INFORMATION THEORY: SOURCES, DIRICHLET SERIES, REALISTIC ANALYSIS OF ALGORITHMS

Brigitte Vallée
GREYC Laboratory
(CNRS and University of Caen, France)

Talk based on joint works with
Eda Cesaratto, Julien Clément, Jim Fill,
Philippe Flajolet, and Mathieu Roux

Séminaire de Combinatoire énumérative et analytique,
IHP, Paris, 3 février 2011

Description of a framework which

– unifies the analyses for text algorithms and searching/sorting algorithms

– provides a general model for sources
– shows the importance of the Dirichlet generating functions
– explains the importance of tameness for sources

– defines a natural subclass of sources, the dynamical sources
– provides sufficient conditions for tameness of dynamical sources

– provides probabilistic analyses for algorithms built on tame sources.

Plan of the talk.

– General motivations

– Models for a source

– The Dirichlet generating function of the source

– Conclusion and possible extensions.

Plan of the talk.

- General motivations
- Models for a source
- The Dirichlet generating function of the source
- Conclusion and possible extensions.

The classical framework for analysis of algorithms

in two main algorithmic domains:

Text algorithms – Sorting or Searching algorithms.

– In text algorithms, algorithms deal with words

What is a word ?

.... a sequence of symbols from the same alphabet

– In sorting or searching algorithms, algorithms deal with keys.

The set of keys must be ordered.

What is a key?

.... a sequence of symbols from the same alphabet

Key or word? the same object... but,

– for comparing two words, importance of the structure of words

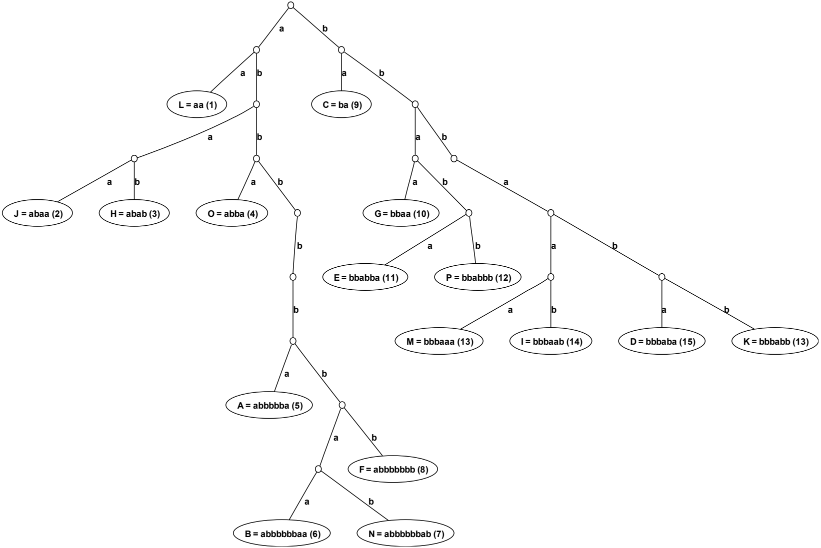
– for comparing two keys, transparence of the structure of keys

only their relative order plays a role.

Text algorithms and dictionaries : The trie structure

An example : A trie built on a set of 16 words.

A = abbbbbaaabab B = abbbbbbaabaa C = baabbbabbbba D = bbbababbbaab E = bbabbaababbb
F = abbbbbbbbabb G = bbaabbabbaba H = ababbbabbbab I = bbbaabbbbbbb J = abaabbbbaabb
K = bbbabbbbbbaa L = aaaabbabaaba M = bbbaaabbbbbb N = abbbbbbabbaa O = abbababababbbb P = bbabbbaaaabb



Probabilistic study of the `Trie` structure.

Main parameter on a node n_w labelled with prefix w:

$N_w :=$ the number of words which begin with prefix w.

$N_w :=$ the number of words which go through the node n_w

The size, and the path length of a trie equal

$$R = \sum_{w \in \Sigma^\star} \mathbf{1}_{[N_w \geq 2]} \qquad\qquad T = \sum_{w \in \Sigma^\star} \mathbf{1}_{[N_w \geq 2]} \ N_w;$$

Role of $p_w :=$ the probability that a word begins with prefix w.

Classical analyses of the main algorithms for searching or sorting

The unit cost is the key–comparison.

The behaviour of the algorithm (wrt to key–comparisons) only depends on the relative order between the keys.
It is sufficient to restrict to the case when $\Omega = [1..n]$.
The input set is then $\mathfrak{S}_n$, with uniform probability.

Then, the analysis of all these algorithms is very well known, with respect to the number of key–comparisons performed in the worst-case, or in the average case.

A more realistic framework for sorting or searching.

Keys are viewed as words. The domain Ω of keys is a subset of $\Sigma^{\mathbb{N}}$,

$$\Sigma^{\mathbb{N}} = \{\text{the infinite words on some ordered alphabet } \Sigma\}.$$

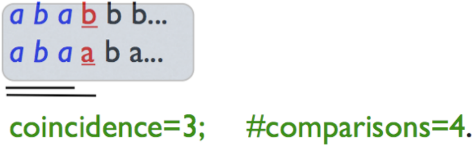
The words are compared [wrt the lexicographic order].
The realistic unit cost is now the symbol–comparison.

The realistic cost of the comparison between two words A and B,

$$A = a_1\, a_2\, a_3 \ldots a_i \ldots \qquad \text{and} \qquad B = b_1\, b_2\, b_3 \ldots b_i \ldots$$

equals $k+1$, where k is the length of their largest common prefix

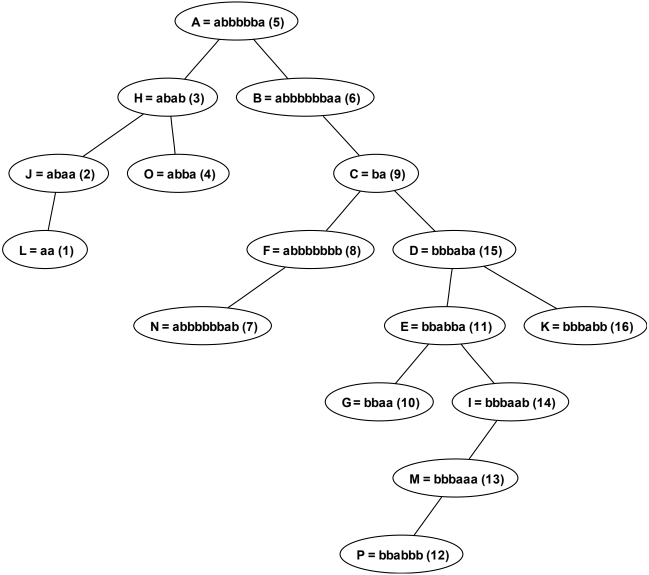
$$k := \max\{i;\quad \forall j \quad i;\quad a_j = b_j\} = \text{the coincidence}$$



Now, sorting or searching algorithms are viewed as text algorithms.

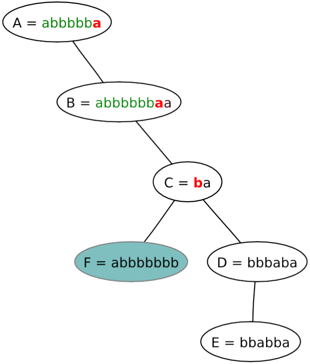
An example : The BST (binary search tree) built on a sequence of words

A = abbbbbaaabab B = abbbbbbaabaa C = baabbbabbbba D = bbbababbbaab E = bbabbaababbb
F = abbbbbbbbabb G = bbaabbabbaba H = abababbabbbab I = bbbaabbbbbbb J = abaabbbbaabb
K = bbbabbbbbbaa L = aaaabbabaaba M = bbbaaabbbbbb N = abbbbbbabbaa O = abbababbabbbb P = bbabbbaaaabb



An example : The cost of the insertion of the key F into the BST

F =abbbbbbb

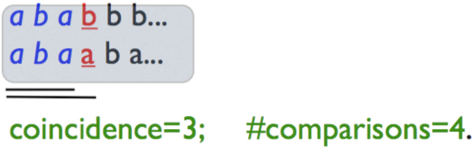


Number of symbol comparisons needed ?

$=$ 7 for comparing to A
$+$ 8 for comparing to B
$+$ 1 for comparing to C

Total $=$ 16

The realistic cost of the comparison between two words A and B
is related to their coincidence $c(A; B)$



The coincidence $c(A; B)$ satisfies $c(A; B) \quad k$
if and only if A and B begin with the same prefix of length k

Importance of $p_w :=$ the probability that a word begins with prefix w.

Now, we work inside an unifying framework
where searching and sorting algorithms are viewed as text algorithms.

In this context, the probabilistic behaviour of algorithms heavily depends on the mechanism which produces words.

A source:= a mechanism which produces symbols from alphabet Σ,
one for each time unit.

When (discrete) time evolves, a source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

For $w \in \Sigma^\star$, $p_w :=$ probability that a word begins with the prefix w.
The set $\{p_w ;\ w \in \Sigma^\star\}$ defines the source S.

Fundamental role of the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s; \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s$$

Remark: $\Lambda_k(1) = 1$ for any k, $\quad \Lambda(1) = \mathit{1}$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

For instance, the entropy $h_{\mathcal{S}}$, the coincidence $c_{\mathcal{S}}$

$$h(S) := \lim_{k \to \infty} \frac{1}{k} \sum_{w \in \Sigma^k} p_w \log p_w = \quad \lim_{k \to \infty} \Lambda_k'(1)$$

$$\Pr[c_{\mathcal{S}} \quad k+1] = \sum_{w \in \Sigma^k} p_w^2 = \Lambda_k(2)$$

– they intervene in the probabilistic analyses of text algorithms and (also) sorting and searching algorithms.

Three main steps for the analysis
of the mean number S_n of symbol comparisons

(1) First step (algebraic).

The Poisson model P_Z deals with a variable number N of words:
N is a random variable which follows a Poisson law of parameter Z.

We first obtain nice expressions for the mean number $\widetilde{S}(Z)$ ....

(2) Second step (algebraic).

It is possible to return to the model where the number n of words is fixed.
We obtain a nice exact formula for S_n ....

from which it is not easy to obtain the asymptotics...

(3) Third step (analytic).

Then, the Rice formula provides the asymptotics of S_n $(n \to \infty)$,

as soon as the source is "tame"

After the second step, an exact formula for S_n ....

$$S_n = \sum_{k=2}^{n} (\quad 1)^k \binom{n}{k} \$(k)$$

...which involves the series $\$$ at integer values k.

For the mean path length (`Trie` or `BST`),

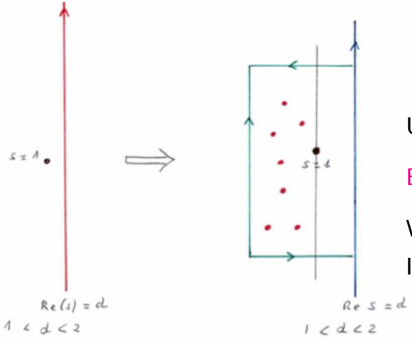
$\$(s)$ is closely related to the Dirichlet series of the probabilities,

$$\$_T(s) = s\Lambda(s) \qquad \$_B(s) = 2\frac{\Lambda(s)}{s(s \quad 1)} \qquad \text{where} \quad \Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$$

Asymptotic analysis.

The residue formula transforms the sum into an integral with $1 < d < 2$.

$$S_n = \sum_{k=2}^{n} (\quad 1)^k \binom{n}{k} \$(k) = \frac{1}{2i} \int_{d-i\infty}^{d+i\infty} \$(s) \frac{n!\,(\quad 1)^{n+1}}{s(s \quad 1) ::: (s \quad n)} ds;$$



We shift the integral on the left,
Usually, the first singularities occur at $\Re s = 1$.

Behaviour of $\varpi(s)$ [or $\Lambda(s)$] near $\Re s = 1$?

Where are the red singularities closest to $\Re s = 1$?
Is $\Lambda(s)$ of polynomial growth on the green contour?

Importance of the existence of a region R
– which contains only $s = 1$ as a pole – where $\Lambda(s)$ is of polynomial growth.
Tameness of the source

Case of Trie(n) [Clément, Flajolet, V. 2001]

Theorem 1. For any tame source,
the mean path length T_n of a trie built on n words independently drawn from the source satisfies

$$T_n \qquad \frac{1}{h_{\mathcal{S}}}\, n \log n:$$

and involves the entropy $h_{\mathcal{S}}$ of the source S, defined as

$$h_{\mathcal{S}} := \lim_{k\to\infty} \left[\frac{1}{k} \sum_{w\in\Sigma^k} p_w \log p_w \right];$$

where p_w is the probability that a word begins with prefix w.

Case of QuickSort(n) or BST(n) [Clément, Fill, Flajolet, V. 2009]

Theorem 2. For any tame source,
the mean number S_n of symbol comparisons used by `QuickSort`(n)
(or the mean number of symbols comparisons used to built the BST) on
n words of the source satisfies

$$B_n \quad \frac{1}{h_{\mathcal{S}}}\, n \log^2 n:$$

and involves the entropy $h_{\mathcal{S}}$ of the source S, defined as

$$h_{\mathcal{S}} := \lim_{k\to\infty}\left[\frac{-1}{k}\sum_{w\in\Sigma^k} p_w \log p_w\right];$$

where p_w is the probability that a word begins with prefix w.

Compared to $K_n \quad 2n \log n$, there is an extra factor equal to $1{=}(2h_{\mathcal{S}}) \log n$

Compared to $T_n \quad (1{=}h_{\mathcal{S}})\, n \log n;$ there is an extra factor of $\log n$.

Plan of the talk.

- General motivations
- Models for a source
- The Dirichlet generating function of the source
- Conclusion and possible extensions.

The parametrization of a general source

A general source S produces infinite words

on an ordered alphabet $\Sigma := fa_1;:::;a_r g$.

The set of infinite words produced by S is $L(S) \quad \Sigma^{\mathbb{N}}$.

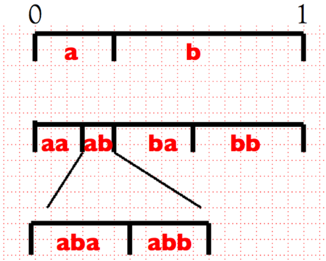
For w 2 $\Sigma^\star$, $p_w :=$ probability that a word begins with the prefix w.

The set $fp_w; \; w$ 2 $\Sigma^\star g$ defines the source S.

We assume: $\quad {}_k := \sup fp_w; \; w$ 2 $\Sigma^k g$! 0 for k ! 1

For each length k, we consider the probabilities p_w with w 2 Σ^k

sorted with respect to the lexicographic order on Σ^k.



For each $w \in \Sigma^k$ $\quad p_w^{(-)} := \sum_{\substack{\alpha\in\Sigma^k,\\ \alpha<w}} p_\alpha$

For any $X \in \mathcal{L}(\mathcal{S})$, let

If $X := \lim_{k\to\infty} w_k, \quad F(X) := \lim_{k\to\infty} p_{w_k}^{(-)}$

Then $F(X) := \Pr[Y < X]$

F is the distribution function on $\mathcal{L}(\mathcal{S})$.

The function $F : L(S) \to [0,1]$ is continuous

and strictly increasing outside an exceptional (denumerable set)

Outside this exceptional set, each infinite word X of $L(S)$ is written as

$X = M(u)$ with $M : [0,1] \to L(S)$.

The real u is the parameter of the word X.

The map M provides a parametrization of the source S.

Via the mapping M,

[Drawing of words X in S]   [Uniform drawing of parameters u in $[0,1]$]

For any finite prefix $w \in \Sigma^\star$,

the set $\{u;\ \ M(u)\ \text{ begins with }\ w\}$ is an interval

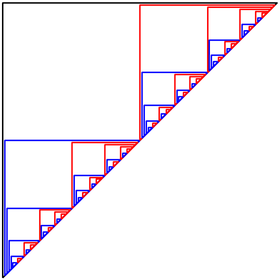
This is the fundamental interval of w. Its length equals p_w.

For any finite prefix w 2 $\Sigma^\star$,

the set *fu; M(u) begins with wg* is an interval

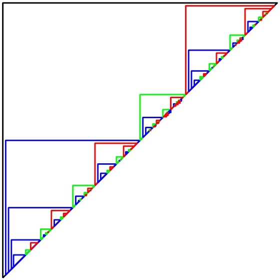
This is the fundamental interval of w. Its length equals p_w.

Instances of fundamental intervals for two memoryless sources.



Memoryless source on $\{a, b\}$

$p_a = 1/2$, $p_b = 1/2$



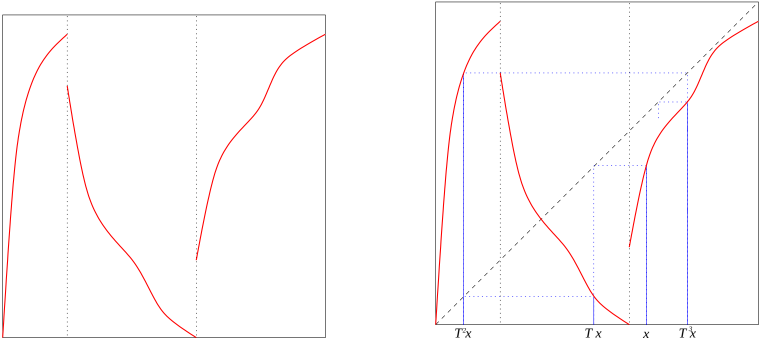
Memoryless source on $\{a, b, c\}$

$p_a = 1/2$, $p_b = 1/6$, $p_c = 1/3$

In the memoryless case, there are regular splittings.

A general class of "natural" sources: dynamical sources associated to a "natural" parametrization

With a shift map $T : I \to I$ and an encoding map $\sigma : I \to \Sigma$,
the emitted word is $M(u) = (\sigma u, \sigma Tu, \sigma T^2 u, \dots, \sigma T^k u, \dots)$
namely, the encoded trajectory of u



A dynamical system, with $\Sigma = \{a, b, c\}$ and a word $M(u) = (c, b, a, c \dots)$.

A dynamical source $=$ a source built with a dynamical system

A dynamical system $(I; S)$ is defined by four elements:

– a finite alphabet Σ,

– a topological partition of $I :=]0; 1[$ with open intervals $I_{m, m \in \Sigma}$,

– an encoding mapping equal to m on each I_m,

– a shift mapping T

s.t. $T|_{\mathcal{I}_m}$ is a bijection of class C^2 from I_m to $J_m := T(I_m)$.

This gives rise to a source: on an input u of I, it outputs the word

$$M(u) := (u; Tu; T^2u; \ldots).$$

When an initial density –and an initial distribution F– is chosen on I,

this induces (via M) a probabilistic model on Σ^∞

$=$ a dynamical source S_F.

Strong relations between the geometry of the system and the probabilistic properties of the source.

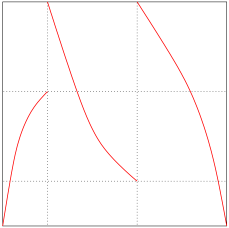
Correlations between symbols are mainly due to two geometric characteristics: the position of the branches and the shape of the branches.

– the position of the branches: the position of $T(I_m)$ wrt I_ℓ; it describes the set $s(m)$ of possible successors of the symbol m.

Particular cases: – Complete systems $T(I_m) = I$

– Markovian systems $T(I_m) = \text{union of some } I_\ell$

give rise to a finite characterization of $s(m)$.



A markovian system

Generally speaking, importance of topological mixing: "There is a word of length n which begins with b and ends with e".

Strong relations between the geometry of the system and the probabilistic properties of the source.

Correlations between symbols are also due to the shape of the branches.

– the shape of the branches, is described by their derivatives;
it explains how the distribution evolves.

Less correlated systems correspond to systems with affine branches.

Generally speaking, importance of expansiveness:

the derivative T' satisfies $\forall u \in I \quad |T'(u)| \quad > 1.$

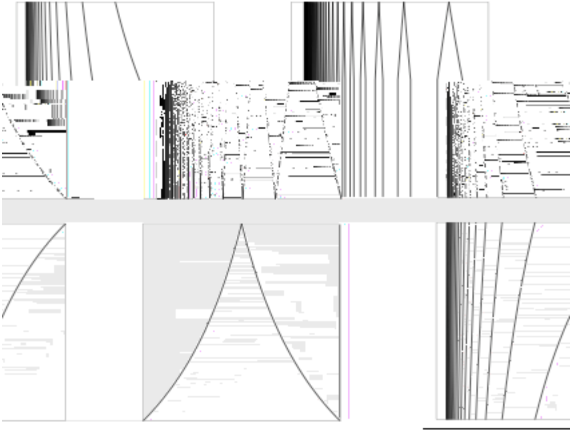
When true, this implies a chaotic behaviour for trajectories.

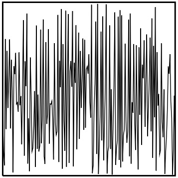
When this condition is violated at only one indifferent point,

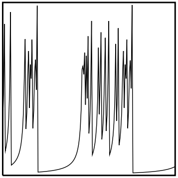
$$[T(u) = u; |T'(u)| = 1]$$

this leads to intermittency phenomena.

Four Euclidean dynamical sources, Two different behaviours







Particular cases: simple sources and affine branches

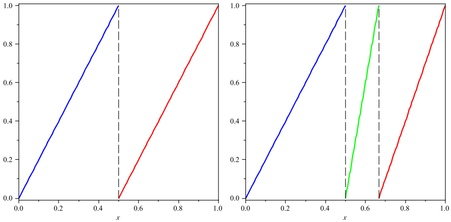
A memoryless source

$:=$ a complete system with affine branches and uniform initial density

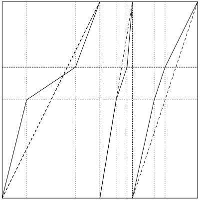
A Markov chain

$:=$ a Markovian system with affine branches,

with an initial density which is constant on each I_m.



Two memoryless sources



a Markov chain.

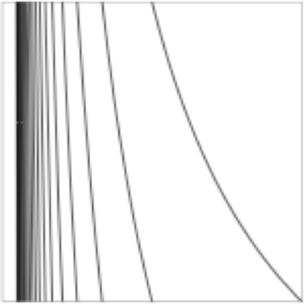
General case of interest.

A complete –or a Markovian– system

– with a possible infinite denumerable alphabet

– topologically mixing – and expansive.

Main instance: the Euclidean source defined with $T(x) := \frac{1}{x} \quad \lfloor \frac{1}{x} \rfloor$



Plan of the talk.

– General motivations

– Models for a source

– The Dirichlet generating function of the source

– Conclusion and possible extensions.

A main analytical object related to any source:

the Dirichlet series of probabilities, $\quad \Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

Memoryless sources, with probabilities (p_i)

$$\Lambda(s) = \frac{1}{1 \quad (s)} \qquad \text{with} \qquad (s) = \sum_{i=1}^{r} p_i^s$$

Markov chains, defined by – the vector R of initial probabilities (r_i)

– and the transition matrix $P := (p_{i,j})$

$$\Lambda(s) =^t \mathbf{1}(I \quad P(s))^{-1} R(s) \qquad \text{with} \quad P(s) = (p_{i,j}^s); \quad R(s) = (r_i^s):$$

A general dynamical source

$\Lambda(s)$ closely related to $(I \quad \mathbb{H}_s)^{-1}$

where $\mathbb{H}_s$ is the (secant) transfer operator of the dynamical system.

Probabilistic analysis of text algorithms
when the text is generated by a dynamical source.

A dynamical source

Geometric properties of the branches + Spectral properties of the transfer operator + Analytical properties of the Quasi-Inverse of the transfer operator

+

Analytical properties of the Dirichlet generating function $\Lambda(s)$

+

Probabilistic analysis of text algorithms

Fundamental probabilities for a dynamical source (complete case).

The fundamental probability $p_w =$ the probability that $M(u)$ begins with w

For any $w \in \Sigma^\star$, $\qquad \mathcal{I}_w := \{u; \quad M(u)$ begins with the prefix $w\}$

Here, in the complete dynamical case, for any k and any $w \in \Sigma^k$:

– the restriction $T^k|_{\mathcal{I}_w}$ is a C^2 bijection of $\mathcal{I}_w$ onto $\mathcal{I}$

– Its inverse mapping h_w is a C^2 bijection from $\mathcal{I}$ to $\mathcal{I}_w = [h_w(0), h_w(1)]$.

With a change of variables, this provides an alternative expression for p_w:

$$p_w = \int_{\mathcal{I}_w} f(t)dt = \int_{\mathcal{I}} |h'_w(x)| \; f \circ h_w(x) \; dx.$$

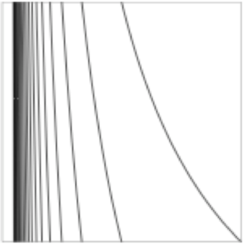
which involves the operator $\mathbf{H}_{[w]}$

$$\mathbf{H}_{[w]}[f](x) := |h'_w(x)| \; f \circ h_w(x)$$

via the relation

$$p_w = \int_{\mathcal{I}} \mathbf{H}_{[w]}[f](t)dt$$

The density transformer and the transfer operator



The operator $\mathbf{H} := \sum_{a\in\Sigma} \mathbf{H}_{[a]}$

with $\mathbf{H}_{[a]}[f](x) = |h_a'(x)| \cdot f \circ h_a(x)$

is the density transformer of the dynamical system.

It describes the evolution of the density after one iteration.

For a density f on $[0,1]$,

$\mathbf{H}[f]$ is the density on $[0,1]$ after one iteration.

Transfer operator (Ruelle)

$$\mathbf{H}_s := \sum_{a\in\Sigma} \mathbf{H}_{s,[a]} \qquad \text{with} \quad \mathbf{H}_{s,[a]}[f](x) = |h_a'(x)|^s\, f \circ h_a(x).$$

The k-th iterate satisfies:

$$\mathbf{H}_s^k = \sum_{w\in\Sigma^k} \mathbf{H}_{s,[w]} \qquad \text{with} \quad \mathbf{H}_{s,[w]}[f](x) = |h_w'(x)|^s\, f \circ h_w(x)$$

Generation of p_w^s.

For an inverse branch h of any depth, and an initial distribution F, consider the "secants" $H; L$ of $h; F$,

$$H(x; y) := \left|\frac{h(x) \quad h(y)}{x \quad y}\right|; \qquad L(x; y) = \left|\frac{F(x) \quad F(y)}{x \quad y}\right|:$$

and the component "secant" operator $\mathbb{H}_{s,[w]}$, defined as

$$\mathbb{H}_{s,[w]}[L](x; y) := H_w^s(x; y) \quad L(h_w(x); h_w(y)):$$

This operator generates p_w^s:

$$p_w^s = jF(h_w(1)) \quad F(h_w(0)j^s = \left|\frac{h_w(1) \quad h_w(0)}{1 \quad 0}\right|^s \left|\frac{F(h_w(1)) \quad F(h_w(0))}{h_w(1) \quad h_w(0)}\right|^s$$

$$= \mathbb{H}_{s,[w]}[L^s](1; 0)$$

"On the diagonal" $\qquad \mathbb{H}_{s,[w]}[L](x; x) = \mathbf{H}_{s,[w]}[f](x)$

The diagonal of the secant is the tangent

The secant operator is an extension of the plain (tangent) operator

For $w, w' \in \Sigma^\star$, the multiplicative relation $\mathbb{H}_{[s, w\cdot w']} = \mathbb{H}_{s,[w']} \circ \mathbb{H}_{s,[w]}$ generalizes the equality $p^s_{w\cdot w'} = p^s_w \cdot p^s_{w'}$,

no longer true when the source has memory.

The Dirichlet series of fundamental probabilities

$$\Lambda_k(s) := \sum_{w\in\Sigma^k} p^s_w, \qquad \Lambda(s) := \sum_{w\in\Sigma^\star} p^s_w$$

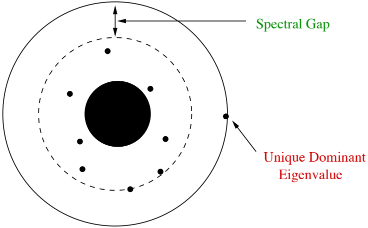
are "generated" by the secant transfer operator $\mathbb{H}_s$ [V. 2000]

$$\Lambda_k(s) = \mathbb{H}^k_s[L^s](0,1), \quad \Lambda(s) = (I - \mathbb{H}_s)^{-1}[L^s](0,1).$$

Singularities of $s \mapsto \Lambda(s)$ are essential in the analysis.
Singularities of $(I - \mathbb{H}_s)^{-1}$ are related to spectral properties of $\mathbb{H}_s$.
For $s = 1$, $\mathbb{H}_1$ is an extension of $\mathbf{H}$ and has an eigenvalue equal to 1.

A source is decomposable if there exists a Banach space $\mathcal{L}$ for which $\mathbb{H}_s : \mathcal{L} \longrightarrow \mathcal{L}$ possesses for real $s > s_0$ (with $s_0 < 1$)

- a unique dominant eigenvalue $\lambda(s)$
- and a spectral gap.



Spectrum of $\mathbb{H}_s$ for a decomposable source

Sufficient conditions for decomposability:

A markovian and expansive system on a finite alphabet is decomposable.

In this case, the function s 7! $\Lambda(s)$

- is analytic in the plane $<(s) > 1$,
- and it has a simple pole at $s = 1$.

$$\Lambda(s) \quad {}_{s\to 1} \quad \frac{1}{1 \quad\quad (s)} \quad {}_{s\to 1} \quad \frac{1}{'(1)(s \quad 1)}$$

The entropy of the source h_S is related to (s), namely $h_S = \quad '(1)$

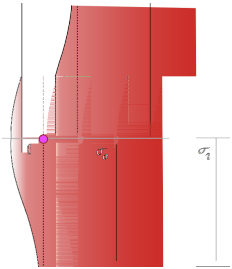
And on the left of the vertical line $<s = 1$?

It is important for the analysis to deal with a region R where $\Lambda(s)$ is tame

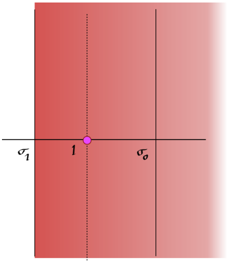
– it is analytic

– it is of polynomial growth when $=s$! 1

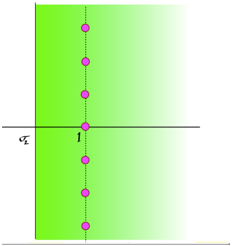
Different possible regions R on the left of $<s = 1$ where $\Lambda(s)$ is tame.



Situation 1
Hyperbolic region

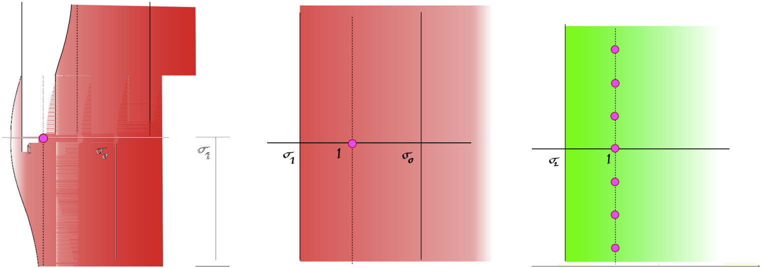


Situation 2
Vertical strip



Situation 3
Vertical strip with holes

Different possible regions on the left of $<s = 1$ where $\Lambda(s)$ is tame.



Situation 1	Situation 2	Situation 3
Hyperbolic region	Vertical strip	Vertical strip with holes

For which simple sources do these different situations occur?

For memoryless sources relative to probabilities $(p_1; p_2; : : : ; p_r)$

– S2 is impossible

– S3 occurs when all the ratios $\log p_i = \log p_j$ are rational

– S1 occurs if there exists a ratio $\log p_i = \log p_j$

which is badly approximable by rationals.

In this case

$$\Lambda(s) = \frac{1}{1 \quad (s)} \qquad \text{with} \qquad (s) = p_1^s + p_2^s = \dots p_r^s.$$

Case $r = 2$. Suppose that there exists $t \neq 0$ for which $\quad (1 + it) = 1.$
Then $t \neq 0$ is solution of $p_1 p_1^{it} + p_2 p_2^{it} = 1.$

Then, as $p_1 + p_2 = 1$, this implies:

$$1 = p_1 |p_1^{it}| + p_2 |p_2^{it}|, \qquad |p_1 p_1^{it} + p_2 p_2^{it}| = 1$$

The "converse of the triangular inequality" entails $p_1^{it} = p_2^{it} = 1$

$$\Longrightarrow \qquad t \log p_1 = 2 q_1 \ ; \qquad t \log p_2 = 2 q_2$$

$$\Longrightarrow \qquad \frac{\log p_1}{\log p_2} \in \mathbb{Q}$$

$$\Longrightarrow \qquad s \mapsto \quad (s) \text{ is periodic with period } it$$

Memoryless sources – Case when there are poles close to $\Re s = 1$

Knowing $p_1 + p_2 = 1$,

Look for a solution s of $p_1^s + p_2^s = 1$ when $s = \sigma + it$, with σ close to 1

$$p_1^\sigma p_1^{it} + p_2^\sigma p_2^{it} = 1 \quad \Longrightarrow \quad p_1^{it} \approx 1, \quad p_2^{it} \approx 1$$

$$\Longrightarrow \quad \exists q_1, q_2 \in \mathbb{Z} \quad \text{for which} \quad t \approx \frac{2\pi}{\log p_1} q_1, \quad t \approx \frac{2\pi}{\log p_2} q_2$$

Thus

$$\frac{\log p_2}{\log p_1} \approx \frac{q_2}{q_1}$$

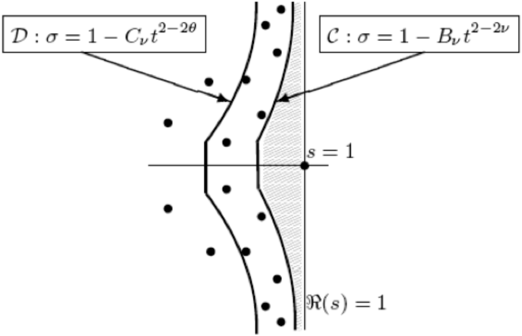
The poles of $\Lambda(s)$ close to $\Re s = 1$

are related to good rational approximations of $\log p_2 / \log p_1$

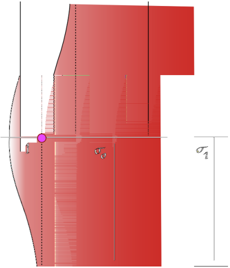
A number x is diophantine with exponent if there exists C for which

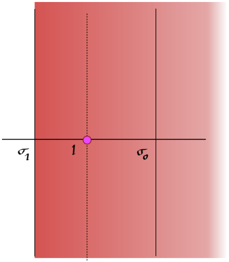
$$\left|x \quad \frac{a}{b}\right| > \frac{C}{b^{\mu}} \qquad 8a; b\ 2\ \mathbb{Z}$$

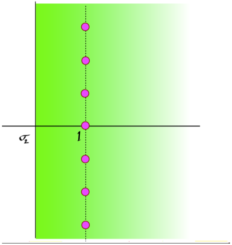
Consider $\Lambda(s) = 1{=}(1 \quad p_1^s \quad p_2^s)$. If $\log p_1 {=} \log p_2$ is –diophantine, then, for any ; with < < , the tameness region has an hyperbolic form: [Flajolet-Roux-V. 2010]



Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.







Situation 1
Hyperbolic region
Arithmetic condition

Situation 2
Vertical strip
Geometric condition

Situation 3
Vertical strip with holes
Periodicity condition

For which general dynamical sources do these different situations occur?

– S3 occurs only if the source is conjugated to a simple source.

– S2 occurs when " the branches are not too often of the same shape".

– S1 occurs if a extension of the following condition holds:

"there exists a ratio $\log p_i = \log p_j$ which is badly approximable by rationals".

Situation 2- Existence of a vertical strip where $\Lambda(s)$ is tame

There exists a condition, the condition UNI, which expresses that
" the branches of the dynamical system are not too often of the same form"

Theorem [Dolgopyat-Baladi-Cesaratto-V]. For a good dynamical system [complete, expansive, with bounded distortion], which satisfies the condition *UNI*, there exists a vertical strip where its Dirichlet series $\Lambda(s)$ is tame.

Dolgopyat (98) proves the result for the plain transfer operator, in the case of a finite number of branches

– Baladi and V. (03) extend the result for an infinite number of branches

– Cesaratto and V. (09) extend the result to the secant transfer operator.

Description of the UNI Condition

A distance Δ between two inverse branches of the same depth:

$$\Delta(h;k) := \inf_{x\in\mathcal{I}} \Psi'_{h,k}(x); \quad \text{with} \quad \Psi_{h,k}(x) := \log \frac{|h'(x)|}{|k'(x)|}$$

Contraction ratio . $:= \limsup\left(\{\max|h'(x)|; h \in H^n; x \in I\}\right)^{1/n}$.

Probability $\Pr_n$ on $H^n \quad H^n$. $\Pr_n(h;k) := |h(I)| \; |k(I)|$

For a system C^2–conjugated with a piecewise-affine system :
For any ˆ with $< \hat{} < 1$, for any n, $\Pr_n[\Delta < \hat{}^n] = 1$

Condition UNI.
For any ˆ with $< \hat{} < 1$, for any n, $\Pr_n[\Delta < \hat{}^n] \ll \hat{}^n$

Situation 3- Existence of a hyperbolic region where $\Lambda(s)$ is tame

The condition DIOP extends the arithmetic condition

"There exists a ratio $\log p_i = \log p_j$ which is diophantine"

to the general dynamical case.

Condition DIOP: There exists a ratio $c(h; k)$ which is diophantine.

For a complete system, each branch h has a fixed point denoted by $h^\star$.

We consider the ratios between the derivatives $jh'(h^\star)j$ at the fixed point

$$c(h; k) := \frac{\log jh'(h^\star)j}{\log jk'(k^\star)j}$$

Theorem [Dolgopyat-Roux-V.] For a good dynamical system [complete, expansive, with bounded distortion], which satisfies the condition *DIOP*, there exists an hyperbolic region where $\Lambda(s)$ is tame.

Dolgopyat (98) proves the result for the plain transfer operator, in the case of a finite number of branches – Roux and V. (2010) extend the result : for an infinite number of branches and for the secant transfer operator.

Plan of the talk.

– General motivations

– Models for a source

– The Dirichlet generating function of the source

– Conclusion and possible extensions.

Conclusions.

Description of a framework which

– unifies the analyses for text algorithms and searching/sorting algorithms

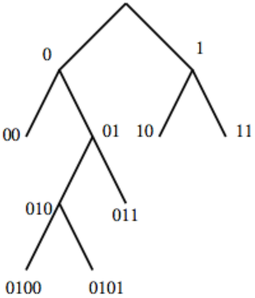
– provides a general model for sources
– shows the importance of the Dirichlet generating functions
– explains the importance of tameness for sources

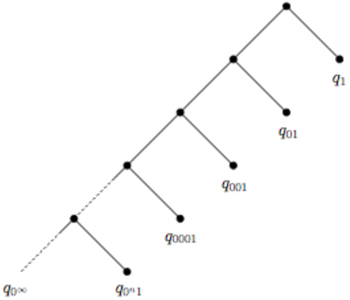
– defines a natural subclass of sources, the dynamical sources
– provides sufficient conditions for tameness of dynamical sources

– provides probabilistic analyses for algorithms built on tame sources.

Possible extensions and work in progress
I– Classification of sources

- Place of dynamical sources amongst general sources:
- A dynamical source = limit of Markov chains with increasing order?
- Comparing dynamical sources with Markov chains of variable length





Possible extensions and work in progress

II– Realistic analyses of other algorithms and other structures

– Analysis of other sorting algorithms

  – Analysis of `Insertion` Sort easy

  – Analysis of `QuickSelect` already done

  – And `Selection` algorithm ?

– Analysis of the DST structure?