INFORMATION THEORY: SOURCES, DIRICHLET SERIES, REALISTIC ANALYSIS OF DATA STRUCTURES

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

Talk also based on joint works with
Viviane Baladi, Eda Cesaratto, Julien Clément,
Jim Fill, Philippe Flajolet

WORDS 2011, Prague, September 2011

Description of a framework which

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

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

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

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 data structures built on tame sources.

Plan of the talk.

– General motivations: Dirichlet generating functions and tameness

– An important class of sources: dynamical sources.

– Tameness in the case of dynamical sources

– Conclusion and possible extensions.

Plan of the talk.

– General motivations: Dirichlet generating functions and tameness.

– An important class of sources: dynamical sources.

– Tameness in the case of dynamical sources

– Conclusion and possible extensions.

The classical framework for analysis of algorithms

in two main algorithmic domains:

Text algorithms – Sorting or Searching algorithms.

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

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

A word or a key are both a sequence of symbols ... but

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

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

A word or a key are both a sequence of symbols ... but

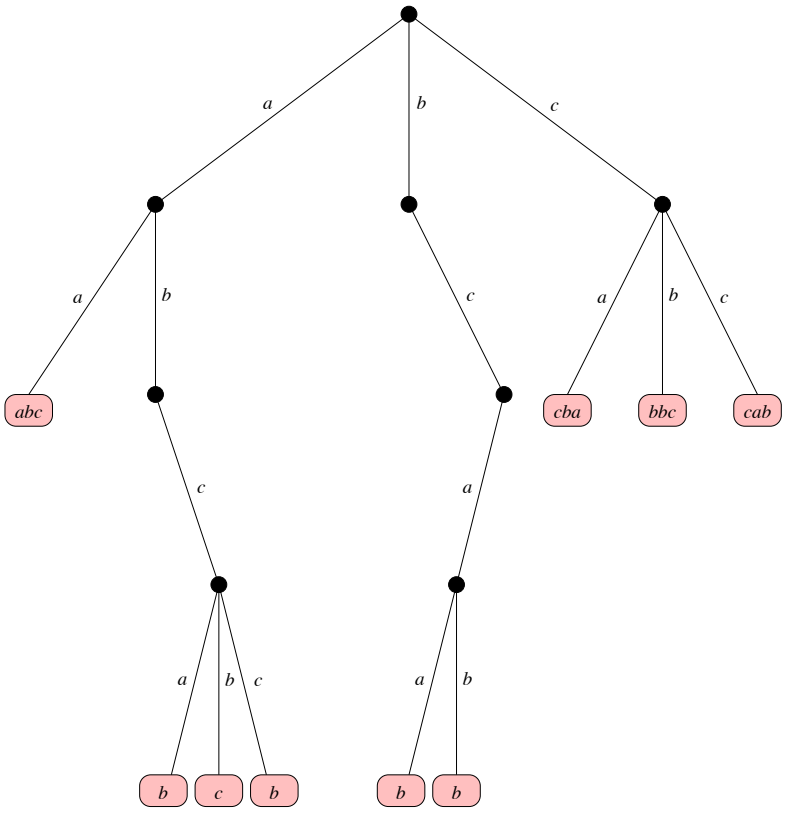
– for comparing two words, importance of the structure of words

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

only their relative order plays a role.

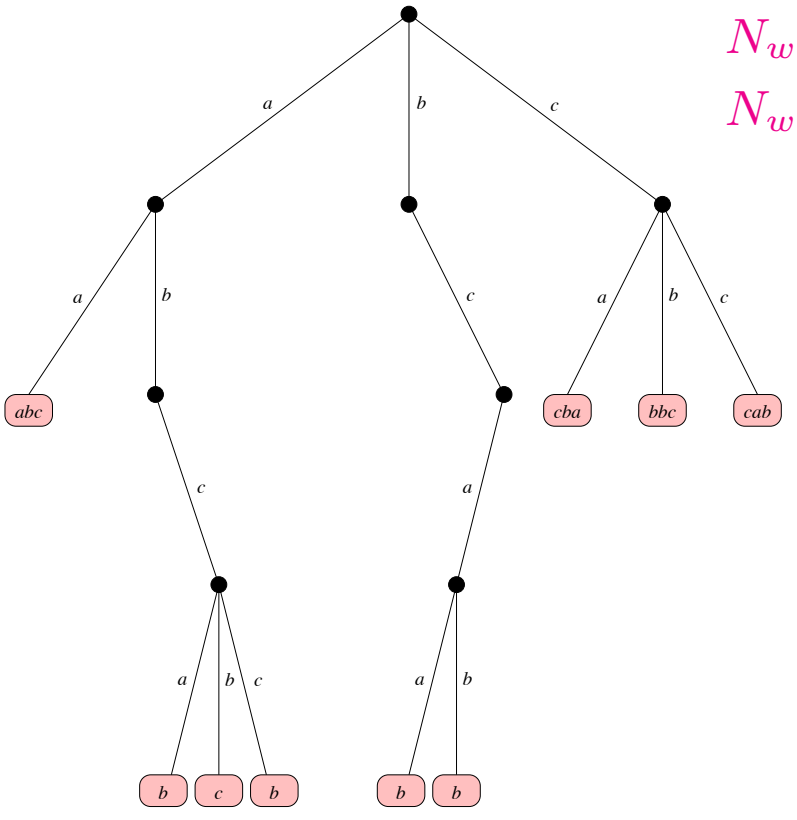
Text algorithms and dictionaries : The trie structure

Probabilistic study



Text algorithms and dictionaries : The trie structure

Probabilistic study



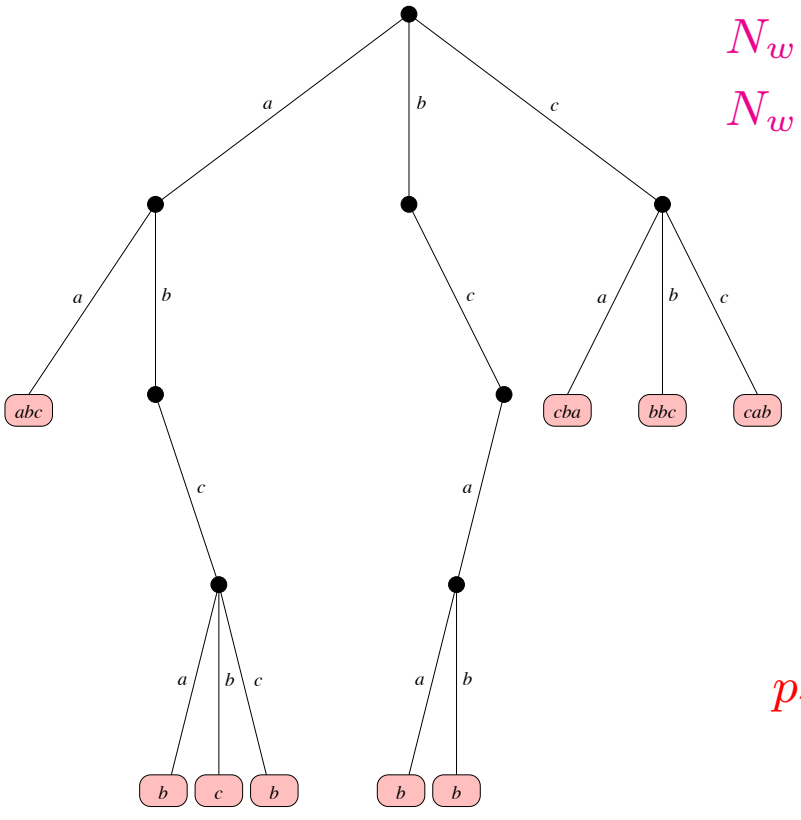
Main parameter on a node n_w labelled with prefix w:

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

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

Text algorithms and dictionaries : The trie structure

Probabilistic study



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

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

Central role of

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

A realistic framework for sorting or searching.

Keys are viewed as words and are compared [wrt the lexicographic order].

The realistic unit cost is now the symbol–comparison.

A realistic framework for sorting or searching.

Keys are viewed as words and 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

$$k := \max\{i; \quad \forall j \leq i, \quad a_j = b_j\} = \text{the coincidence } c(A, B)$$

A realistic framework for sorting or searching.

Keys are viewed as words and are compared [wrt the lexicographic order].

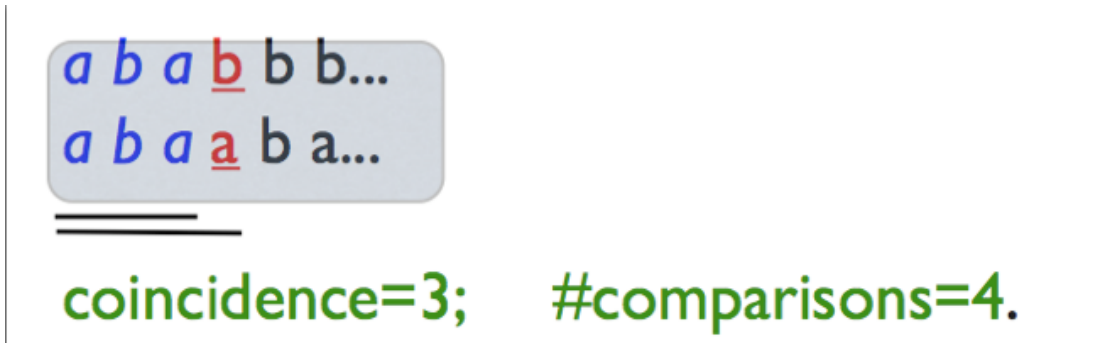
The realistic unit cost is now the symbol–comparison.

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

$$k := \max\{i; \quad \forall j \leq i, \quad a_j = b_j\} = \text{the coincidence } c(A,B)$$



A realistic framework for sorting or searching.

Keys are viewed as words and are compared [wrt the lexicographic order].

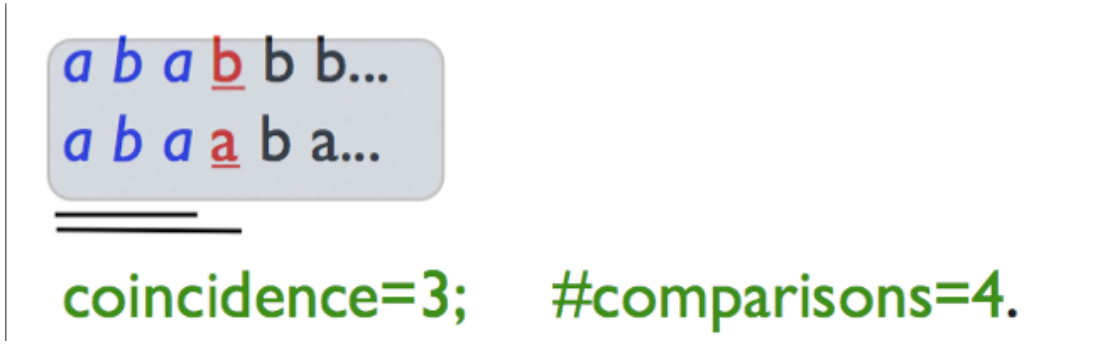
The realistic unit cost is now the symbol–comparison.

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

$$k := \max\{i; \quad \forall j \leq i, \quad a_j = b_j\} = \text{the coincidence } c(A,B)$$



The probabilistic study of the coincidence deals with

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

$\Pr[c(A,B) \geq k] = \Pr[A \text{ and } B \text{ begin with the same } w \text{ of length } k]$

A realistic framework for sorting or searching.

Keys are viewed as words and are compared [wrt the lexicographic order].

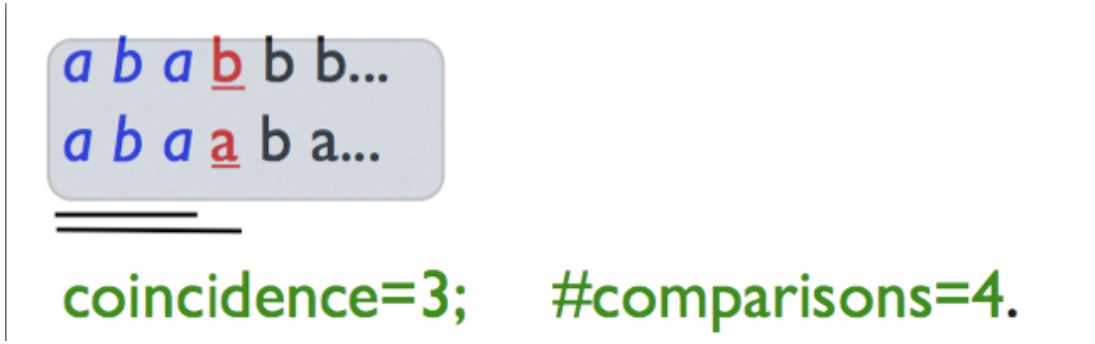
The realistic unit cost is now the symbol–comparison.

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

$$k := \max\{i; \quad \forall j \leq i, \quad a_j = b_j\} = \text{the coincidence } c(A,B)$$

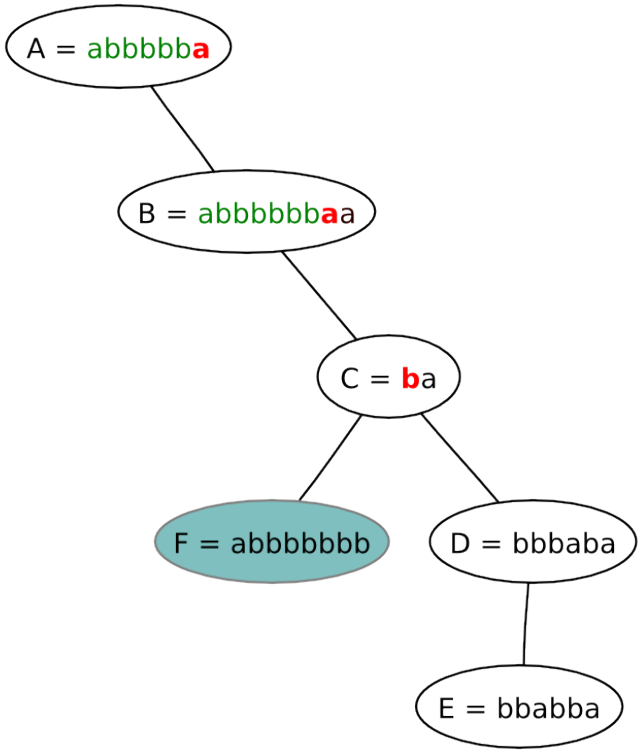


The probabilistic study of the coincidence deals with

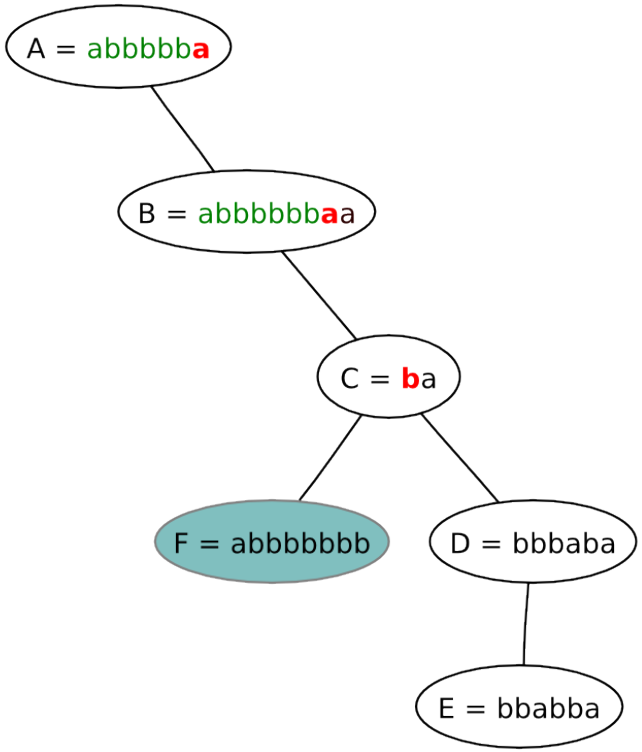
$p_w :=$ the probability that a word begins with prefix w.

$$\Pr[c(A,B) \geq k] = \Pr[A \text{ and } B \text{ begin with the same } w \text{ of length } k] = \sum_{|w|=k} p_w^2$$

The example of the binary search tree (BST)

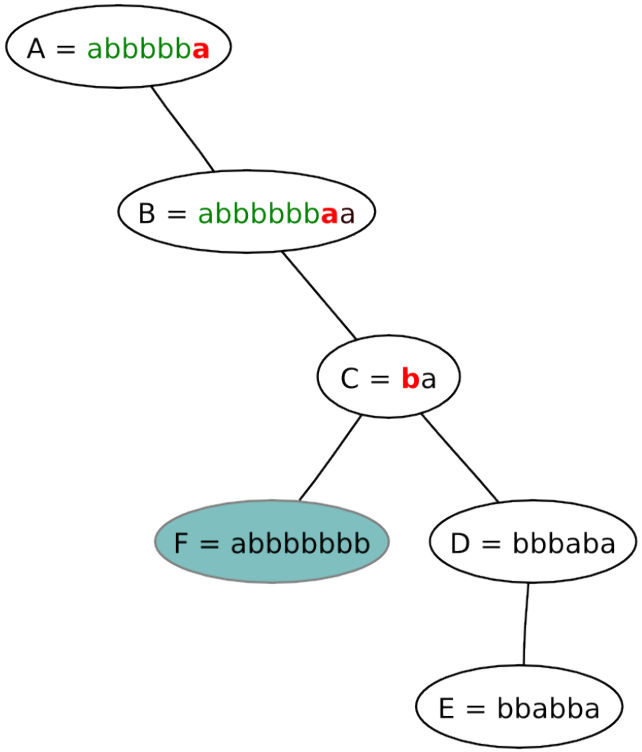


The example of the binary search tree (BST)



Number of symbol comparisons needed for inserting F = abbbbbbb.

The example of the binary search tree (BST)

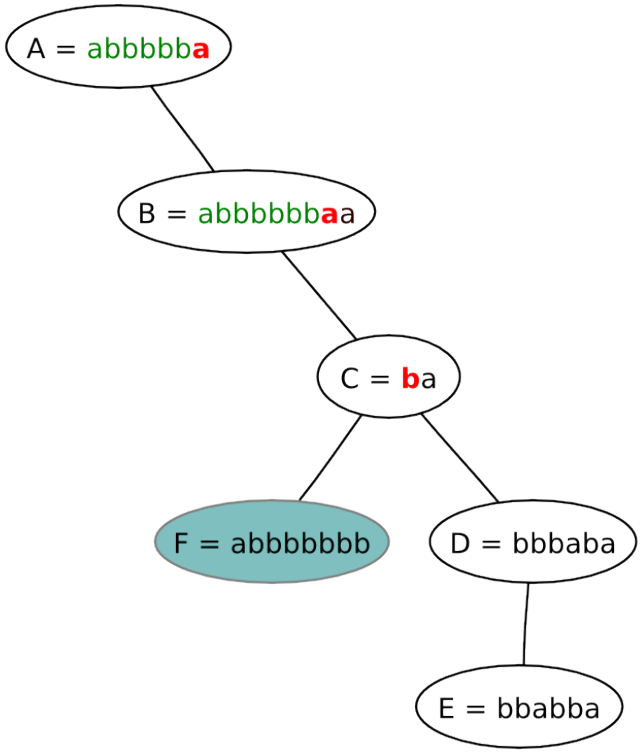


Number of symbol comparisons

needed for inserting F = abbbbbbb.

= 7 for comparing to A $c(F, A) = 6$

The example of the binary search tree (BST)



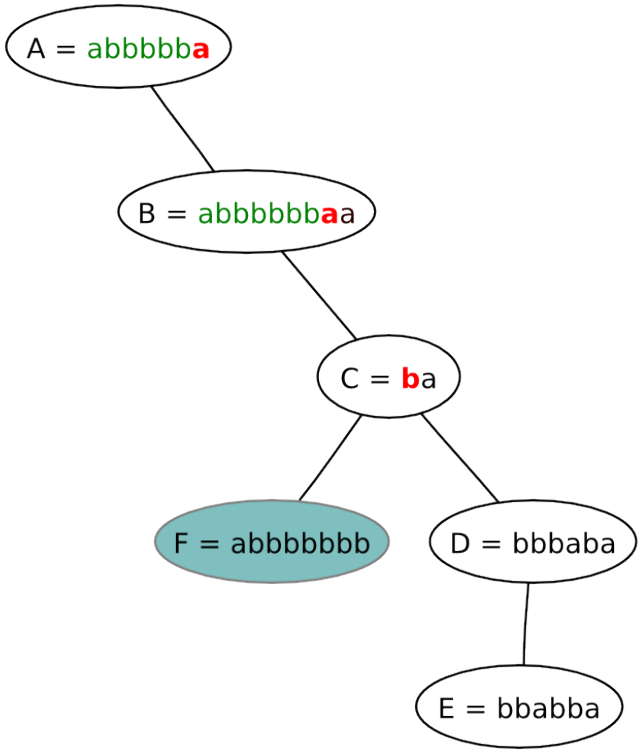
Number of symbol comparisons

needed for inserting F = abbbbbbb.

$= 7$ for comparing to A $\quad c(F, A) = 6$

$+ 8$ for comparing to B $\quad c(F, B) = 7$

The example of the binary search tree (BST)



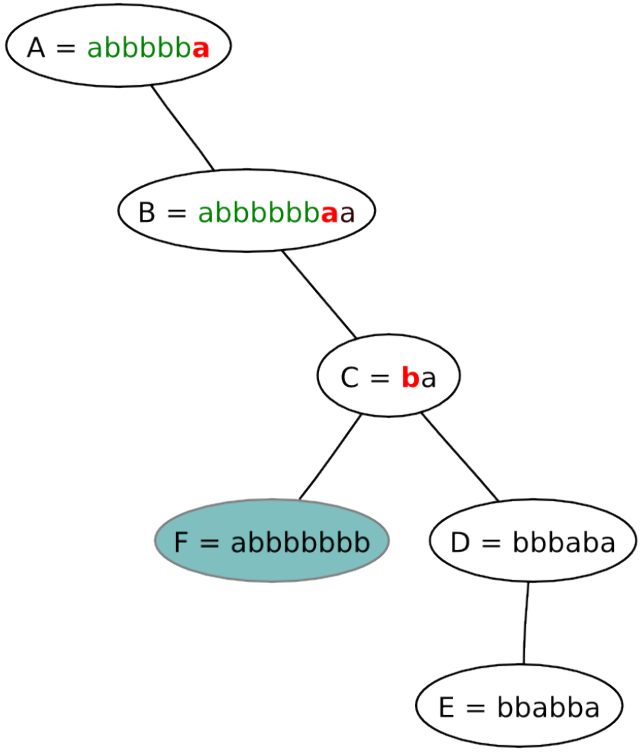
Number of symbol comparisons
needed for inserting F = abbbbbbb.

$= 7$ for comparing to A $\quad c(F, A) = 6$

$+ 8$ for comparing to B $\quad c(F, B) = 7$

$+ 1$ for comparing to C $\quad c(F, C) = 0$

The example of the binary search tree (BST)

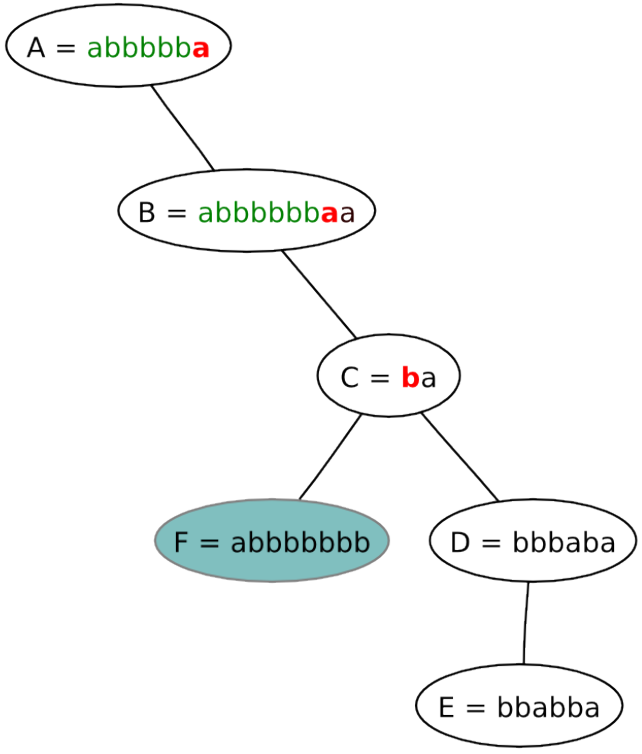


Number of symbol comparisons
needed for inserting F = abbbbbbb.

$= 7$ for comparing to A $\quad c(F, A) = 6$
$+ 8$ for comparing to B $\quad c(F, B) = 7$
$+ 1$ for comparing to C $\quad c(F, C) = 0$

Total $= 16$
To be compared to
the number of key comparisons $[= 3]$

The example of the binary search tree (BST)



Number of symbol comparisons
needed for inserting F = abbbbbbb.

$= 7$ for comparing to A $\quad c(F, A) = 6$
$+ 8$ for comparing to B $\quad c(F, B) = 7$
$+ 1$ for comparing to C $\quad c(F, C) = 0$

Total $= 16$
To be compared to
the number of key comparisons $[= 3]$

This defines the symbol-path-length of a BST based on the coincidence
We perform a probabilistic study of this symbol path-length

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

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.

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 $\Sigma^{\mathbb{N}}$.

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 $\{p_w, \quad w \in \Sigma^\star\}$ defines the source $\mathcal{S}$.

Fundamental role of the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

Fundamental role of the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

Fundamental role of the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w \in \Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k \geq 0} \Lambda_k \right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

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

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

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

Fundamental role of the Dirichlet generating functions of the source

$$\Lambda(s) := \sum_{w\in\Sigma^\star} p_w^s, \qquad \Lambda_k(s) = \sum_{w\in\Sigma^k} p_w^s, \qquad \left[\Lambda = \sum_{k\geq 0} \Lambda_k\right]$$

Remark: $\Lambda_k(1) = 1$ for any k, $\qquad \Lambda(1) = \infty$.

– they encapsulate the main probabilistic properties of the source

– they translate them into analytic properties

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

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

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

– they intervene in probabilistic analysis of algorithms and data structures.

Exact average-case analysis for Tries or BST's

$S_n^{(X)}$:= the mean path-length for the Trie $[X = T]$
or the mean symbol path-length of the BST $[X = B]$
when built on n words independently drawn from the same source.

Exact average-case analysis for Tries or BST's

$S_n^{(X)}$:= the mean path-length for the Trie $[X = T]$
or the mean symbol path-length of the BST $[X = B]$
when built on n words independently drawn from the same source.

For each case [$X = T$ or $X = B$] an exact formula for $S_n^{(X)}$

$$S_n^{(X)} = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_X(k)$$

which involves a series ϖ_X at integer values k.

Clément, Flajolet, V. (2001) for $X = T$, Clément, Fill, Flajolet, V. (2009) for $X = B$

Exact average-case analysis for Tries or BST's

$S_n^{(X)}$:= the mean path-length for the Trie $[X = T]$
or the mean symbol path-length of the BST $[X = B]$
when built on n words independently drawn from the same source.

For each case [$X = T$ or $X = B$] an exact formula for $S_n^{(X)}$

$$S_n^{(X)} = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_X(k)$$

which involves a series ϖ_X at integer values k.

Clément, Flajolet, V. (2001) for $X = T$, Clément, Fill, Flajolet, V. (2009) for $X = B$

This series $\varpi_X(s)$ is closely related to the Dirichlet series of the source

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

Exact average-case analysis for Tries or BST's

$S_n^{(X)}$:= the mean path-length for the Trie $[X = T]$
or the mean symbol path-length of the BST $[X = B]$
when built on n words independently drawn from the same source.

For each case [$X = T$ or $X = B$] an exact formula for $S_n^{(X)}$

$$S_n^{(X)} = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_X(k)$$

which involves a series ϖ_X at integer values k.

Clément, Flajolet, V. (2001) for $X = T$, Clément, Fill, Flajolet, V. (2009) for $X = B$

This series $\varpi_X(s)$ is closely related to the Dirichlet series of the source

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

Nice exact formulae, not easy to deal with, due to the alternating signs

Asymptotic analysis.

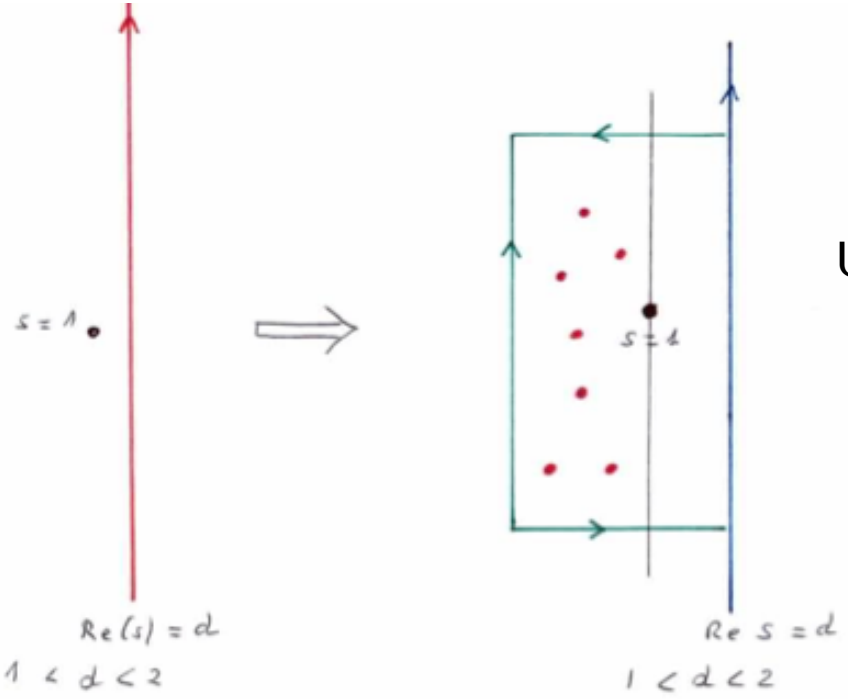
The residue formula transforms the sum into an integral with $1 < d < 2$.

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

Asymptotic analysis.

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

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



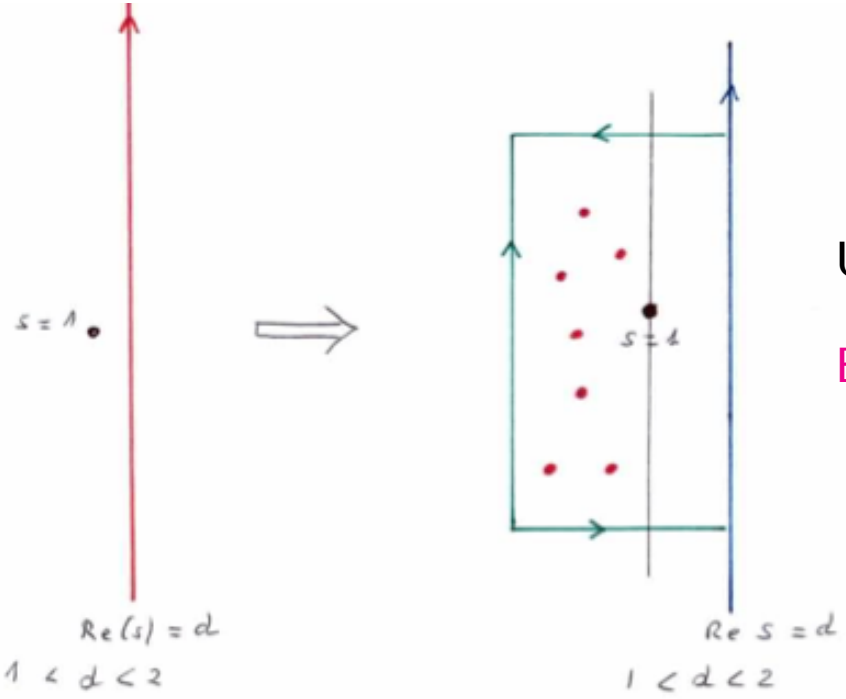
We shift the integral on the left,

Usually, the first singularities occur at $\Re s = 1$.

Asymptotic analysis.

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

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



We shift the integral on the left,

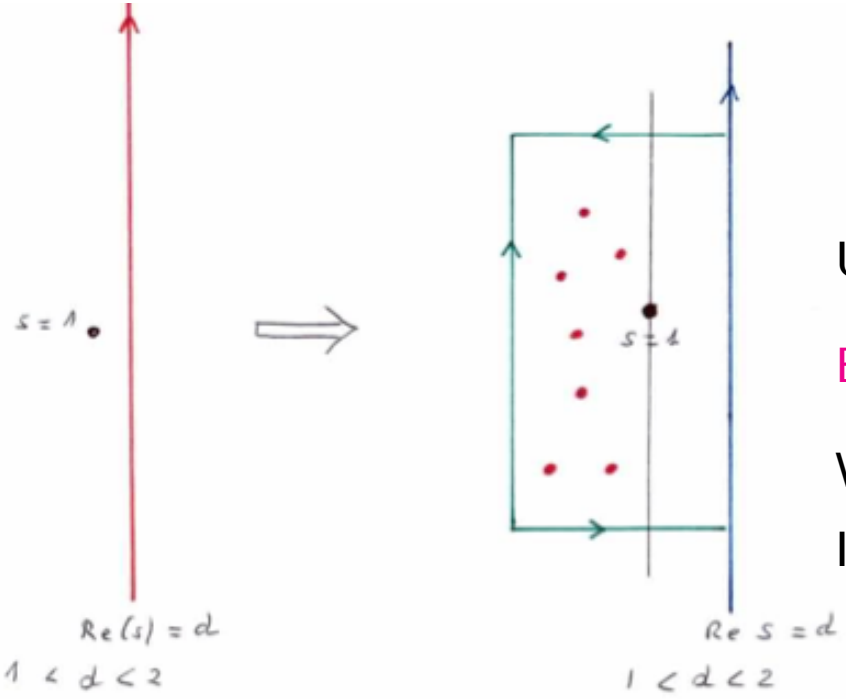
Usually, the first singularities occur at $\Re s = 1$.

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

Asymptotic analysis.

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

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



We shift the integral on the left,

Usually, the first singularities occur at $\Re s = 1$.

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

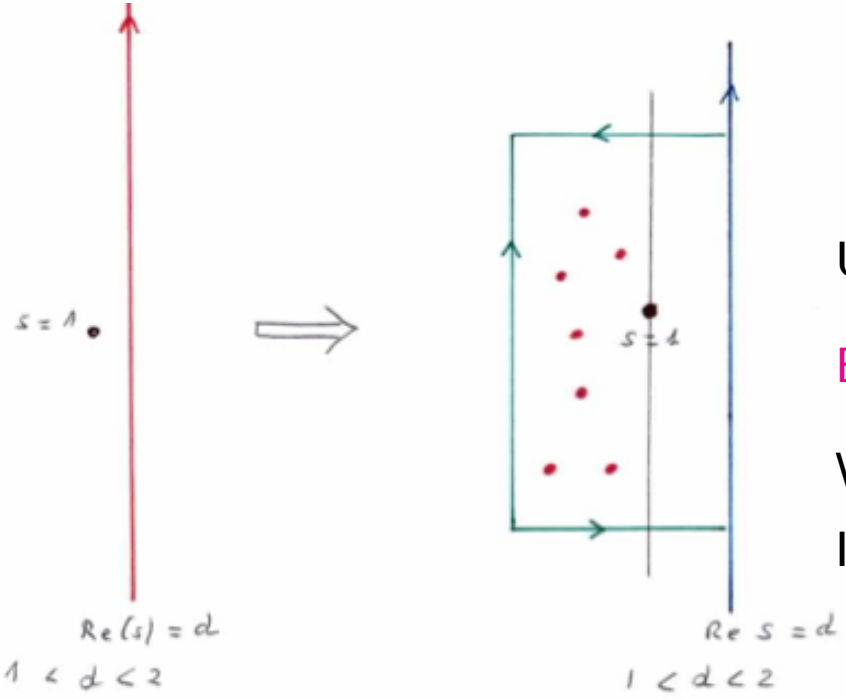
Where are the red singularities closest to $\Re s = 1$?

Is $\Lambda(s)$ of polynomial growth on the green contour?

Asymptotic analysis.

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

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



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 $\mathcal{R}$

– which contains only $s = 1$ as a pole – where $\Lambda(s)$ is of polynomial growth.

Tameness of the source

Main results

[Clément, Flajolet, V. (2001), Clément, Flajolet, Fill, V. (2009)]

Consider n words independently drawn from the same tame source. Then:

The mean path-length T_n of the Trie satisfies	The mean symbol path-length B_n of the BST satisfies
$T_n \sim \frac{1}{h_{\mathcal{S}}}\, n \log n.$	$B_n \sim \frac{1}{h_{\mathcal{S}}}\, n \log^2 n.$

Here, $h_{\mathcal{S}}$ is the entropy $h_{\mathcal{S}}$ of the source $\mathcal{S}$, 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.

Plan of the talk.

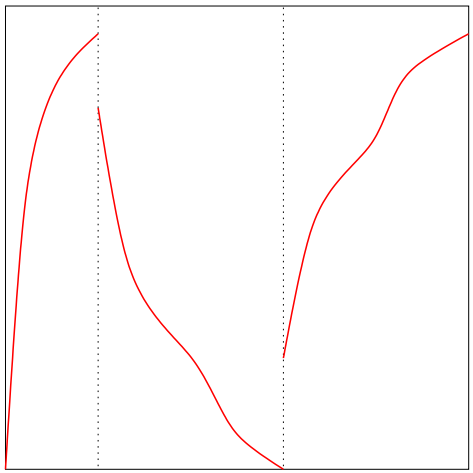
– General motivations: Dirichlet generating functions and tameness

– An important class of "natural" sources: dynamical sources
= sources associated to dynamical systems

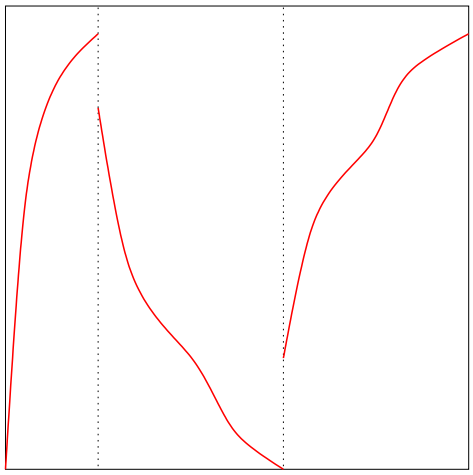
– Tameness in the case of dynamical sources

– Conclusion and possible extensions.

A dynamical source = a source built with a dynamical system [V. 1998]



A dynamical source = a source built with a dynamical system [V. 1998]



A dynamical system $(\mathcal{I}, T)$ is defined by

– an alphabet Σ denumerable (possibly infinite),

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

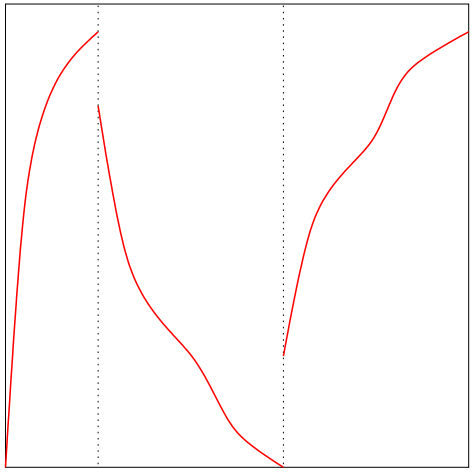
– an encoding mapping σ equal to m on each $\mathcal{I}_m$,

– a shift mapping T

– each $T|_{\mathcal{I}_m}$ is a bijection of class $\mathcal{C}^2$ on $\mathcal{I}_m$

– The local inverse of $T|_{\mathcal{I}_m}$ is denoted by h_m.

A dynamical source = a source built with a dynamical system [V. 1998]



A dynamical system $(\mathcal{I}, T)$ is defined by

– an alphabet Σ denumerable (possibly infinite),

– a topological partition of $\mathcal{I} :=]0,1[$

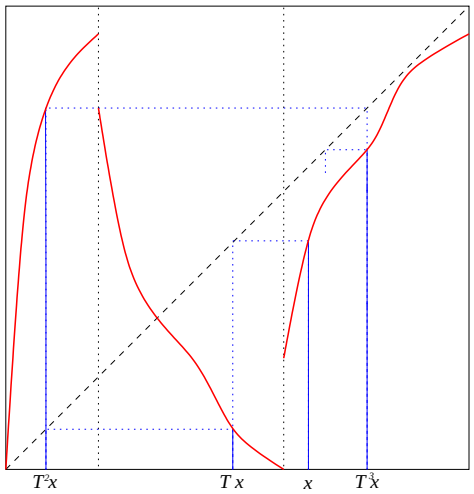
with open intervals $\mathcal{I}_{m, m\in\Sigma}$,

– an encoding mapping σ equal to m on each $\mathcal{I}_m$,

– a shift mapping T

– each $T|_{\mathcal{I}_m}$ is a bijection of class $\mathcal{C}^2$ on $\mathcal{I}_m$

– The local inverse of $T|_{\mathcal{I}_m}$ is denoted by h_m.



$M(x) = (c, b, a, c \ldots)$

This gives rise to a source:

On an input x of $\mathcal{I}$, it outputs the word

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

When an initial density is chosen on $\mathcal{I}$,

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

= a dynamical source

Strong relations between the geometry of the system,

the correlations between symbols and the probabilistic properties of the source.

Two geometric characteristics of the system:

- The position of the branches $T(\mathcal{I}_k)$ w.r.t $\mathcal{I}_m$
- The shape of the branches defined by the derivative of h_m

Particular cases: simple sources and affine branches

Strong relations between the geometry of the system,

the correlations between symbols and the probabilistic properties of the source.

Two geometric characteristics of the system:

- The position of the branches $T(\mathcal{I}_k)$ w.r.t $\mathcal{I}_m$
- The shape of the branches defined by the derivative of h_m

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 $\mathcal{I}_m$.

Strong relations between the geometry of the system,

the correlations between symbols and the probabilistic properties of the source.

Two geometric characteristics of the system:

- The position of the branches $T(\mathcal{I}_k)$ w.r.t $\mathcal{I}_m$
- The shape of the branches defined by the derivative of h_m

Particular cases: simple sources and affine branches

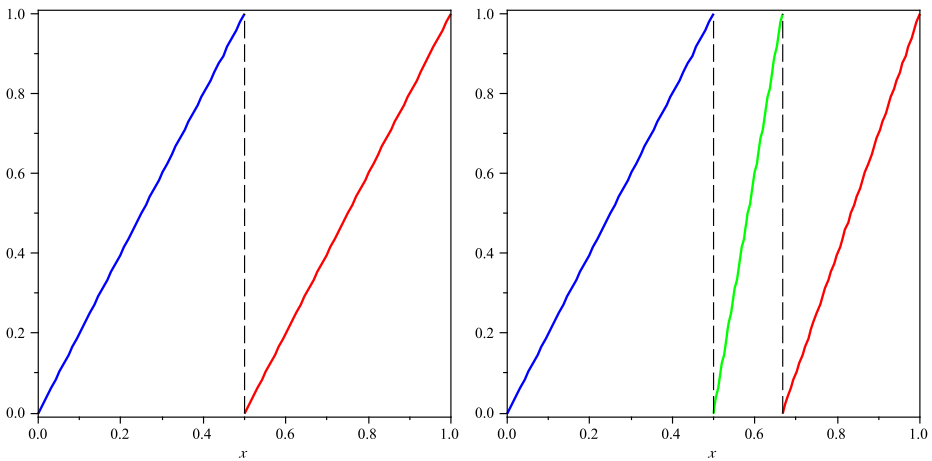
A memoryless source

:= 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 $\mathcal{I}_m$.



Strong relations between the geometry of the system,

the correlations between symbols and the probabilistic properties of the source.

Two geometric characteristics of the system:

- The position of the branches $T(\mathcal{I}_k)$ w.r.t $\mathcal{I}_m$
- The shape of the branches defined by the derivative of h_m

Particular cases: simple sources and affine branches

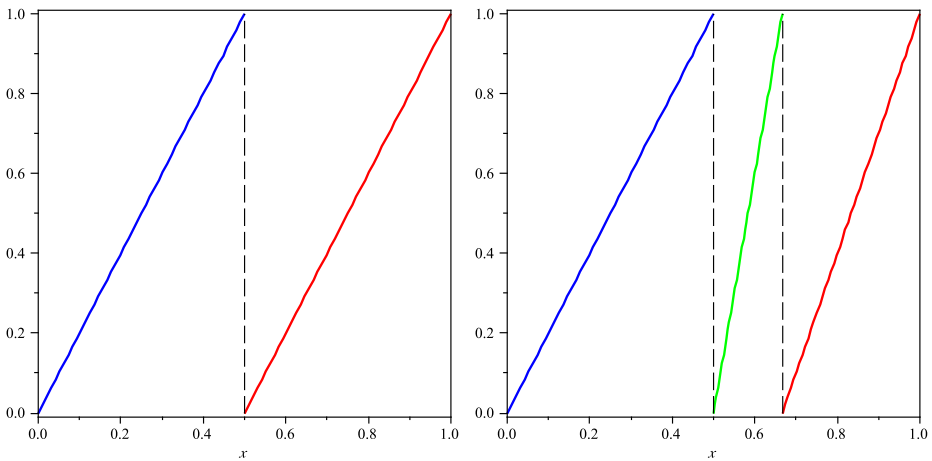
A memoryless source

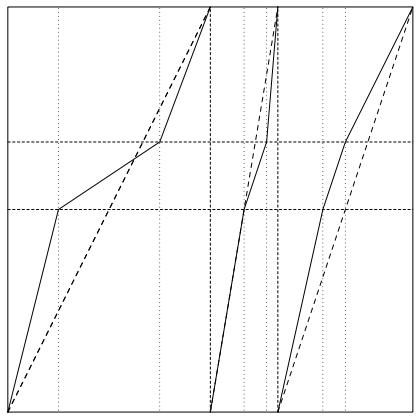
:= a complete system with affine branches and uniform initial density

A Markov chain

:= a Markovian system with affine branches,

with an initial density which is constant on each $\mathcal{I}_m$.





General case of interest = the Good Class gathers

– Complete systems: $T(\mathcal{I}_m) = \mathcal{I}$

– with a possible infinite denumerable alphabet

– with expansive branches : $|T'(x)| \geq \rho > 1$.

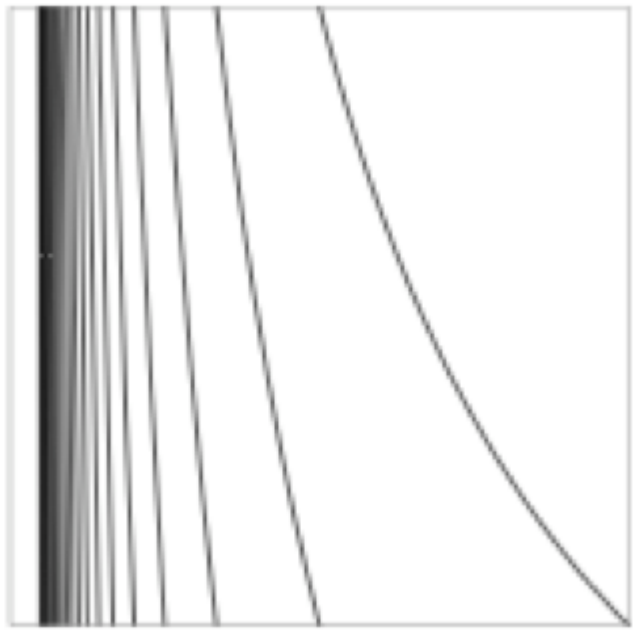
General case of interest = the Good Class gathers

– Complete systems: $T(\mathcal{I}_m) = \mathcal{I}$

– with a possible infinite denumerable alphabet

– with expansive branches : $|T'(x)| \geq \rho > 1$.

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



A main analytical object related to any source:

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

A main analytical object related to any source:

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

Memoryless sources, with probabilities (p_i)

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

A main analytical object related to any source:

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

Memoryless sources, with probabilities (p_i)

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

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

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

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

A main analytical object related to any source:

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

Memoryless sources, with probabilities (p_i)

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

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

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

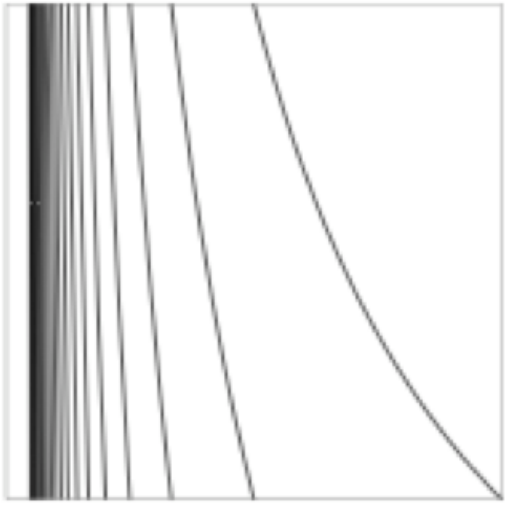
$$\Lambda(s) = 1 + {}^t\mathbf{1}(I - P(s))^{-1} R(s) \qquad \text{with} \quad P(s) = (p_{i,j}^s), \quad R(s) = (r_i^s).$$

A general dynamical source

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

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

The density transformer and the transfer operators



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

with $\mathbf{H}_{[m]}[f](x) = |h'_m(x)| \cdot f \circ h_m(x)$

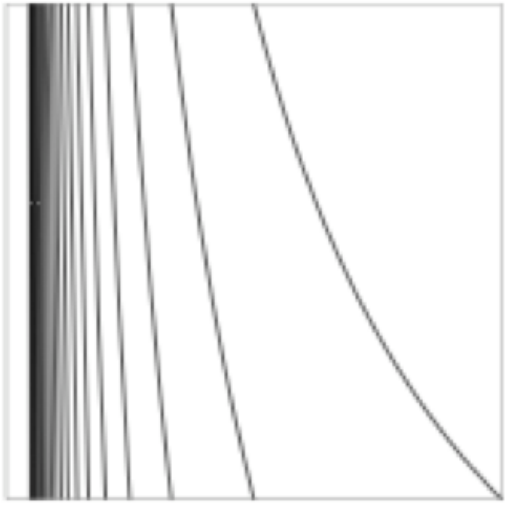
is the density transformer of the dynamical system.

It describes the evolution of the density.

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

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

The density transformer and the transfer operators



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

with $\mathbf{H}_{[m]}[f](x) = |h'_m(x)| \cdot f \circ h_m(x)$

is the density transformer of the dynamical system.

It describes the evolution of the density.

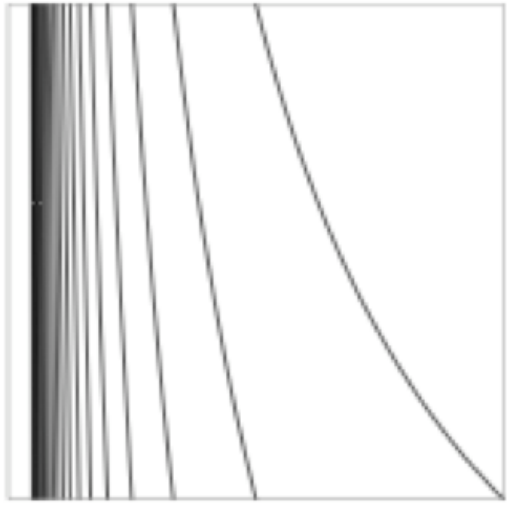
For a density f on $[0,1]$,

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

Transfer operator (Ruelle) [tangent version]

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

The density transformer and the transfer operators



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

with $\mathbf{H}_{[m]}[f](x) = |h'_m(x)| \cdot f \circ h_m(x)$

is the density transformer of the dynamical system.

It describes the evolution of the density.

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

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

Transfer operator (Ruelle) [tangent version]

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

Transfer operator (Vallée, 2000) [secant version]

$$\mathbb{H}_s := \sum_{m\in\Sigma} \mathbb{H}_{s,[m]} \quad \text{with} \quad \mathbb{H}_{s,[m]}[F](x,y) = \left|\frac{h_m(x) - h_m(y)}{x-y}\right|^s F(h_m(x), h_m(y))$$

Alternative expression of $\Lambda(s)$ in the dynamical case.

Alternative expression of $\Lambda(s)$ in the dynamical case.

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

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

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

with L the secant of the distribution function F.

Alternative expression of $\Lambda(s)$ in the dynamical case.

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

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

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

with L the secant of the distribution function F.

Singularities of $s \mapsto \Lambda(s)$ are essential in the analysis.

Singularities of $(I - \mathbb{H}_s)^{-1}$ are related to spectral properties of $\mathbb{H}_s$.

Alternative expression of $\Lambda(s)$ in the dynamical case.

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

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

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

with L the secant of the distribution function F.

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.
For a system of the Good Class, $s \mapsto \Lambda(s)$ has a simple pole at $s = 1$

Plan of the talk.

– General motivations: Dirichlet generating functions and tameness

– An important class of sources: dynamical sources.

– Tameness of dynamical sources

– Conclusion and possible extensions.

What happens on the left of the vertical line $\Re s = 1$?

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

– it is analytic (except for $s = 1$) and of polynomial growth ($\Im s \to \infty$)

What happens on the left of the vertical line $\Re s = 1$?

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

– it is analytic (except for $s = 1$) and of polynomial growth ($\Im s \to \infty$)

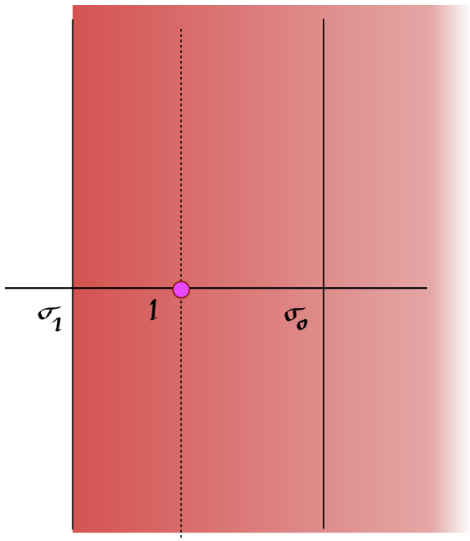
Different possible regions $\mathcal{R}$ on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.

What happens on the left of the vertical line $\Re s = 1$?

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

– it is analytic (except for $s = 1$) and of polynomial growth ($\Im s \to \infty$)

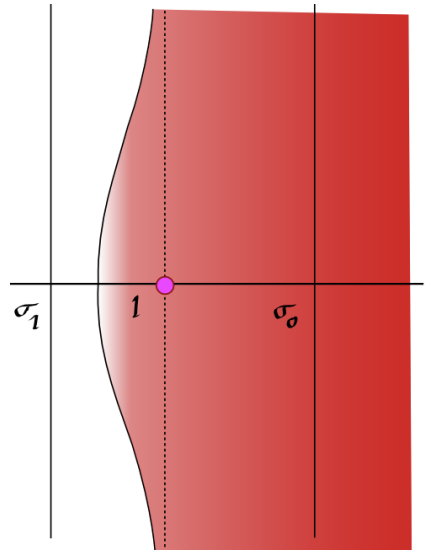
Different possible regions $\mathcal{R}$ on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

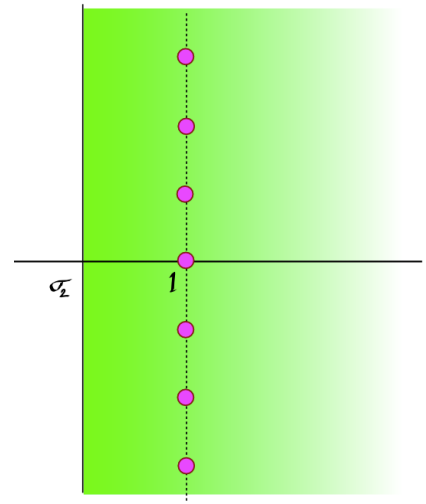
$1 - \sigma \leq a$



Situation 2

Hyperbolic region

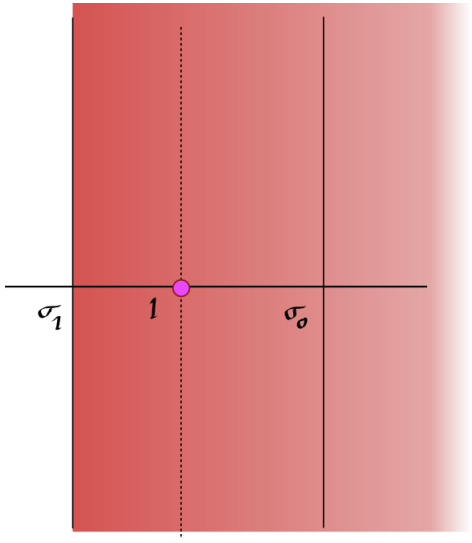
$1 - \sigma \leq t^{-\alpha}$



Situation 3

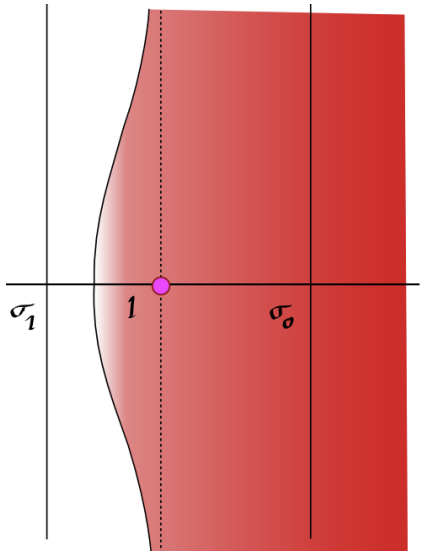
Vertical strip with holes

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



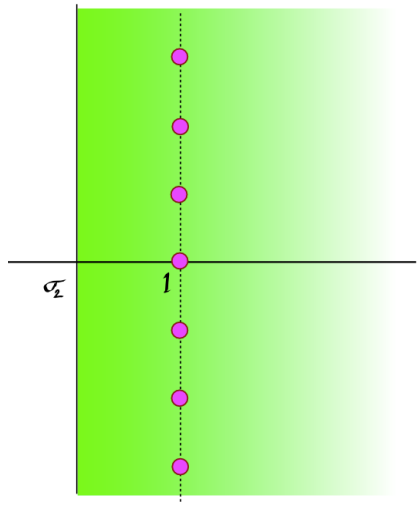
Situation 1

Vertical strip



Situation 2

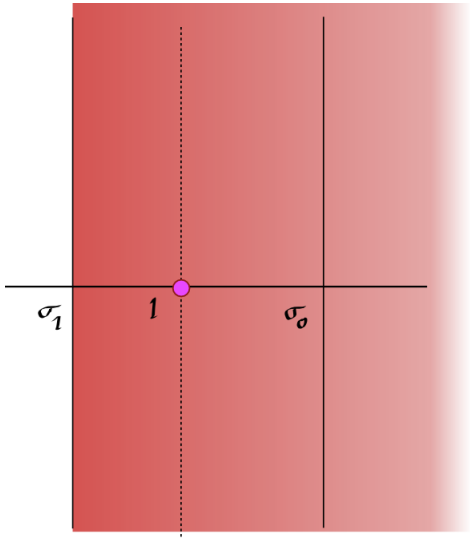
Hyperbolic region



Situation 3

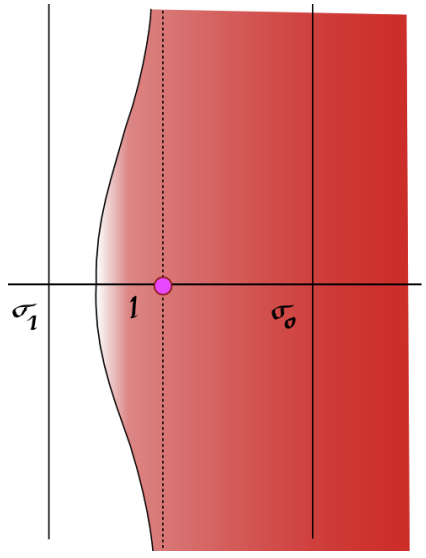
Vertical strip with holes

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



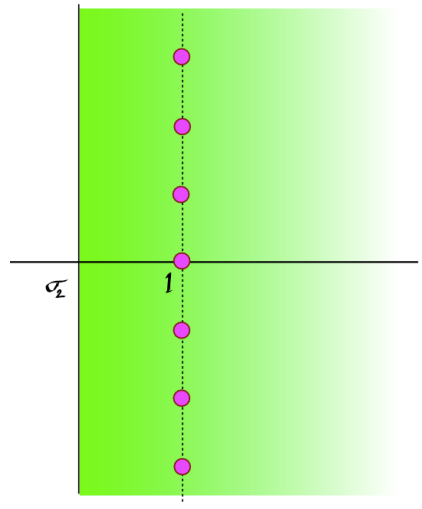
Situation 1

Vertical strip



Situation 2

Hyperbolic region

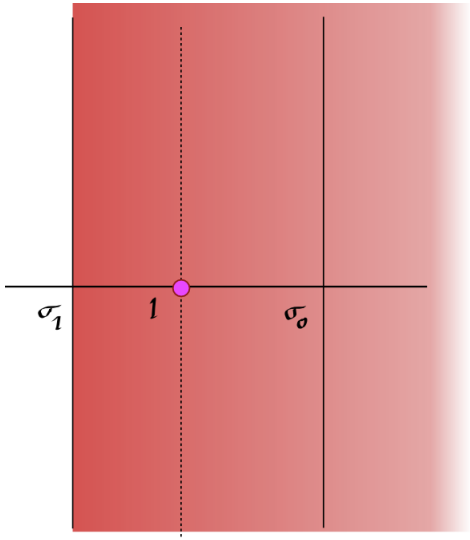


Situation 3

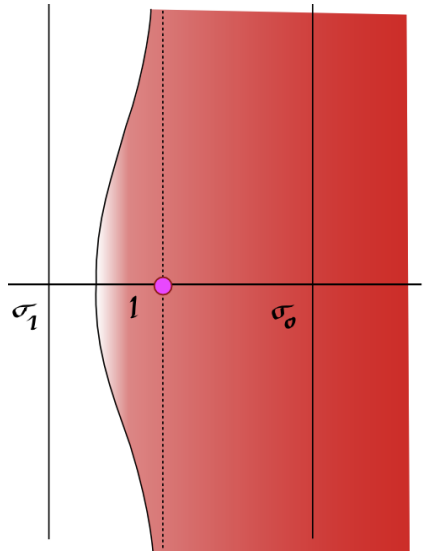
Vertical strip with holes

For which simple sources do these different situations occur?

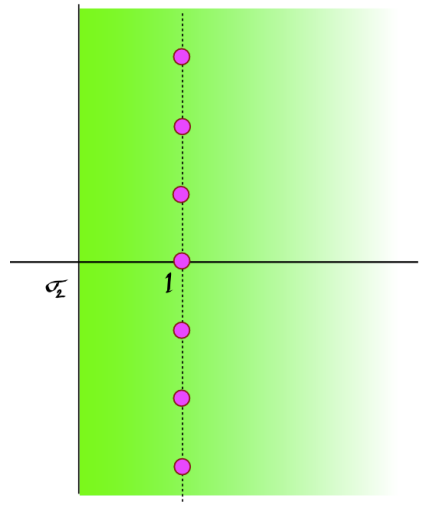
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1
Vertical strip



Situation 2
Hyperbolic region



Situation 3
Vertical strip with holes

For which simple sources do these different situations occur?

For memoryless sources relative to probabilities $(p_1, p_2, \ldots, p_r)$

– S1 is impossible

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

– S2 occurs if there exists a ratio $\log p_i / \log p_j$
which is "diophantine" [badly approximable by rationals]

Memoryless sources $\Lambda(s) = \dfrac{1}{1-\lambda(s)}$ with $\lambda(s) = p_1^s + p_2^s$ $[r = 2]$

Memoryless sources $\Lambda(s) = \dfrac{1}{1-\lambda(s)}$ with $\lambda(s) = p_1^s + p_2^s$ $[r = 2]$

The tameness of Λ depends on arithmetical properties of $\log p_2 / \log p_1$ which influence $\mathcal{Z}$:= the set of poles on $\Re s = 1, s \neq 1$

Memoryless sources $\Lambda(s) = \dfrac{1}{1-\lambda(s)}$ with $\lambda(s) = p_1^s + p_2^s$ $[r=2]$

The tameness of Λ depends on arithmetical properties of $\log p_2/\log p_1$
which influence $\mathcal{Z}$:= the set of poles on $\Re s = 1, s \neq 1$

(i) $\mathcal{Z} \neq \emptyset \iff \log p_2/\log p_1$ is rational

(ii) If $\mathcal{Z} = \emptyset$, then the poles of $\Lambda(s)$ close to $\Re s = 1$
are created by good rational approximations of $\log p_2/\log p_1$

Memoryless sources $\Lambda(s) = \dfrac{1}{1-\lambda(s)}$ with $\lambda(s) = p_1^s + p_2^s$ $[r=2]$

The tameness of Λ depends on arithmetical properties of $\log p_2/\log p_1$
which influence $\mathcal{Z}$:= the set of poles on $\Re s = 1, s \neq 1$

(i) $\mathcal{Z} \neq \emptyset \iff \log p_2/\log p_1$ is rational

(ii) If $\mathcal{Z} = \emptyset$, then the poles of $\Lambda(s)$ close to $\Re s = 1$
are created by good rational approximations of $\log p_2/\log p_1$

The irrationality exponent $\mu(x)$ of a number x equals μ if, for any $\nu > \mu$,
the set of pairs $(a,b) \in \mathbb{Z}^2$ for which $\left|x - \dfrac{a}{b}\right| \leq \dfrac{1}{b^\nu}$ is finite

x diophantine $\iff \mu(x) < \infty$

Memoryless sources $\Lambda(s) = \dfrac{1}{1-\lambda(s)}$ with $\lambda(s) = p_1^s + p_2^s$ $[r=2]$

The tameness of Λ depends on arithmetical properties of $\log p_2/\log p_1$
which influence $\mathcal{Z}$:= the set of poles on $\Re s = 1, s \neq 1$

(i) $\mathcal{Z} \neq \emptyset \iff \log p_2/\log p_1$ is rational

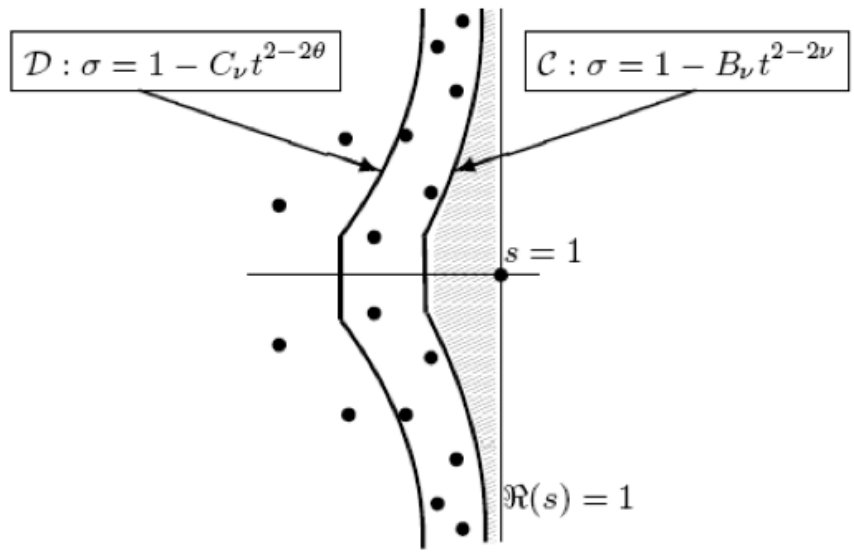
(ii) If $\mathcal{Z} = \emptyset$, then the poles of $\Lambda(s)$ close to $\Re s = 1$
are created by good rational approximations of $\log p_2/\log p_1$

The irrationality exponent $\mu(x)$ of a number x equals μ if, for any $\nu > \mu$,
the set of pairs $(a,b) \in \mathbb{Z}^2$ for which $\left|x - \dfrac{a}{b}\right| \leq \dfrac{1}{b^\nu}$ is finite

x diophantine $\iff \mu(x) < \infty$

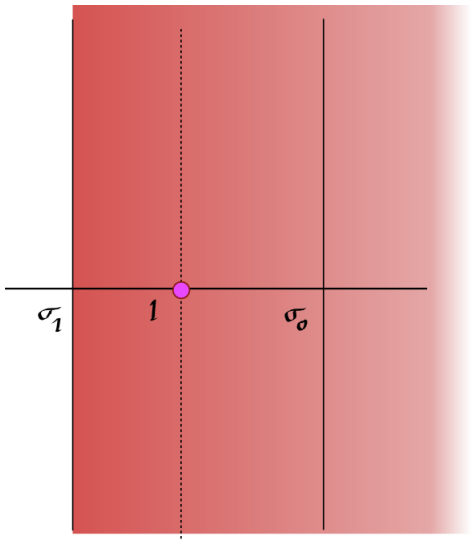
The shape of the tameness region
is related to $\mu(\log p_2/\log p_1)$.

If $\mu(\log p_2/\log p_1) = \mu$
then, for any θ, ν with $\theta < \mu < \nu$,
the tameness region is as shown:
[Flajolet-Roux-V. 2010]



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

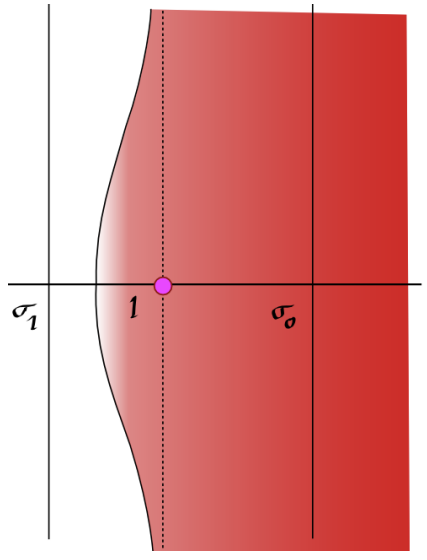
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

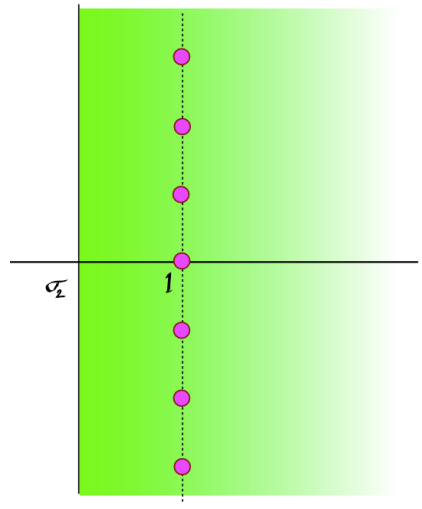
Geometric condition



Situation 2

Hyperbolic region

Arithmetic condition



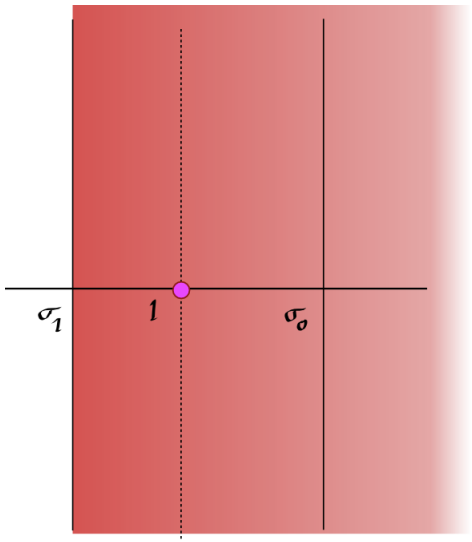
Situation 3

Vertical strip with holes

Periodicity condition

For which general dynamical sources do these different situations occur?

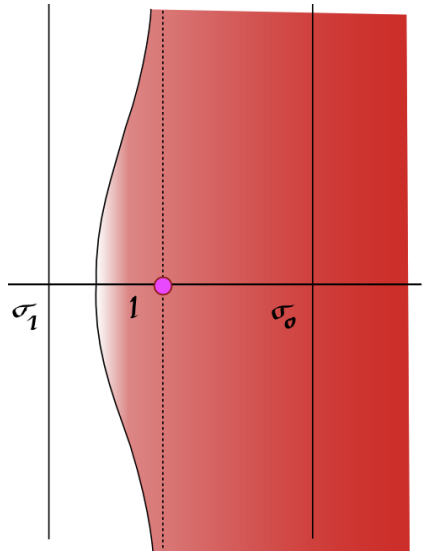
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

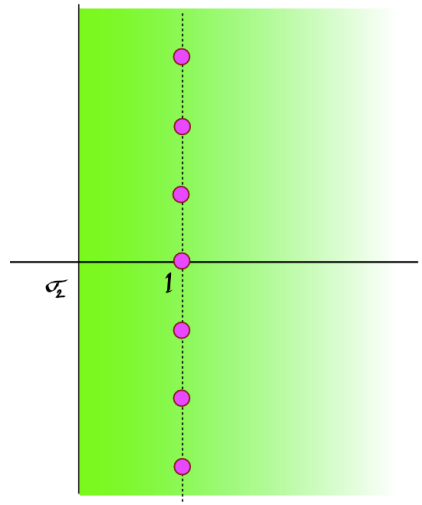
Geometric condition



Situation 2

Hyperbolic region

Arithmetic condition



Situation 3

Vertical strip with holes

Periodicity condition

For which general dynamical sources do these different situations occur?

– S1 occurs when "the branches are not too often of the same shape".

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

– S2 occurs if a extension of the following condition holds:

"there exists a ratio $\log p_i / \log p_j$ which is "diophantine"

Situation 1- Existence of a vertical strip where $\Lambda(s)$ is tame

The condition UNI expresses that

"the branches of the dynamical system are not too often of the same shape"

Situation 1- Existence of a vertical strip where $\Lambda(s)$ is tame

The condition UNI expresses that

"the branches of the dynamical system are not too often of the same shape"

Theorem [Dolgopyat-Baladi-Cesaratto-V].
For a good dynamical system which satisfies the condition UNI,
there exists a vertical strip where $\Lambda(s)$ is tame.

Situation 1- Existence of a vertical strip where $\Lambda(s)$ is tame

The condition UNI expresses that

"the branches of the dynamical system are not too often of the same shape"

> Theorem [Dolgopyat-Baladi-Cesaratto-V].
> For a good dynamical system which satisfies the condition UNI,
> there exists a vertical strip where $\Lambda(s)$ 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.

Situation 2- 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"

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

The derivatives $|h'(h^\star)|$ replace the probabilities of the memoryless case.

Situation 2- 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"

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

The derivatives $|h'(h^\star)|$ replace the probabilities of the memoryless case.

DIOP: There exists a ratio $c(h,k) := \dfrac{\log |h'(h^\star)|}{\log |k'(k^\star)|}$ which is diophantine.

Situation 2- 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"

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

The derivatives $|h'(h^\star)|$ replace the probabilities of the memoryless case.

DIOP: There exists a ratio $c(h,k) := \dfrac{\log |h'(h^\star)|}{\log |k'(k^\star)|}$ which is diophantine.

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

Situation 2- 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"

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

The derivatives $|h'(h^\star)|$ replace the probabilities of the memoryless case.

DIOP: There exists a ratio $c(h,k) := \dfrac{\log |h'(h^\star)|}{\log |k'(k^\star)|}$ which is diophantine.

Theorem [Dolgopyat-Roux-V.]
For a good dynamical system 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: Dirichlet generating functions and ttameness

– An important class of sources: dynamical sources.

– Tameness of dynamical sources

– Conclusion and possible extensions.

Conclusions.

Description of a framework which

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

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

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

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

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

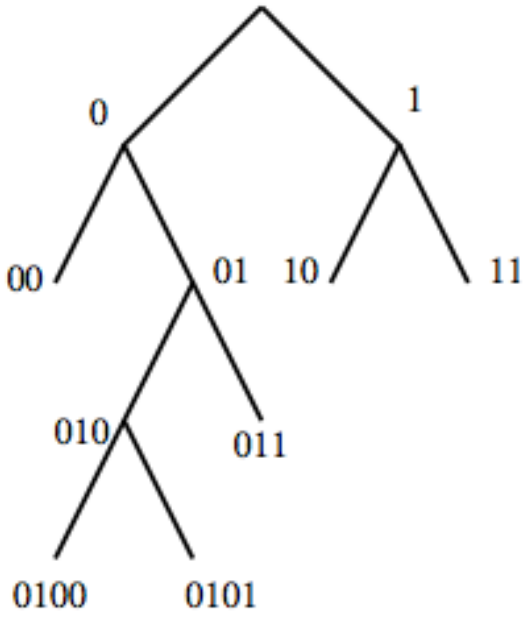
Possible extensions and work in progress

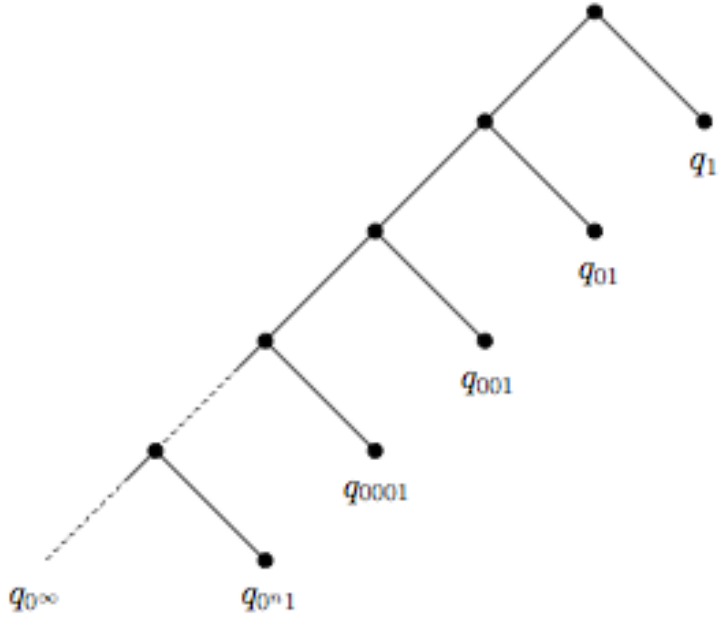
I– Classification of sources

– 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 ?

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?