INFORMATION THEORY: SOURCES and DIRICHLET SERIES, TAMENESS of DYNAMICAL SOURCES

Brigitte Vallée

GREYC Laboratory

(CNRS and University of Caen, France)

INFORMATION THEORY: SOURCES and DIRICHLET SERIES, TAMENESS of DYNAMICAL SOURCES

Brigitte Vallée

GREYC Laboratory

(CNRS and University of Caen, France)

Talk partly based on joint works with

Viviane Baladi, Eda Cesaratto,

Julien Clément, Jim Fill, Philippe Flajolet, Mathieu Roux

INFORMATION THEORY: SOURCES and DIRICHLET SERIES, TAMENESS of DYNAMICAL SOURCES

Brigitte Vallée

GREYC Laboratory

(CNRS and University of Caen, France)

Talk partly based on joint works with

Viviane Baladi, Eda Cesaratto,

Julien Clément, Jim Fill, Philippe Flajolet, Mathieu Roux

Dedicated to Philippe

Plan of the talk.

Part I

– describes a general model for sources

– shows the importance of the Dirichlet generating functions

– explains the importance of tameness in the analyses of text algorithms

Plan of the talk.

Part I

– describes a general model for sources

– shows the importance of the Dirichlet generating functions

– explains the importance of tameness in the analyses of text algorithms

Part II

– defines a natural subclass of sources, the dynamical sources

Plan of the talk.

Part I

– describes a general model for sources

– shows the importance of the Dirichlet generating functions

– explains the importance of tameness in the analyses of text algorithms

Part II

– defines a natural subclass of sources, the dynamical sources

Part III

– provides sufficient conditions for tameness of dynamical sources

Part I

– describes a general model for sources

– shows the importance of the Dirichlet generating functions

– explains the importance of tameness in the analyses of text algorithms

Sources.

A source:= a mechanism which produces symbols from alphabet Σ,
one symbol for each time unit.
When (discrete) time evolves, a source produces (infinite) words of $\Sigma^{\mathbb{N}}$.

Sources.

A source:= a mechanism which produces symbols from alphabet Σ,
one symbol 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 $\mathcal{S}$.

Sources.

A source:= a mechanism which produces symbols from alphabet Σ,
one symbol 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

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

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

The Dirichlet generating functions of the source

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

encapsulate the main probabilistic properties of the source and translate them into analytic properties. Two instances:

The Dirichlet generating functions of the source

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

encapsulate the main probabilistic properties of the source and translate them into analytic properties. Two instances:

– the entropy $h_{\mathcal{S}}$,

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

– the coincidence $c_{\mathcal{S}}$

$c_{\mathcal{S}}(A, B) :=$ the length of the longest common prefix of A and B

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

The Dirichlet generating functions of the source

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

encapsulate the main probabilistic properties of the source and translate them into analytic properties. Two instances:

– the entropy $h_{\mathcal{S}}$,

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

– the coincidence $c_{\mathcal{S}}$

$c_{\mathcal{S}}(A, B) :=$ the length of the longest common prefix of A and B

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

– they intervene in the realistic analyses of sorting and searching algorithms

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

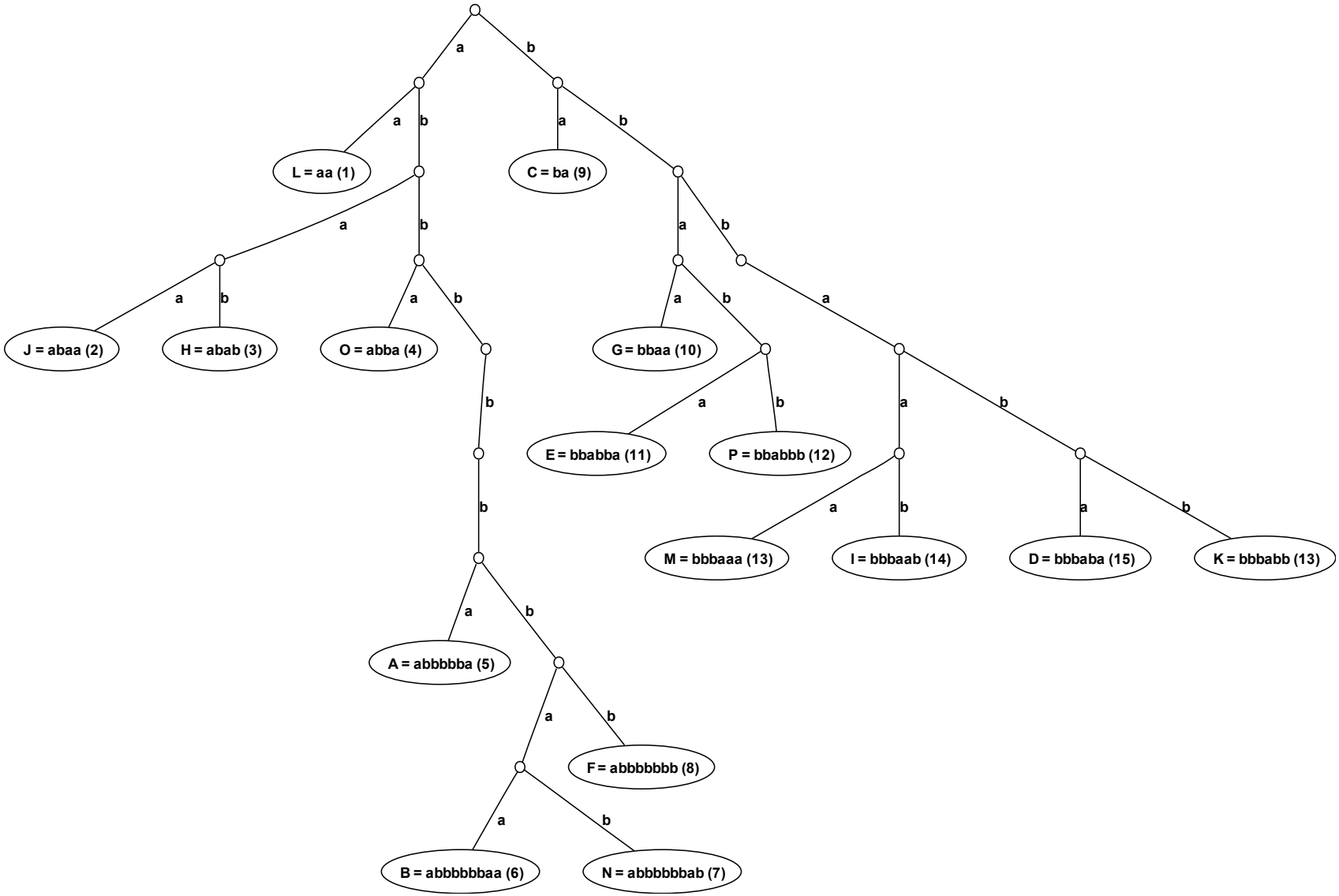
F = abbbbbbbbabb G = bbaabbabbaba H = ababbbabbbab I = bbbaabbbbbbb J = abaabbbbaabb

K = bbbabbbbbbaa L = aaaabbabaaba M = bbbaaabbbbbb N = abbbbbbabbaa O = abbababbbb P = bbabbbaaaabb

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 = abbababbbbb 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

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

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

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

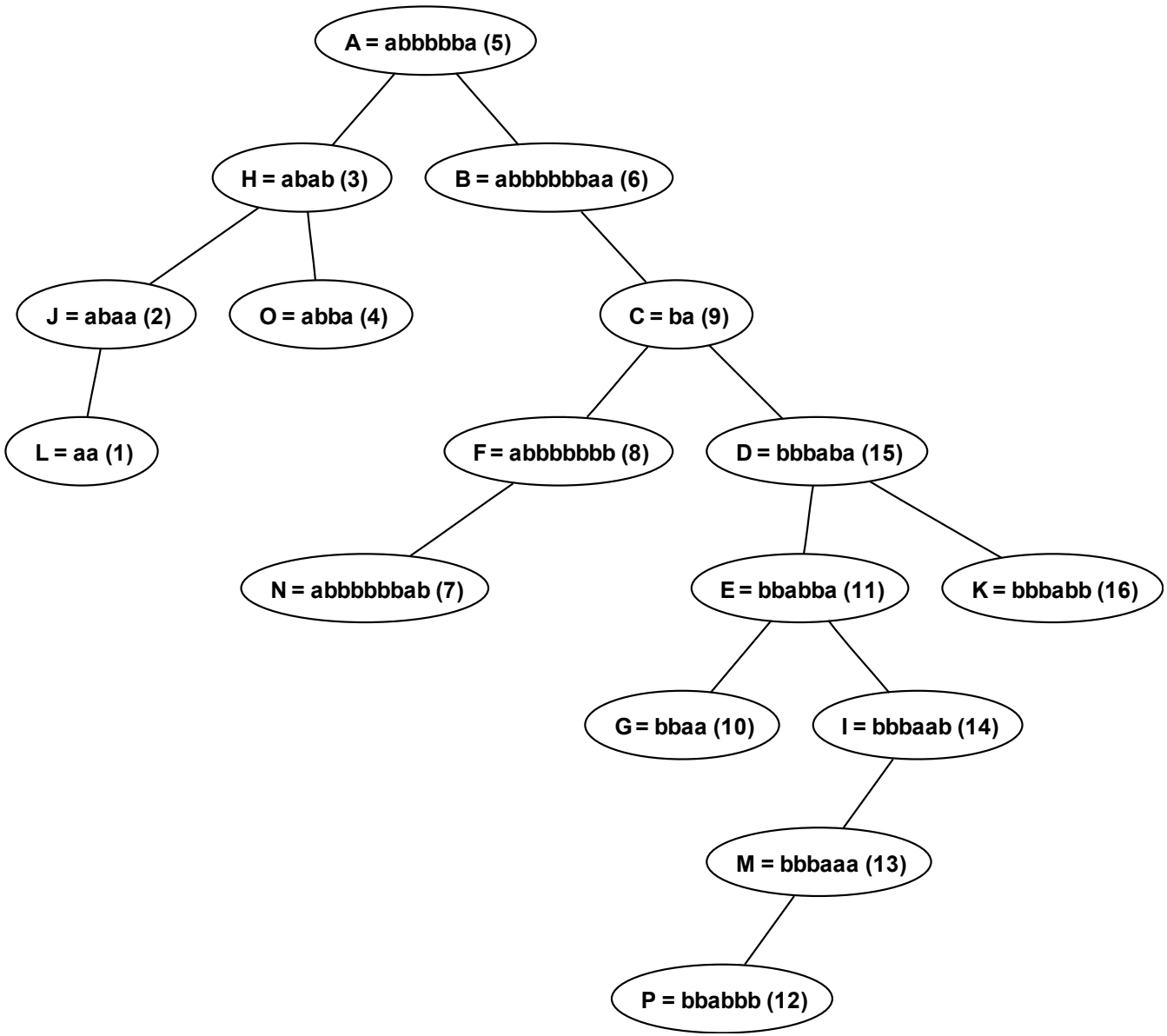
The size, and the path length of a trie equal

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

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

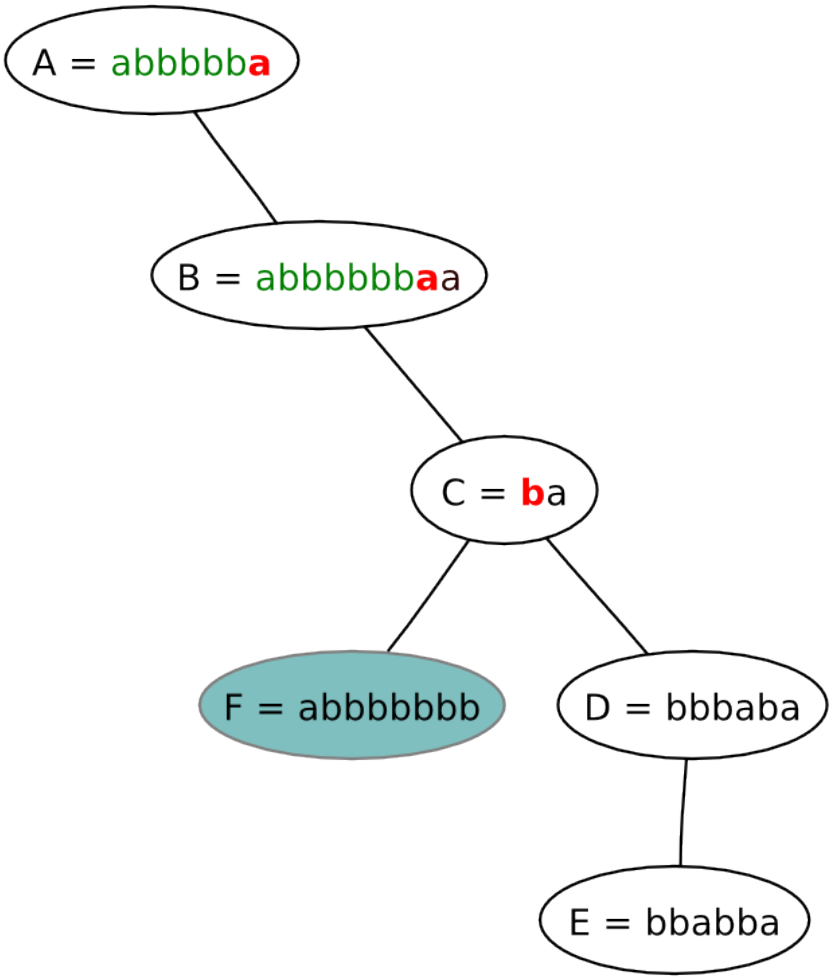
Realistic probabilistic study of the BST (binary search tree) built on words

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 = abbababbbbb P = bbabbbaaaabb

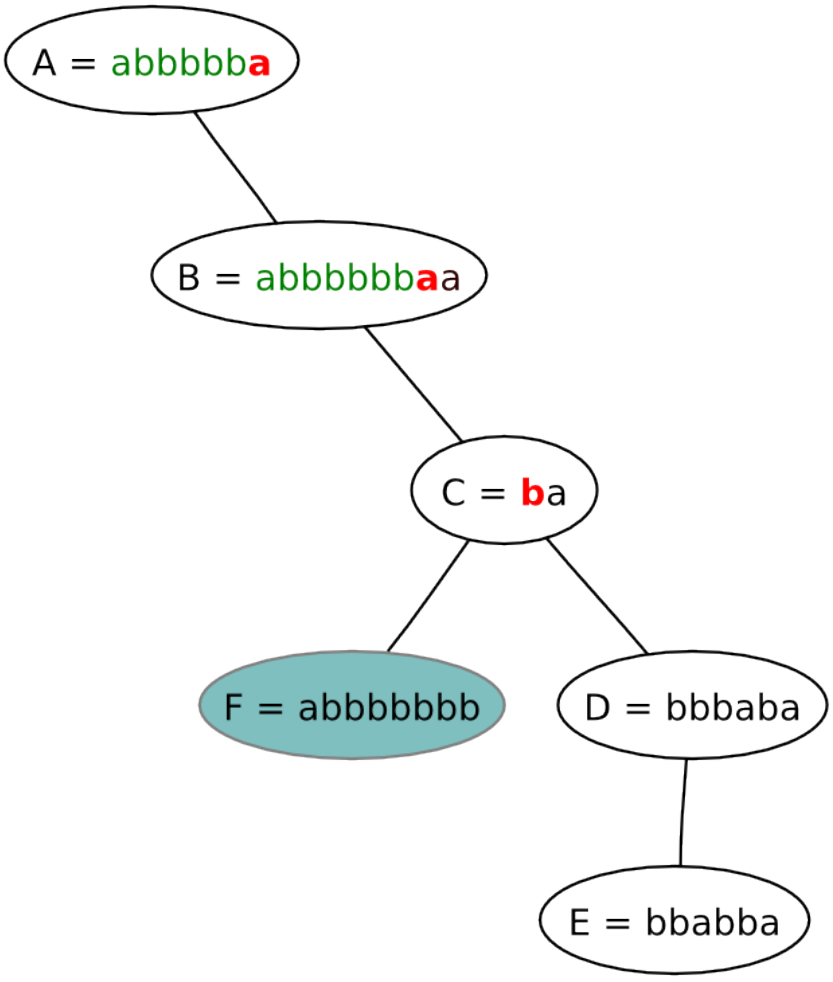


An example : The realistic cost of the insertion of a key into the BST

An example : The realistic cost of the insertion of a key into the BST



An example : The realistic cost of the insertion of a key into the BST

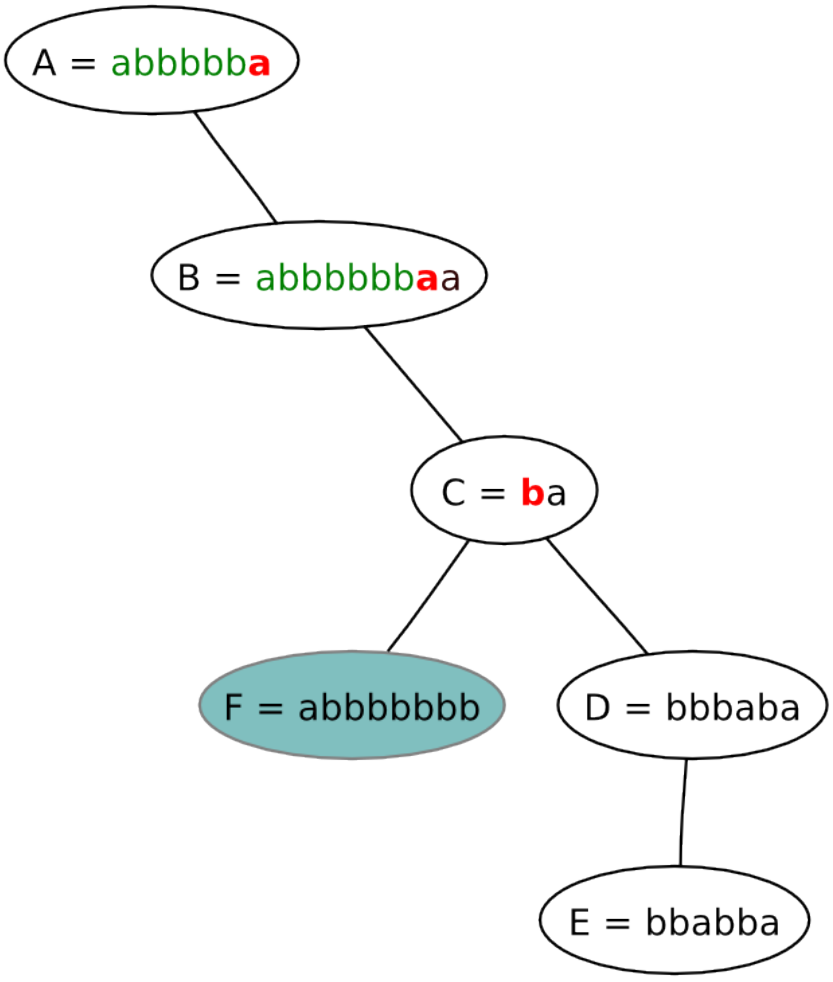


The realistic cost of the comparison between two words A and B is related to their coincidence $c(A, B)$

a b a b b b...
a b a a b a...

coincidence=3; #comparisons=4.

An example : The realistic cost of the insertion of a key into the BST



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

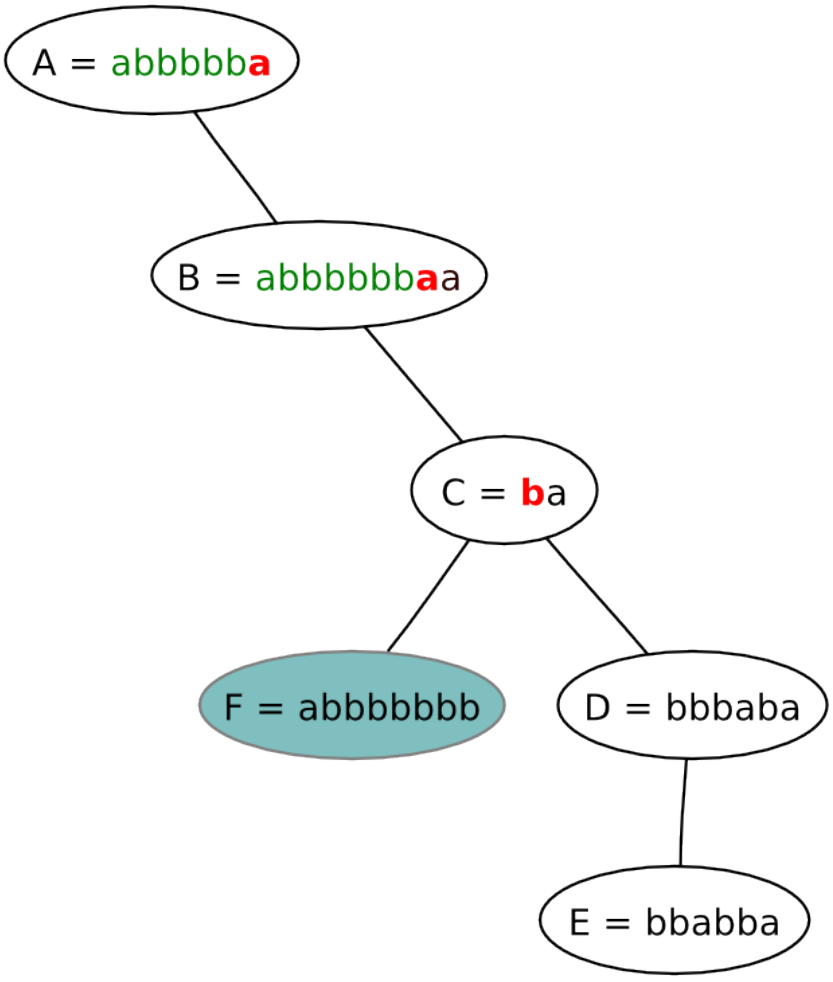
a b a b b b...
a b a a b a...

coincidence=3; #comparisons=4.

The coincidence satisfies $c(A, B) \geq k$ iff A and B begin with the same prefix of length k

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

An example : The realistic cost of the insertion of a key into the BST



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

a b a b b b...
a b a a b a...

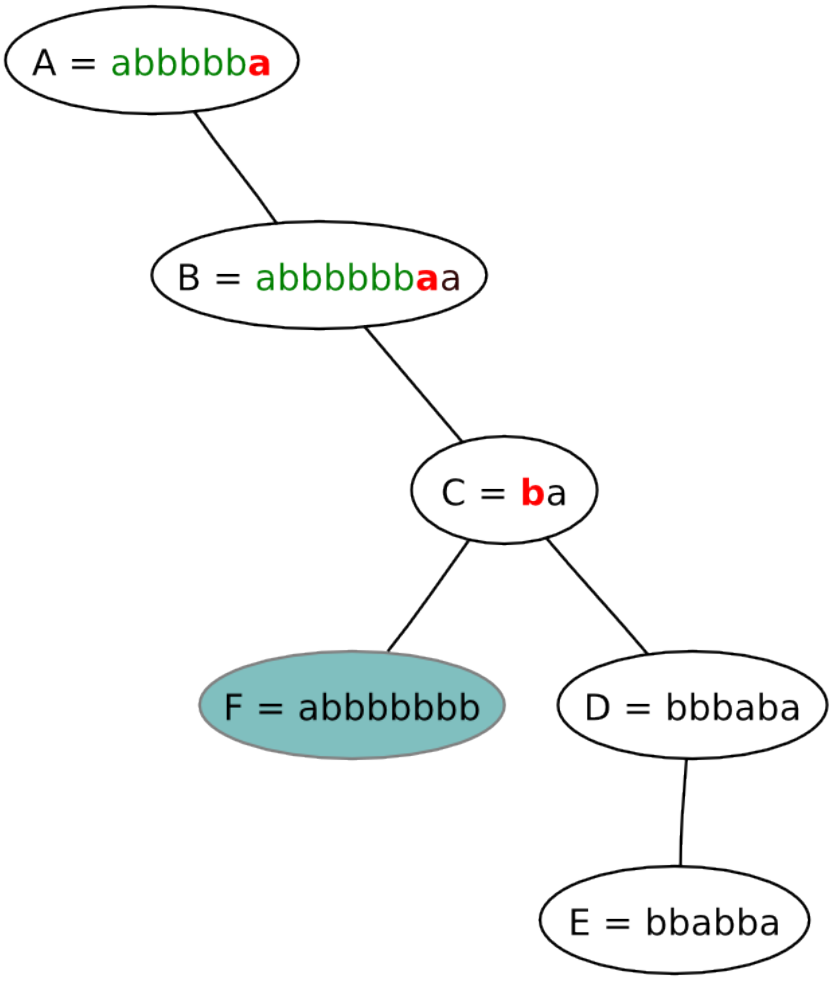
coincidence=3; #comparisons=4.

The coincidence satisfies $c(A, B) \geq k$ iff A and B begin with the same prefix of length k

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

Number of symbol comparisons needed for the insertion of F? $= 16$

An example : The realistic cost of the insertion of a key into the BST



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

a b a b b b...
a b a a b a...

coincidence=3; #comparisons=4.

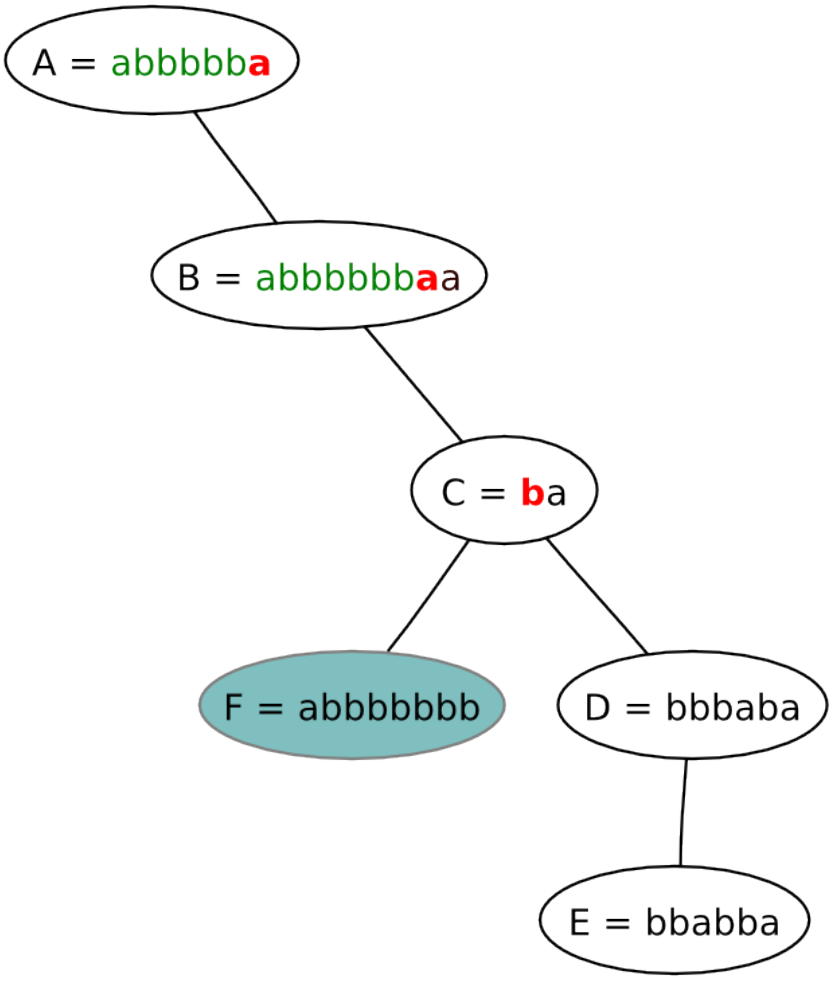
The coincidence satisfies $c(A, B) \geq k$ iff A and B begin with the same prefix of length k

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

Number of symbol comparisons needed for the insertion of F? $= 16$

$= 7$ for comparing to A

An example : The realistic cost of the insertion of a key into the BST



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

a b a b b b...
a b a a b a...

coincidence=3; #comparisons=4.

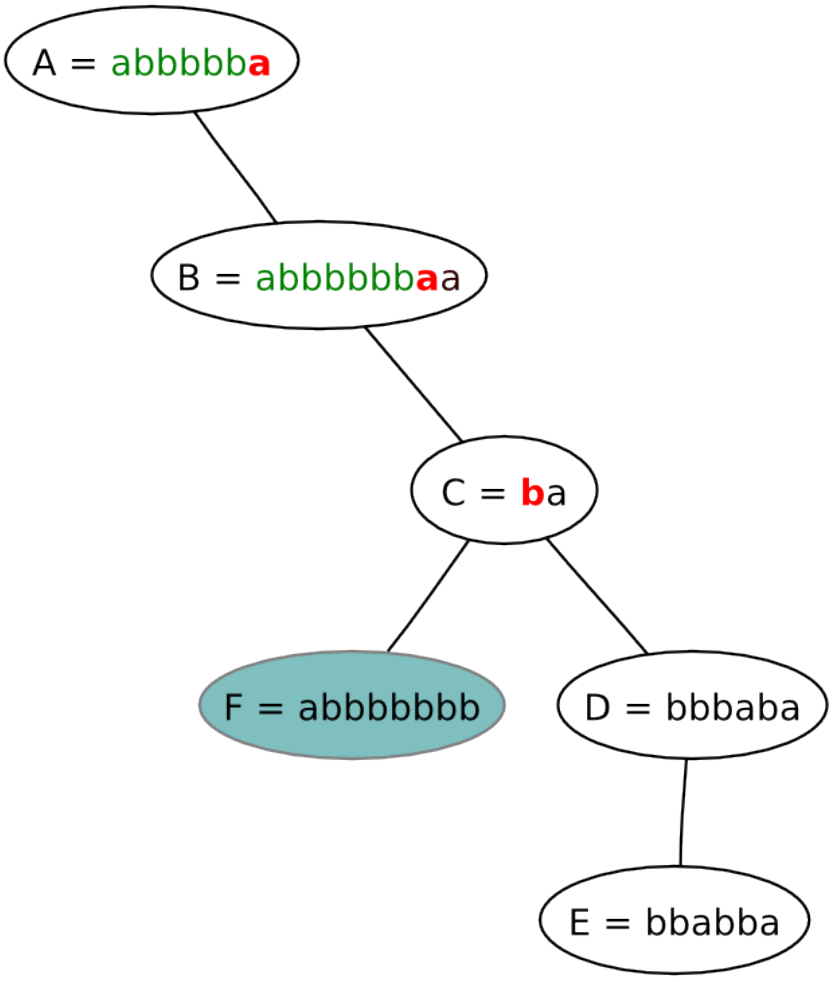
The coincidence satisfies $c(A, B) \geq k$ iff A and B begin with the same prefix of length k

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

Number of symbol comparisons needed for the insertion of F? $= 16$

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

An example : The realistic cost of the insertion of a key into the BST



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

a b a b b b...
a b a a b a...

coincidence=3; #comparisons=4.

The coincidence satisfies $c(A, B) \geq k$ iff A and B begin with the same prefix of length k

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

Number of symbol comparisons needed for the insertion of F? $= 16$

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

$+ 1$ for comparing to C

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.
[CIFIVa 2001] for Tries, [CIFiFIVa 2009] for symbol-BST'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.
[CIFIVa 2001] for Tries, [CIFiFIVa 2009] for symbol-BST's

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.
[CIFIVa 2001] for Tries, [CIFiFIVa 2009] for symbol-BST's

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 formula, not easy to deal with, due to the alternating signs

Asymptotic analysis.

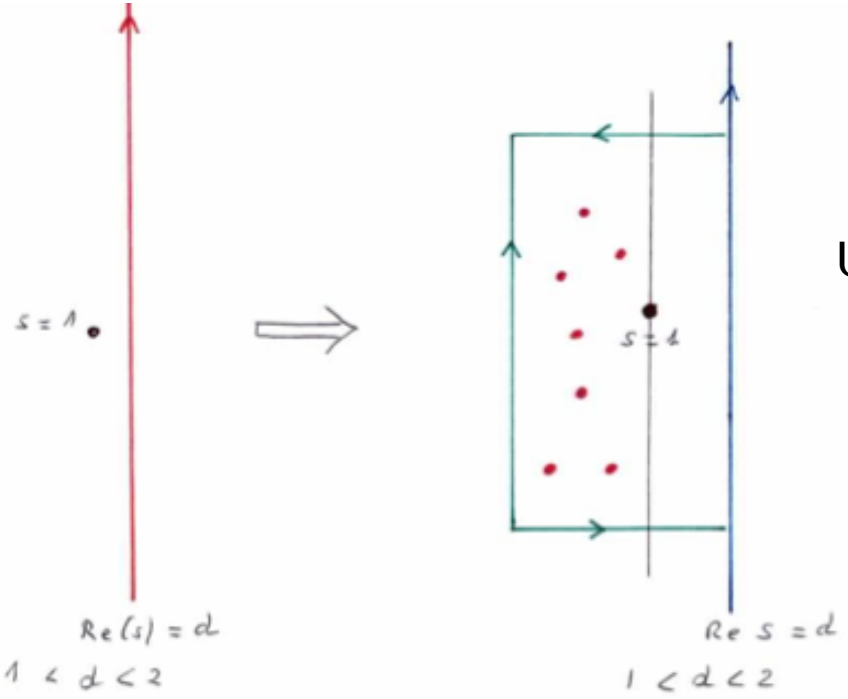
The residue formula transforms the sum into an integral with $1 < d < 2$.

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

Asymptotic analysis.

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

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



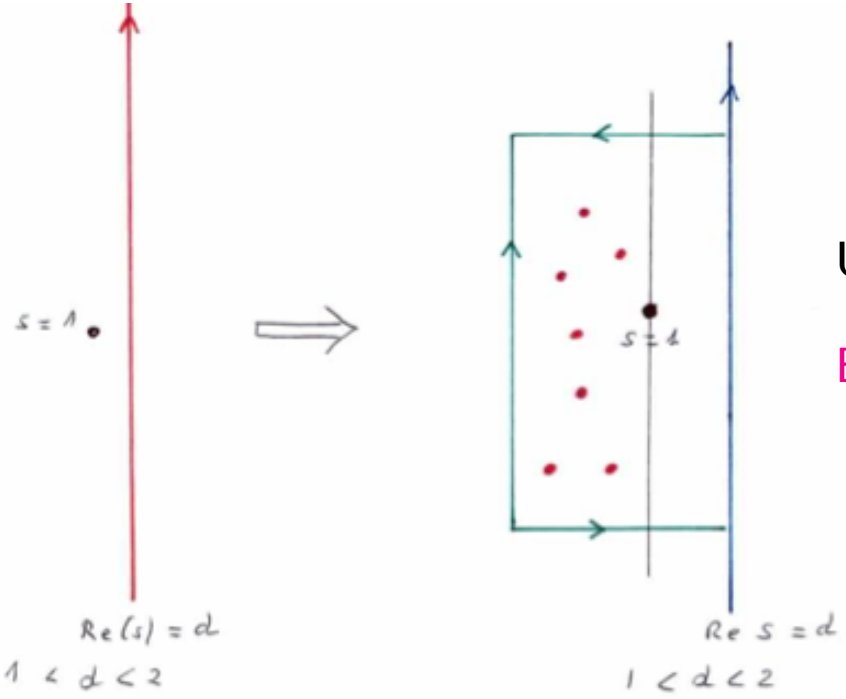
We shift the integral on the left,

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,

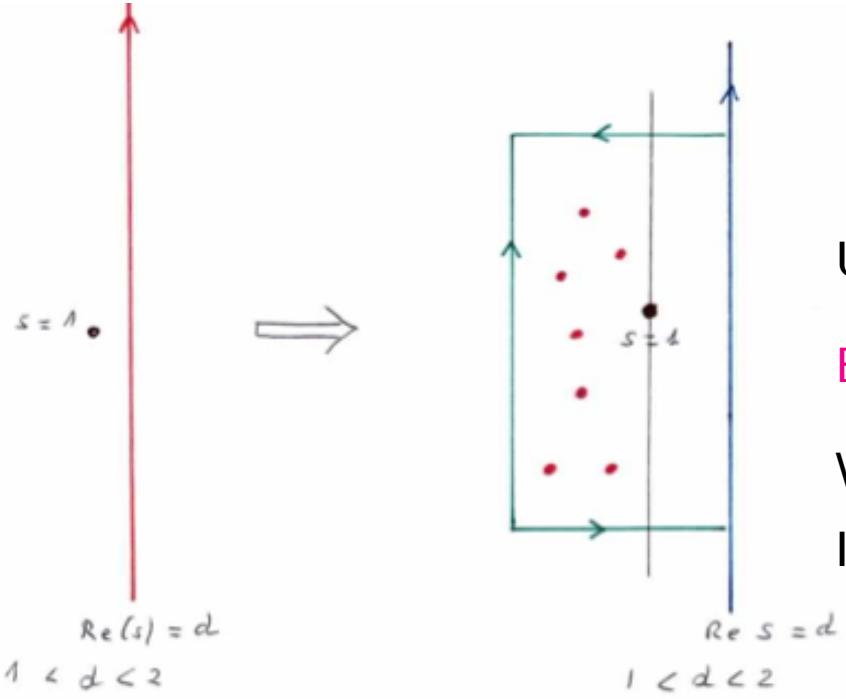
Usually, the first singularities occur at $\Re s = 1$.

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 $\Re s = 1$?

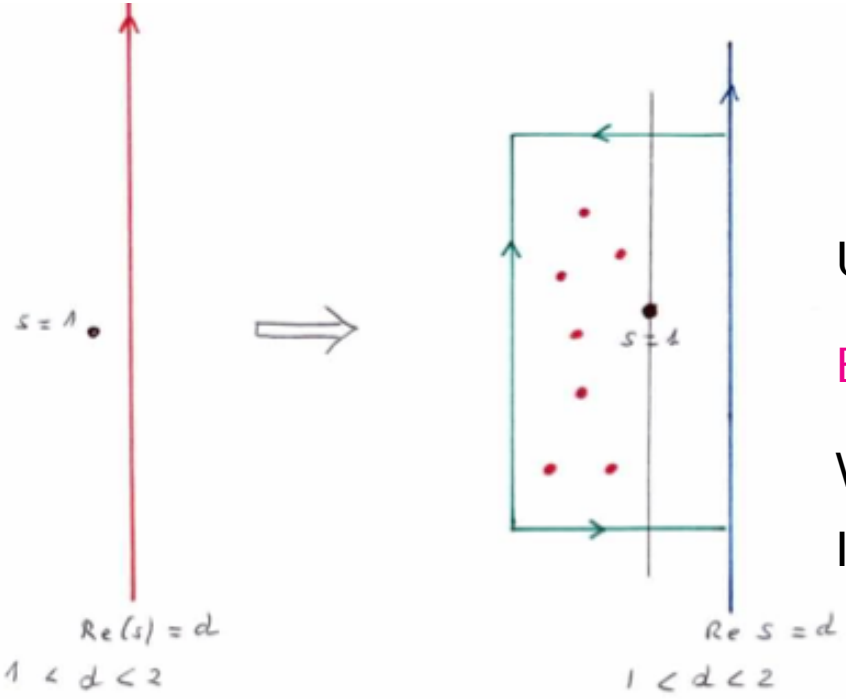
Where are the red singularities closest to $\Re s = 1$?

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

Part II

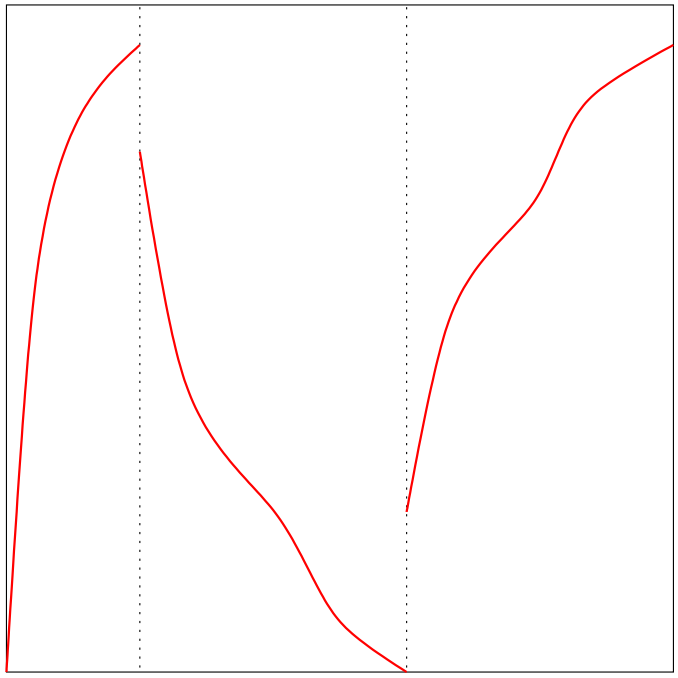
– defines a natural subclass of sources, the dynamical sources

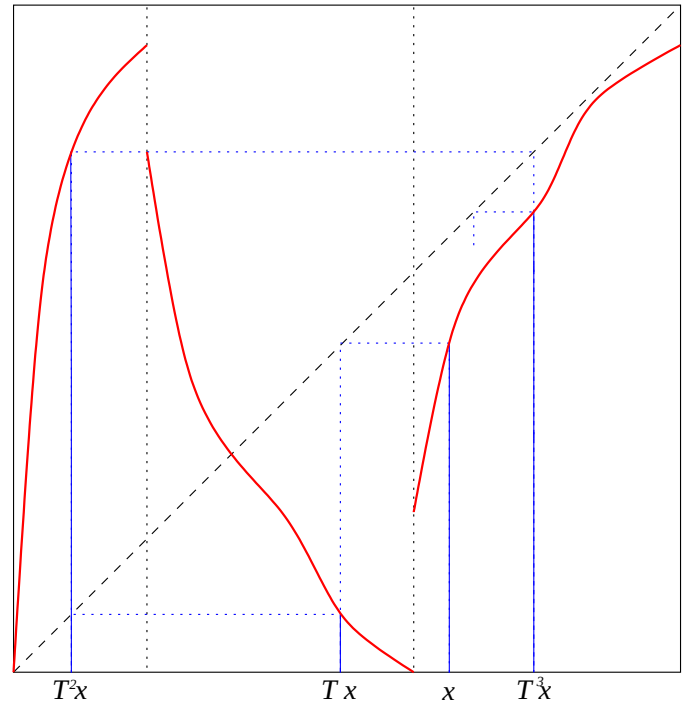
A general class of "natural" sources: dynamical sources associated to dynamical systems [V. 2001]

With a shift map $T : \mathcal{I} \to \mathcal{I}$ and an encoding map $\sigma : \mathcal{I} \to \Sigma$,
the emitted word is $M(x) = (\sigma x, \sigma T x, \sigma T^2 x, \ldots \sigma T^k x, \ldots)$
namely, the encoded trajectory of x

A general class of "natural" sources: dynamical sources associated to dynamical systems [V. 2001]

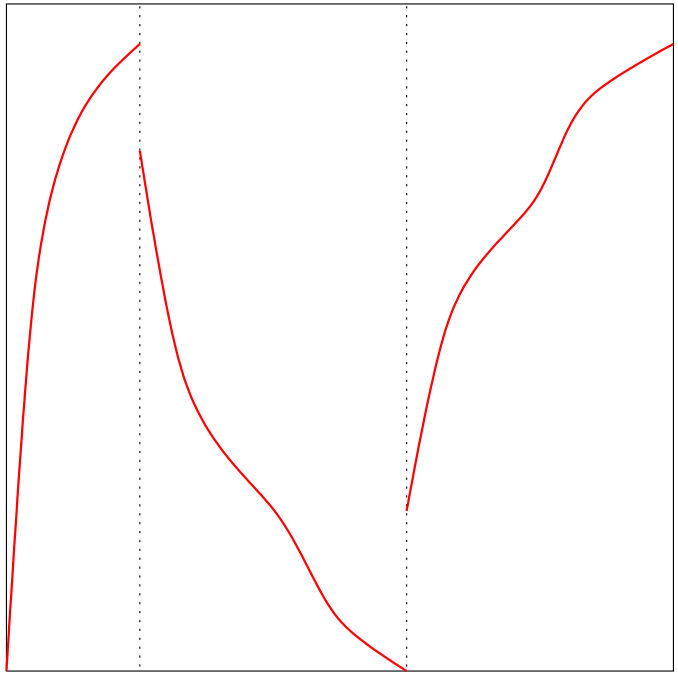
With a shift map $T : \mathcal{I} \to \mathcal{I}$ and an encoding map $\sigma : \mathcal{I} \to \Sigma$,
the emitted word is $M(x) = (\sigma x, \sigma T x, \sigma T^2 x, \ldots \sigma T^k x, \ldots)$
namely, the encoded trajectory of x

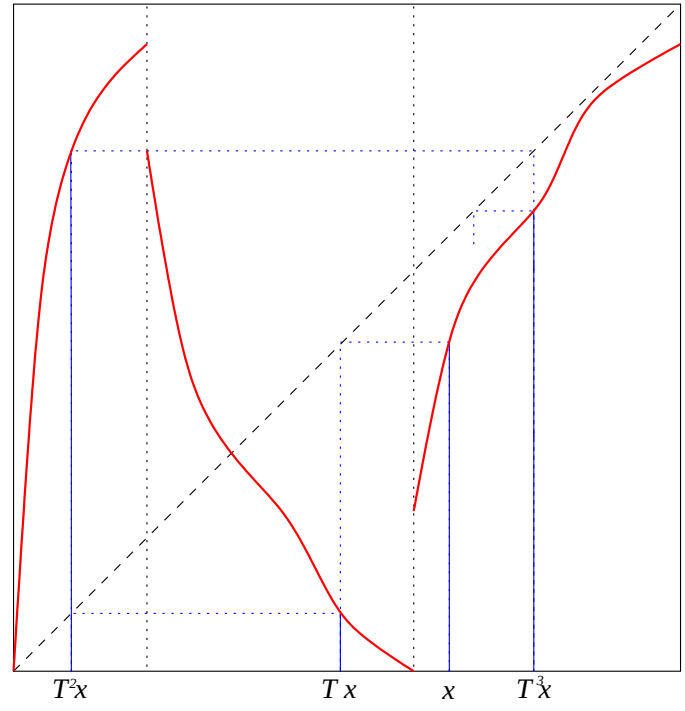




A general class of "natural" sources: dynamical sources associated to dynamical systems [V. 2001]

With a shift map $T : \mathcal{I} \to \mathcal{I}$ and an encoding map $\sigma : \mathcal{I} \to \Sigma$,
the emitted word is $M(x) = (\sigma x, \sigma Tx, \sigma T^2 x, \ldots \sigma T^k x, \ldots)$
namely, the encoded trajectory of x





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

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

A dynamical system $(\mathcal{I}, S)$ is defined by four elements:

– a finite alphabet Σ,

– 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

s.t. $T|_{\mathcal{I}_m}$ 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.

This gives rise to a source: on an input x of $\mathcal{I}$, it outputs the word

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

When an initial density –and an initial distribution F– is chosen on $\mathcal{I}$,

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

$=$ a dynamical source $\mathcal{S}_F$.

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.

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 $T(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$;
it describes the set $s(m)$ of possible successors of the symbol m.

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 $T(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$;
it describes the set $s(m)$ of possible successors of the symbol m.
Particular cases: – Complete systems $T(\mathcal{I}_m) = \mathcal{I}$
– Markovian systems $T(\mathcal{I}_m) = \text{union of some } \mathcal{I}_\ell$

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 $T(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$;
it describes the set $s(m)$ of possible successors of the symbol m.
Particular cases: – Complete systems $T(\mathcal{I}_m) = \mathcal{I}$
– Markovian systems $T(\mathcal{I}_m) = \text{union of some } \mathcal{I}_\ell$

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

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 $T(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$;
it describes the set $s(m)$ of possible successors of the symbol m.
Particular cases: – Complete systems $T(\mathcal{I}_m) = \mathcal{I}$
– Markovian systems $T(\mathcal{I}_m) = \text{union of some } \mathcal{I}_\ell$

– 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.

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

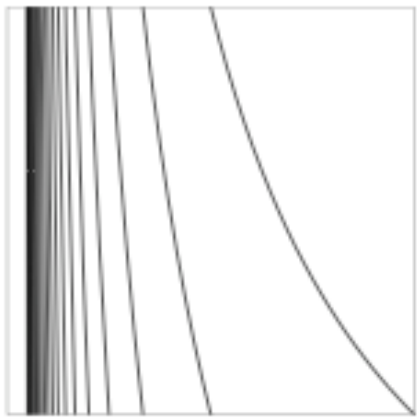
Correlations between symbols are mainly due to two geometric characteristics:
the position of the branches and the shape of the branches.

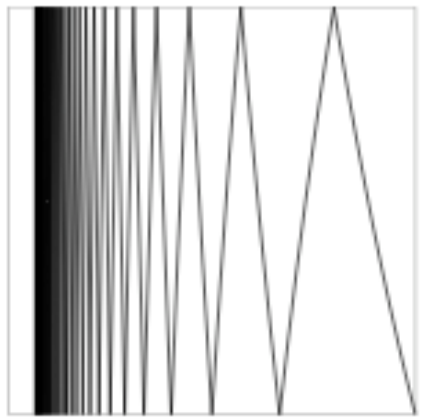
– the position of the branches: the position of $T(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$;
it describes the set $s(m)$ of possible successors of the symbol m.
Particular cases: – Complete systems $T(\mathcal{I}_m) = \mathcal{I}$
– Markovian systems $T(\mathcal{I}_m) = \text{union of some } \mathcal{I}_\ell$

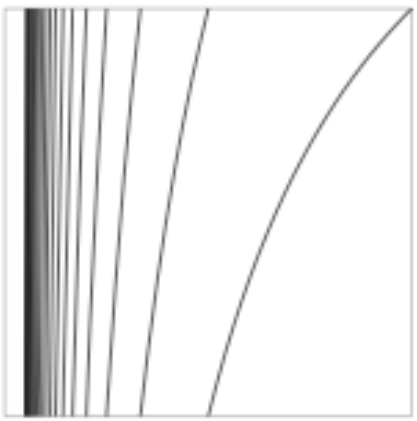
– the shape of the branches, is described by their derivatives;
it explains how the distribution evolves.
Less correlated systems correspond to systems with affine branches.

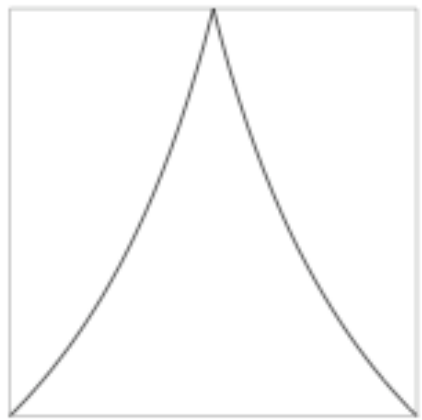
Generally speaking, importance of expansiveness:
the derivative T' satisfies $\forall x \in \mathcal{I} \quad |T'(x)| \geq \delta > 1$.

Four Euclidean dynamical sources, Two different behaviours for trajectories.

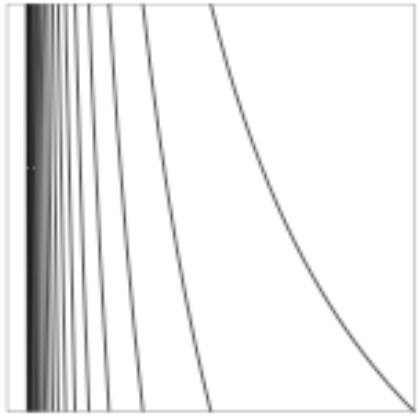


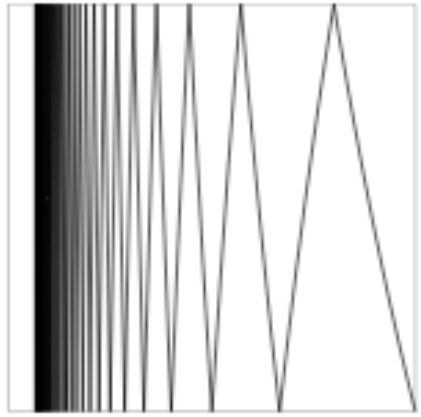


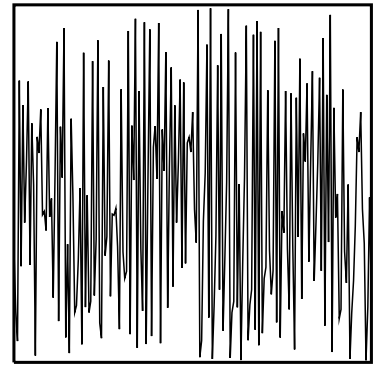


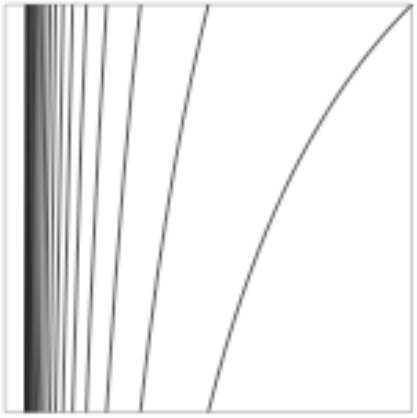


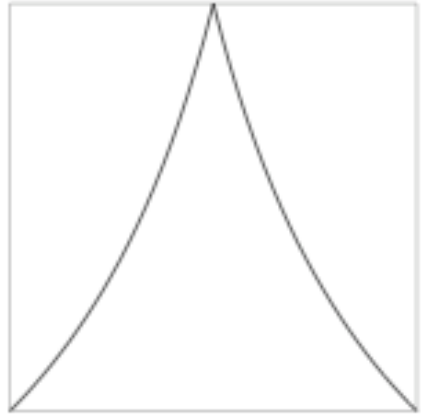
Four Euclidean dynamical sources, Two different behaviours for trajectories.



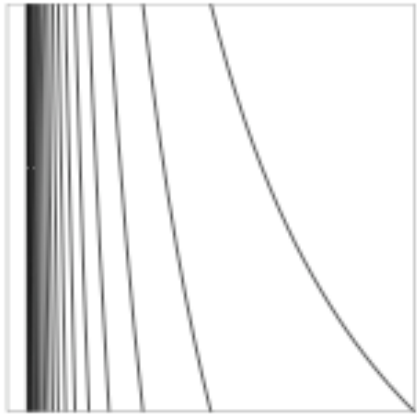


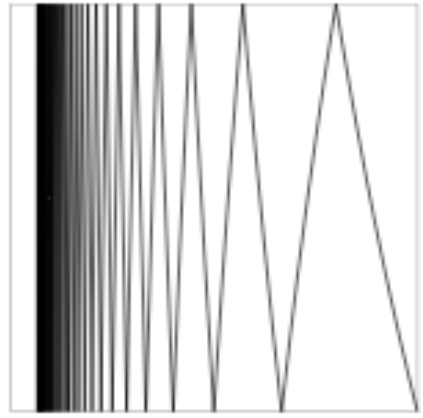


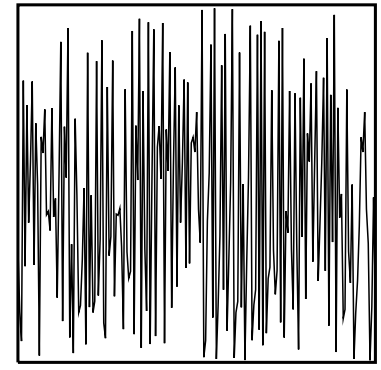


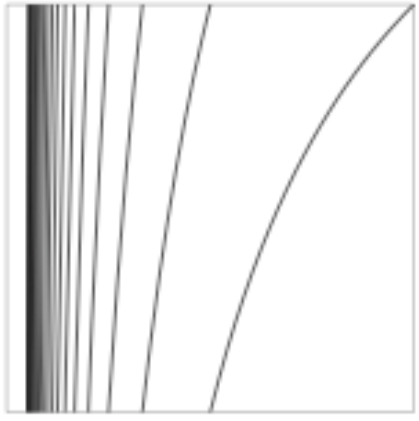


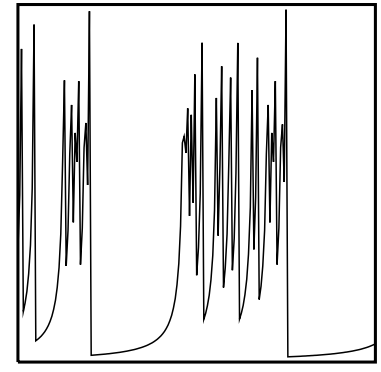
Four Euclidean dynamical sources, Two different behaviours for trajectories.











Particular cases: simple sources and affine branches

Particular cases: simple sources and affine branches

A memoryless source

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

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 $\mathcal{I}_m$.

Particular cases: simple sources and affine branches

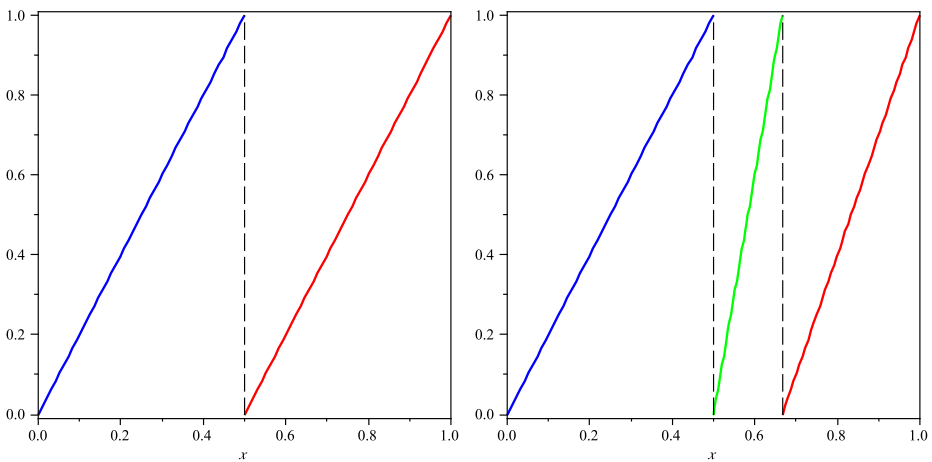
A memoryless source

:= 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 $\mathcal{I}_m$.



Two memoryless sources

Particular cases: simple sources and affine branches

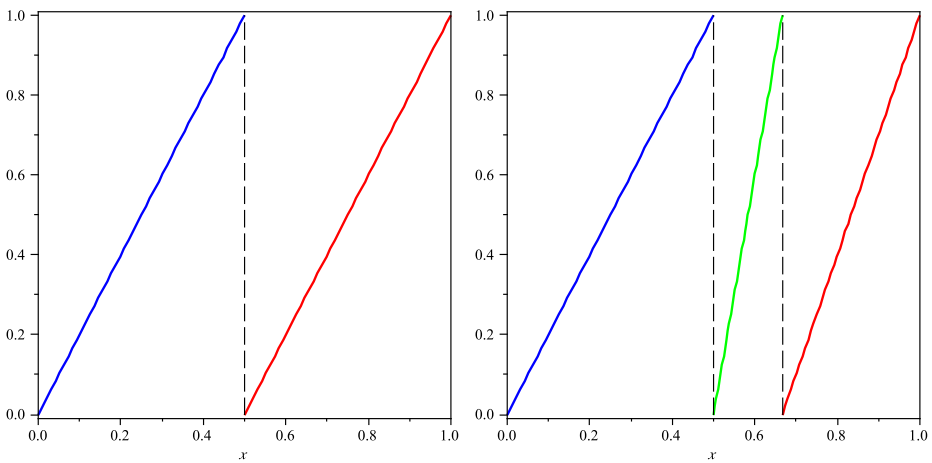
A memoryless source

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

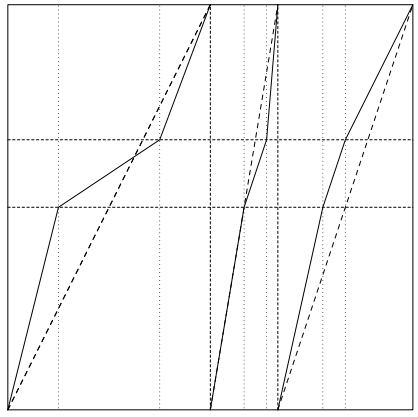
A Markov chain

:= a Markovian system with affine branches,

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



Two memoryless sources



a Markov chain.

General case of interest: the Good Class

A complete –or a Markovian– system

– with a possible infinite denumerable alphabet

– expansive.

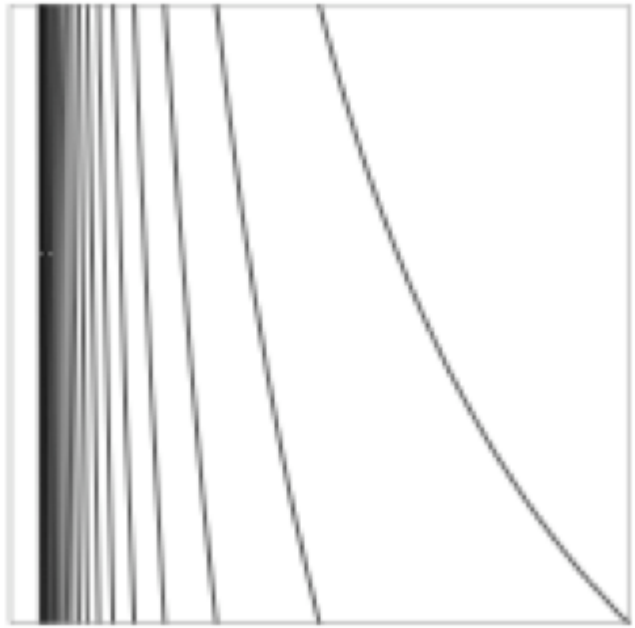
General case of interest: the Good Class

A complete –or a Markovian– system

– with a possible infinite denumerable alphabet

– expansive.

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



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

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

Memoryless sources, with probabilities (p_i)

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

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

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 $P := (p_{i,j})$

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

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

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 $P := (p_{i,j})$

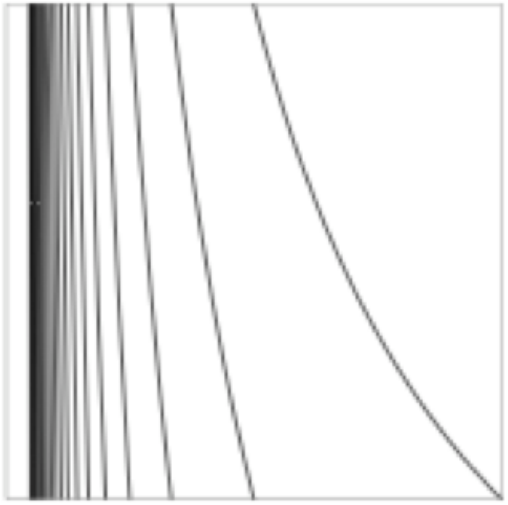
$$\Lambda(s) = 1 + {}^t\mathbf{1}(I - P(s))^{-1} R(s) \qquad \text{with} \quad P(s) = (p_{i,j}^s), \quad R(s) = (r_i^s).$$

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

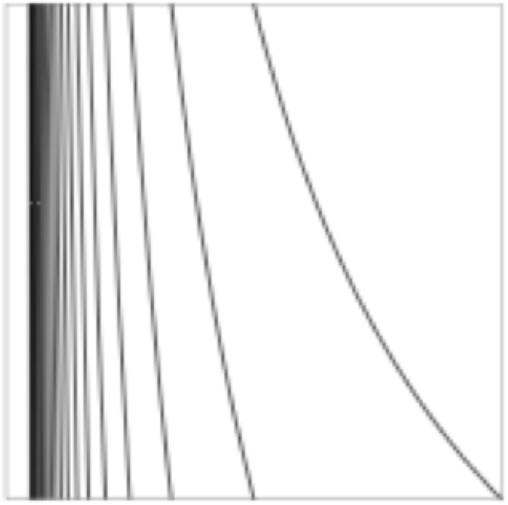
is the density transformer of the dynamical system.

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.

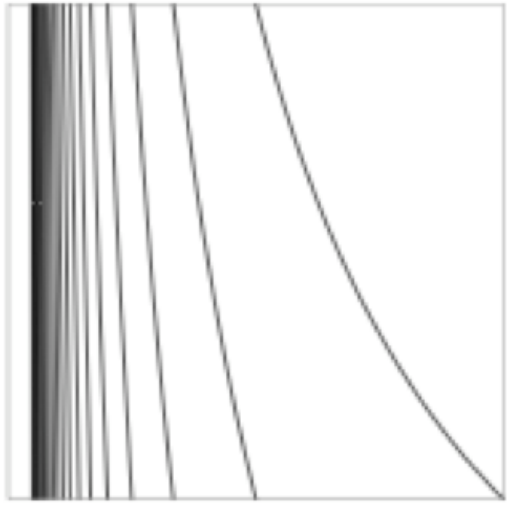
For a density f on $[0,1]$,

$\mathbf{H}[f]$ 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 $s = 1$

Part III

– provides sufficient conditions for tameness of dynamical sources

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 and of polynomial growth when $\Im s \to \infty$

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 and of polynomial growth when $\Im s \to \infty$

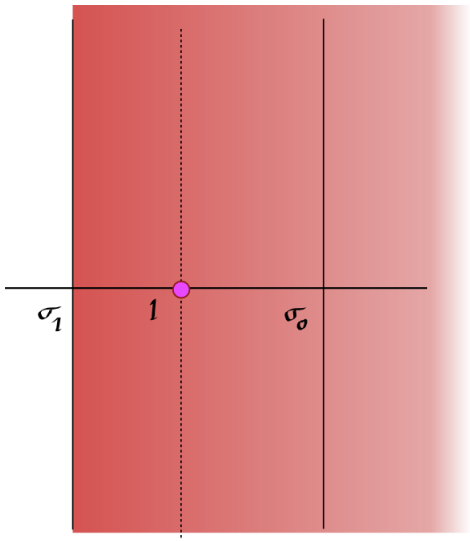
Different possible regions $\mathcal{R}$ on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.

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 and of polynomial growth when $\Im s \to \infty$

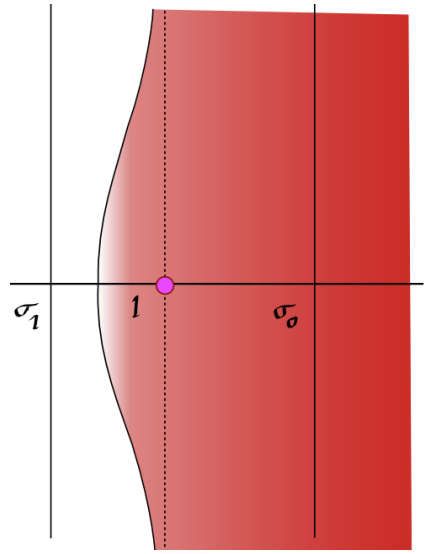
Different possible regions $\mathcal{R}$ on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

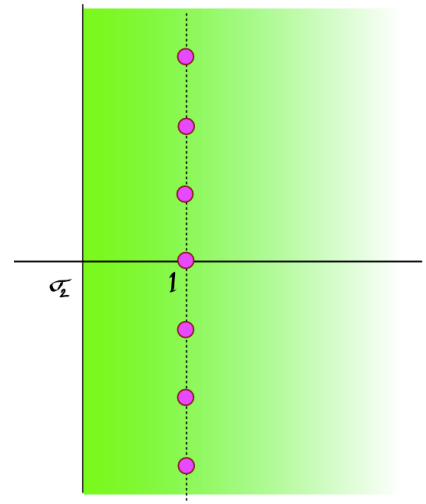
$1 - \sigma \leq a$



Situation 2

Hyperbolic region

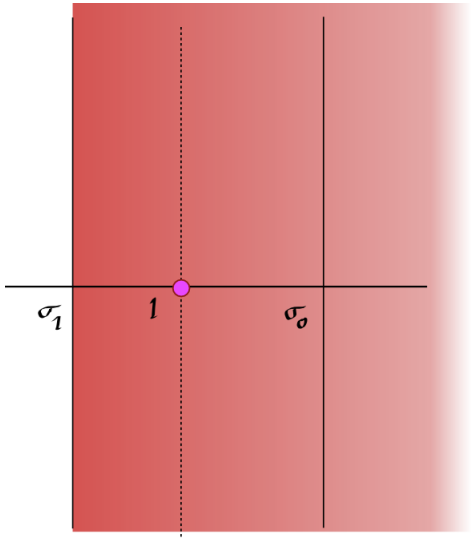
$1 - \sigma \leq t^{-\alpha}$



Situation 3

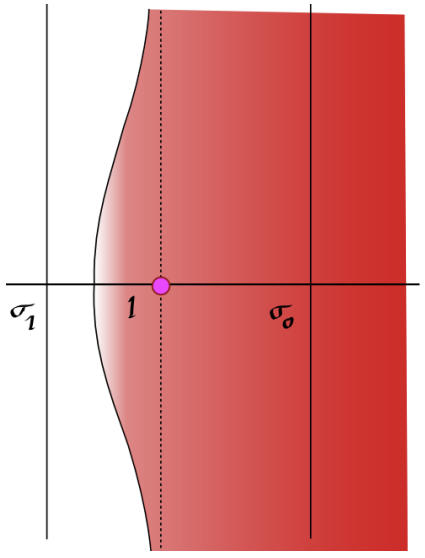
Vertical strip with holes

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



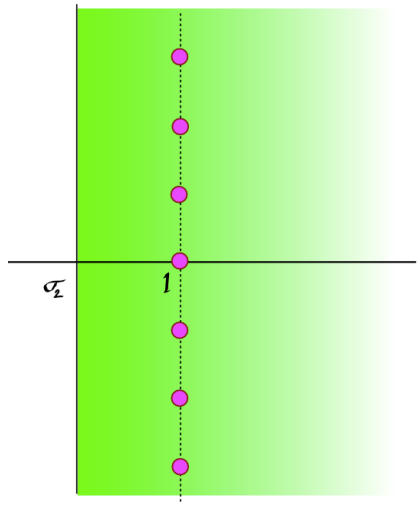
Situation 1

Vertical strip



Situation 2

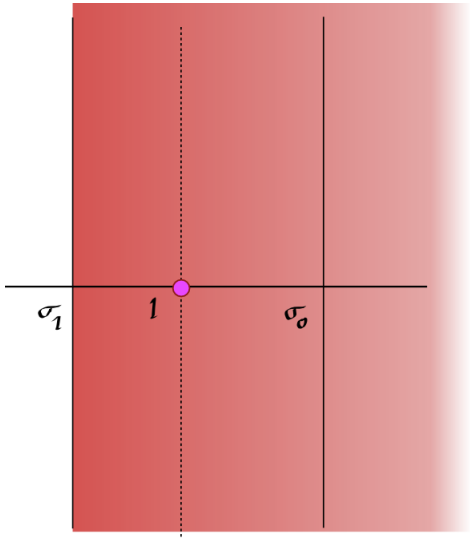
Hyperbolic region



Situation 3

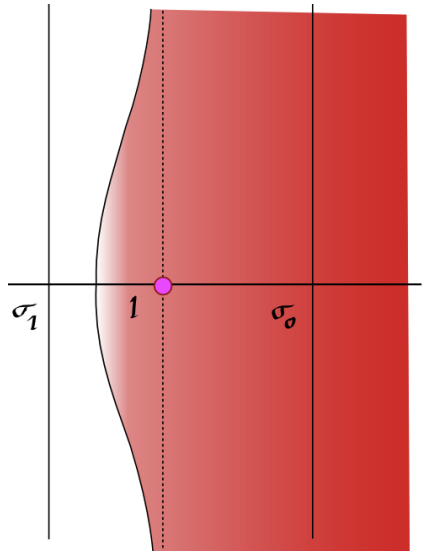
Vertical strip with holes

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



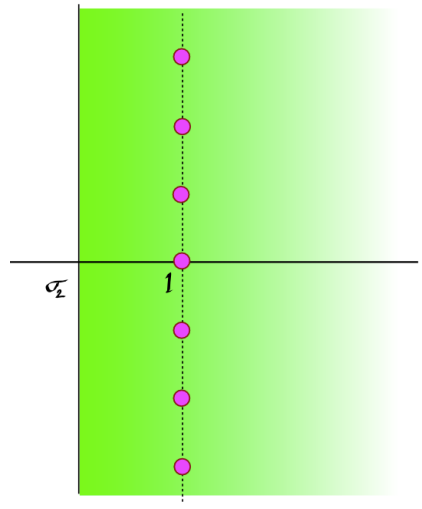
Situation 1

Vertical strip



Situation 2

Hyperbolic region

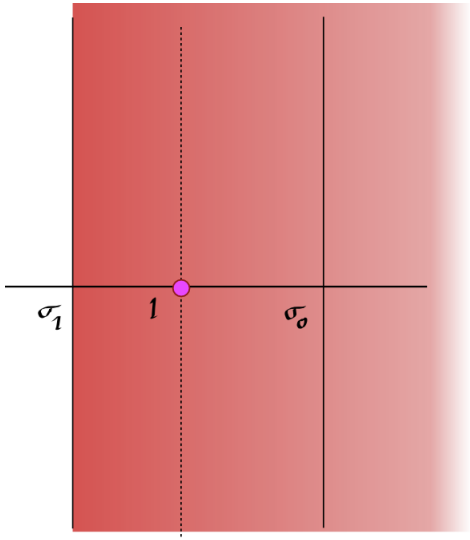


Situation 3

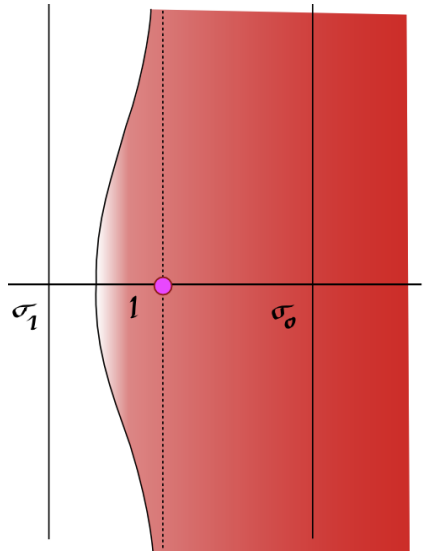
Vertical strip with holes

For which simple sources do these different situations occur?

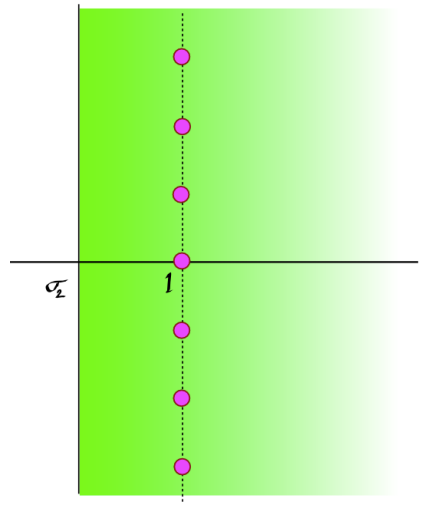
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



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 the behaviour of $\mathcal{Z} := \{t \in \mathbb{R};\quad \lambda(1+it) = 1\}$

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 the behaviour of $\mathcal{Z} := \{t \in \mathbb{R};\quad \lambda(1+it) = 1\}$

(i) $\mathcal{Z} \neq \{0\}$ $\iff$ $\log p_2/\log p_1$ is rational

(ii) If $t \in \mathcal{Z} \setminus \{0\}$ then $s \mapsto \lambda(s)$ is periodic with period it

(iii) If $\mathcal{Z} = \{0\}$, then the poles of $\Lambda(s)$ close to $\Re s = 1$ are related to 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 the behaviour of $\mathcal{Z} := \{t \in \mathbb{R};\quad \lambda(1+it) = 1\}$

(i) $\mathcal{Z} \neq \{0\} \iff \log p_2/\log p_1$ is rational

(ii) If $t \in \mathcal{Z} \setminus \{0\}$ then $s \mapsto \lambda(s)$ is periodic with period it

(iii) If $\mathcal{Z} = \{0\}$, then the poles of $\Lambda(s)$ close to $\Re s = 1$
are related to good rational approximations of $\log p_2/\log p_1$

The irrationality exponent $\theta(x)$ 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 $\iff \theta(x) < \infty$

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 the behaviour of $\mathcal{Z} := \{t \in \mathbb{R};\quad \lambda(1+it) = 1\}$

(i) $\mathcal{Z} \neq \{0\} \iff \log p_2/\log p_1$ is rational

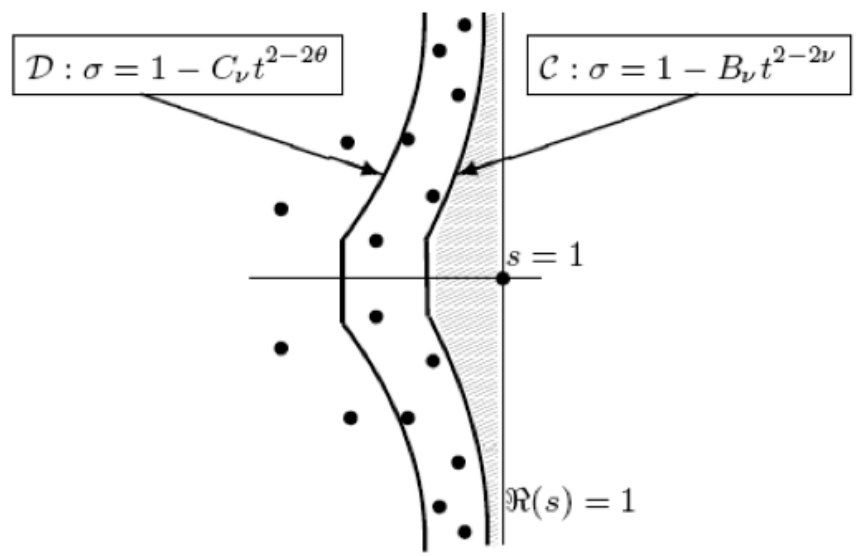
(ii) If $t \in \mathcal{Z} \setminus \{0\}$ then $s \mapsto \lambda(s)$ is periodic with period it

(iii) If $\mathcal{Z} = \{0\}$, then the poles of $\Lambda(s)$ close to $\Re s = 1$
are related to good rational approximations of $\log p_2/\log p_1$

The irrationality exponent $\theta(x)$ 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

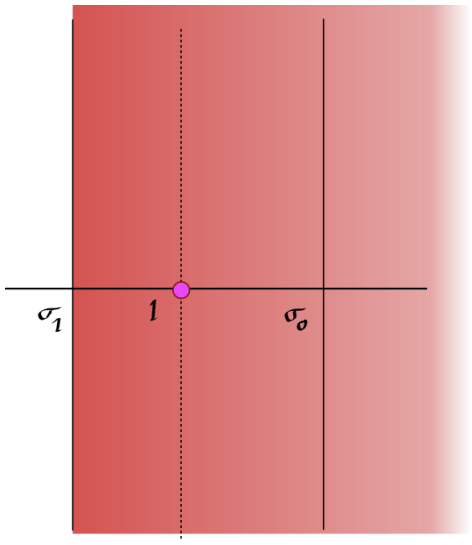
x diophantine $\iff \theta(x) < \infty$

If the irrationality exponent
of $\log p_2/\log p_1$ equals μ
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.

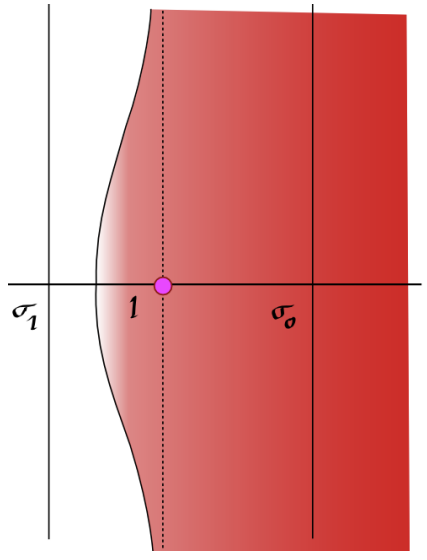
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

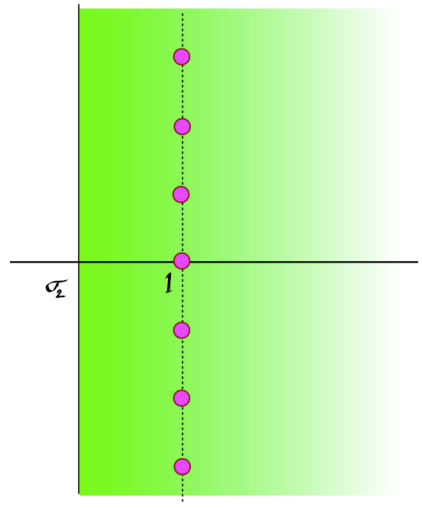
Geometric condition



Situation 2

Hyperbolic region

Arithmetic condition



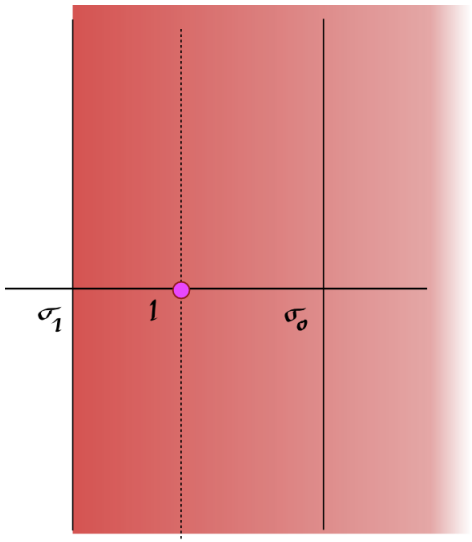
Situation 3

Vertical strip with holes

Periodicity condition

For which general dynamical sources do these different situations occur?

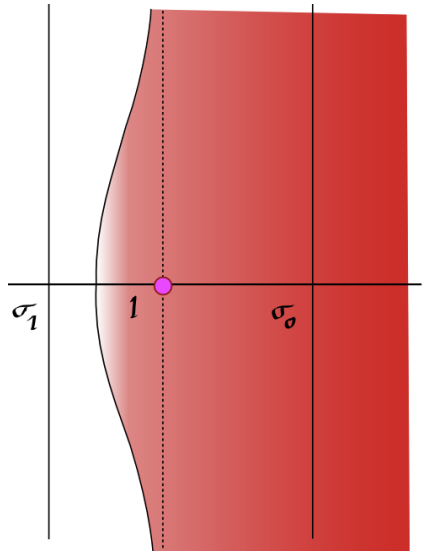
Different possible regions on the left of $\Re s = 1$ where $\Lambda(s)$ is tame.



Situation 1

Vertical strip

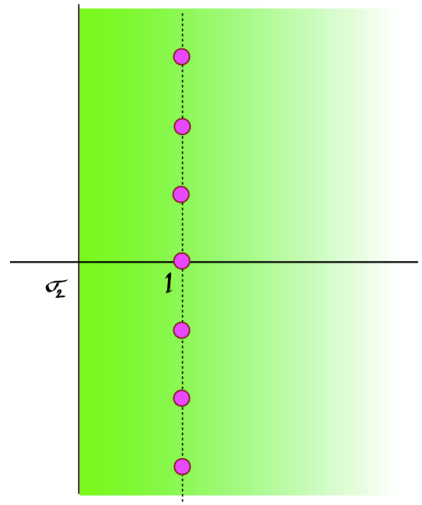
Geometric condition



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.

Conclusion

This talk

– describes a general model for sources

– shows the importance of the Dirichlet generating functions

– explains the importance of tameness in the analyses of text algorithms

– defines a natural subclass of sources, the dynamical sources

– provides sufficient conditions for tameness of dynamical sources

Finally, it provides a precise analysis of text algorithms
when the text is created by a dynamical source.

General result.

When built on a good dynamical source $\mathcal{S}$,

– the mean path-length $T(n)$ of `Trie`,

– the mean symbol path-length $B(n)$ of `Bst`

are both of the form $X(n) = P_X(n) + E(n)$.

General result.

When built on a good dynamical source $\mathcal{S}$,

– the mean path-length $T(n)$ of `Trie`,

– the mean symbol path-length $B(n)$ of `Bst`

are both of the form $X(n) = P_X(n) + E(n)$.

The "principal term" $P_X(n)$ involves the entropy $h(\mathcal{S})$

$$P_T(n) = \frac{1}{h(\mathcal{S})} n \log n + an, \qquad P_B(n) = \frac{1}{h(\mathcal{S})} n \log^2 n + bn \log n + cn,$$

General result.

When built on a good dynamical source $\mathcal{S}$,

– the mean path-length $T(n)$ of `Trie`,

– the mean symbol path-length $B(n)$ of `Bst`

are both of the form $X(n) = P_X(n) + E(n)$.

The "principal term" $P_X(n)$ involves the entropy $h(\mathcal{S})$

$$P_T(n) = \frac{1}{h(\mathcal{S})} n \log n + an, \qquad P_B(n) = \frac{1}{h(\mathcal{S})} n \log^2 n + bn \log n + cn,$$

The order of the "error term" $E(n)$ depends on the tameness of the source

In Situation S1	$E(n) = O(n^{1-\delta})$	$\delta \in [0,1]$
In Situation S2	$E(n) = n \cdot O(\exp[-(\log n)^{\alpha}])$	$\alpha < 1$
In Situation S3	$E(n) = n \cdot \Phi(\log n) + O(n^{1-\delta})$,	$\delta \in [0,1]$ and Φ periodic.