamps, V., Vallée, B. (2004) Erratum to “Dynamical sources in information theory: fundamental intervals and word prefixes”, *Algorithmica* 38, 591–596.
- Chazal, F., Maume-Deschamps, V. (2004) Statistical of Markov dynamical sources: applications to information theory, *DMTCS*, 6, 2004, 283–314.
- Clément, J., Flajolet P., Vallée, B. (2001) Dynamical sources in Information Theory: Analysis of general tries, *Algorithmica*, vol 29 (1/2), 307–369.
- Daudé, H., Flajolet, P. and Vallée, B. (1997) An Average-Case Analysis of the Gaussian Algorithm for Lattice Reduction. *Combinatorics, Probability and Computing* (1997), 6: 397-433 Cambridge University Press
- Flajolet, P., Gourdon, X., and Dumas, P. (1995) Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science* 144, 1–2 (June 1995), 3–58.
- Flajolet, P., Vallée, B. (2000) Continued Fractions, Comparison Algorithms, and Fine Structure Constants, *Proceedings of Canadian Mathematical Society*, Vol 27, 53–82.
- Hwang, H.-K.(1998) On convergence rates in the central limit theorems for combinatorial structures, *European Journal of Combinatorics* 19 (1998) 329–343.
- Lhote, L. (2004). Computation of a Class of Continued Fraction Constants, *Proceedings of ANALCO 2004*, 199–210
- Szpankowski, W. (2001) *Average-case analysis on sequences* John Wiley and sons, New-York.
- Vallée, B. (2001) Dynamical sources in Information Theory: Fundamental Intervals and Word prefixes, *Algorithmica*, vol 29 (1/2), 262–306.