PROBABILISTIC ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

PROBABILISTIC ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

Work based on an idea of Philippe Flajolet,

begun with him at the end of 2010

Dedicated to his memory.

PROBABILISTIC ANALYSIS OF DIGITAL SEARCH TREES BUILT ON A GENERAL SOURCE

Kanal Hun and Brigitte Vallée,

GREYC (CNRS and University of Caen)

Work based on an idea of Philippe Flajolet,

begun with him at the end of 2010

Dedicated to his memory.

AofA 2013, Menorca

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.

- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.

- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.
- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

This realistic analysis is already successful for two other types of trees :

- `tries` and bst, when they are built on general sources
- Why not `dst`, since it is a mixing of these two structures?

Digital Search Tree is a fundamental data structure in Computer Science.

It underlies the compression algorithms of Lempel Ziv type.

It contains the "phrases" created by the algorithm.

This is already analyzed when the text is emitted by simple sources.

- First (seminal) study :

  Flajolet and Sedgewick (1986) for the unbiased binary source.
- Then, for memoryless sources and Markov chains, (1990–2000)

  Works of Jacquet, Louchard, Prodinger, Szpankowski, Tang.

Important to analyze this structure under general models of sources

(more realistic, more correlated)

This realistic analysis is already successful for two other types of trees :

- `tries` and bst, when they are built on general sources
- Why not `dst`, since it is a mixing of these two structures?

This talk : Analysis of Digital Search Trees

when they are built on words emitted by a general source.

(I) Digital Search Trees and simple sources

Description of the dst structure

$\mathcal{Y}$ = a sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

Description of the dst structure

$\mathcal{Y}$ = a sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

For $\mathcal{Y} = \emptyset$, then $\texttt{dst}(\mathcal{Y}) = \emptyset$. Otherwise,

– The first word of the sequence $\mathcal{Y}$ is at the root. Root [dst $(\mathcal{Y})$] := First $(\mathcal{Y})$.

– There are two subtrees built with the sequence $\underline{\mathcal{Y}} := \mathcal{Y} \setminus \{\text{First}(\mathcal{Y})\}$,

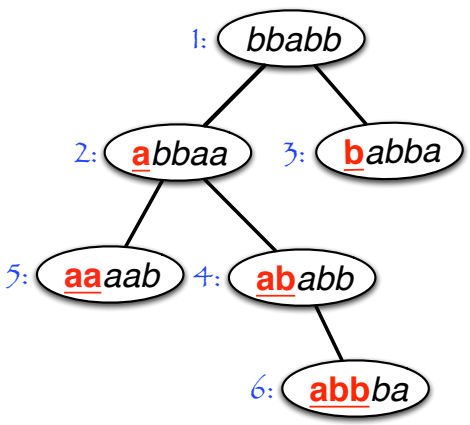
Description of the dst structure

$\mathcal{Y}$ = a sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

For $\mathcal{Y} = \emptyset$, then $\mathtt{dst}(\mathcal{Y}) = \emptyset$. Otherwise,

– The first word of the sequence $\mathcal{Y}$ is at the root. Root [dst $(\mathcal{Y})$] := First $(\mathcal{Y})$.

– There are two subtrees built with the sequence $\underline{\mathcal{Y}} := \mathcal{Y} \setminus \{\mathrm{First}(\mathcal{Y})\}$,



The dst built on six words
$Y_1 = bbabb, Y_2 = abbaa,$
$Y_3 = babba, Y_4 = ababb$
$Y_5 = aaaab, Y_6 = abbba$

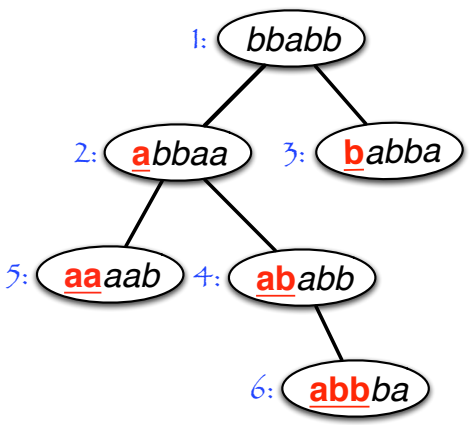
Description of the dst structure

$\mathcal{Y}$ = a sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

For $\mathcal{Y} = \emptyset$, then $\mathtt{dst}(\mathcal{Y}) = \emptyset$. Otherwise,

– The first word of the sequence $\mathcal{Y}$ is at the root. Root [dst $(\mathcal{Y})$] := First $(\mathcal{Y})$.

– There are two subtrees built with the sequence $\underline{\mathcal{Y}} := \mathcal{Y} \setminus \{\mathrm{First}(\mathcal{Y})\}$,



The dst built on six words
$Y_1 = bbabb, Y_2 = abbaa,$
$Y_3 = babba, Y_4 = ababb$
$Y_5 = aaaab, Y_6 = abbba$

The left subtree is built with the subsequence $\mathcal{Y}_{(a)}$ formed with words of $\underline{\mathcal{Y}}$ which begin with a from which the prefix a is removed.

Left [dst $(\mathcal{Y})$] := $\mathtt{dst}(\mathcal{Y}_{(a)})$.

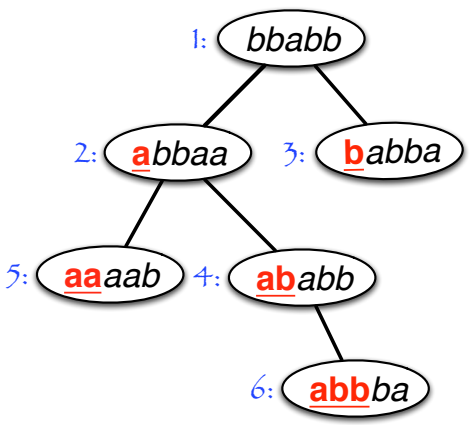
Description of the dst structure

$\mathcal{Y}$ = a sequence of infinite words on the alphabet $\Sigma := \{a, b\}$

For $\mathcal{Y} = \emptyset$, then $\mathtt{dst}(\mathcal{Y}) = \emptyset$. Otherwise,

– The first word of the sequence $\mathcal{Y}$ is at the root. Root [dst $(\mathcal{Y})$] := First $(\mathcal{Y})$.

– There are two subtrees built with the sequence $\underline{\mathcal{Y}} := \mathcal{Y} \setminus \{\mathrm{First}(\mathcal{Y})\}$,



The dst built on six words
$Y_1 = bbabb, Y_2 = abbaa,$
$Y_3 = babba, Y_4 = ababb$
$Y_5 = aaaab, Y_6 = abbba$

The left subtree is built with the subsequence $\mathcal{Y}_{(a)}$ formed with words of $\underline{\mathcal{Y}}$ which begin with a from which the prefix a is removed.

Left [dst $(\mathcal{Y})$] := $\mathtt{dst}(\mathcal{Y}_{(a)})$.

The right subtree is built with the subsequence $\mathcal{Y}_{(b)}$ formed with words of $\underline{\mathcal{Y}}$ which begin with b from which the prefix b is removed.

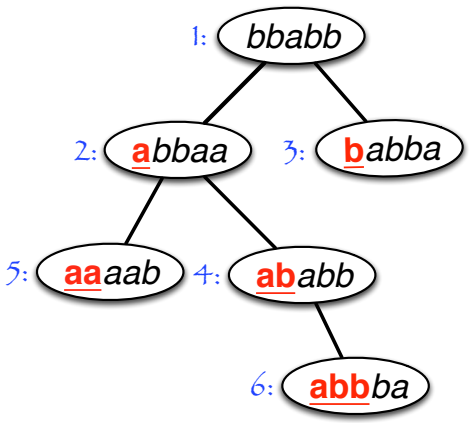
Right [dst $(\mathcal{Y})$] := $\mathtt{dst}(\mathcal{Y}_{(b)})$.

Analysis of the `dst` structure $=$ Description of the shape of the `dst` built on a sequence of n words independently emitted from a source $\mathcal{S}$

Analysis of the dst structure = Description of the shape of the dst

built on a sequence of n words independently emitted from a source $\mathcal{S}$

Level of a node := number of nodes between the node and the root.



Main parameters of interest

– the path length ℓ_n := sum of the levels of the nodes

– the profile $b_{n,k}$:= number of nodes at level k

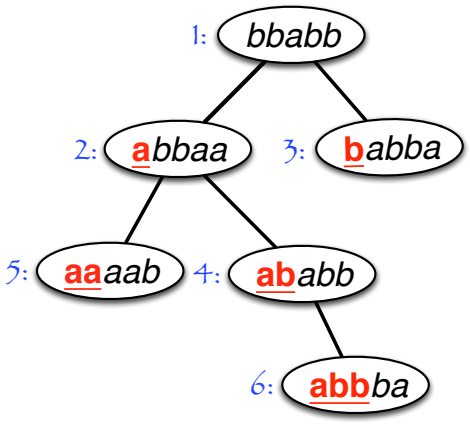
– the typical depth d_n = the level of a random node

$$\ell_n = \sum_k k\, b_{n,k} \qquad \Pr[d_n = k] = \frac{1}{n}\mathbb{E}[b_{n,k}]$$

Analysis of the dst structure = Description of the shape of the dst

built on a sequence of n words independently emitted from a source $\mathcal{S}$

Level of a node := number of nodes between the node and the root.



Main parameters of interest

– the path length ℓ_n := sum of the levels of the nodes

– the profile $b_{n,k}$:= number of nodes at level k

– the typical depth d_n = the level of a random node

$$\ell_n = \sum_k k\, b_{n,k} \qquad \Pr[d_n = k] = \frac{1}{n}\mathbb{E}[b_{n,k}]$$

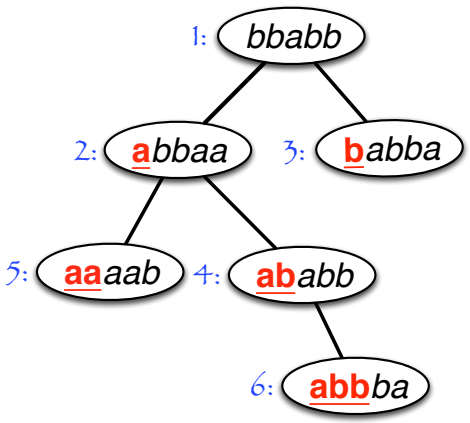
The characteristic function $\mathbb{E}[u^{d_n}]$ of the main parameter d_n satisfies

$$\mathbb{E}[u^{d_n}] = \frac{1}{n}B_n(u), \qquad \text{with} \quad B_n(u) := \sum_{k\geq 0}\mathbb{E}[b_{n,k}]u^k$$

Analysis of the dst structure = Description of the shape of the dst

built on a sequence of n words independently emitted from a source $\mathcal{S}$

Level of a node := number of nodes between the node and the root.



Main parameters of interest

– the path length ℓ_n := sum of the levels of the nodes

– the profile $b_{n,k}$:= number of nodes at level k

– the typical depth d_n = the level of a random node

$$\ell_n = \sum_k k\, b_{n,k} \qquad \Pr[d_n = k] = \frac{1}{n}\mathbb{E}[b_{n,k}]$$

The characteristic function $\mathbb{E}[u^{d_n}]$ of the main parameter d_n satisfies

$$\mathbb{E}[u^{d_n}] = \frac{1}{n}B_n(u), \qquad \text{with} \quad B_n(u) := \sum_{k\geq 0}\mathbb{E}[b_{n,k}]u^k$$

Main object for the study: the Poisson gf $\quad B(z) := e^{-z}\sum_{n\geq 0} B_n(u)\frac{z^n}{n!}$

Simple sources

Simple sources

A source:= a random mechanism which emits symbols from alphabet Σ,

The time is discrete, $X_i :=$ the symbol emitted at time $t = i$

The source is defined by the sequence $(X_0, X_1, \ldots, X_k, X_{k+1} \ldots)$

Simple sources

A source:= a random mechanism which emits symbols from alphabet Σ,

The time is discrete, $X_i :=$ the symbol emitted at time $t = i$

The source is defined by the sequence $(X_0, X_1, \ldots, X_k, X_{k+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Simple sources

A source:= a random mechanism which emits symbols from alphabet Σ,

The time is discrete, $X_i :=$ the symbol emitted at time $t = i$

The source is defined by the sequence $(X_0, X_1, \ldots, X_k, X_{k+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source : The variables X_k are independent,

with the same distribution defined by $p_i := \Pr[X_k = i]$ $(i \in \Sigma)$

$$\lambda(s) := p_1^s + p_2^s + \ldots + p_r^s$$

Simple sources

A source:= a random mechanism which emits symbols from alphabet Σ,

The time is discrete, $\quad X_i :=$ the symbol emitted at time $t = i$

The source is defined by the sequence $(X_0, X_1, \ldots, X_k, X_{k+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source : The variables X_k are independent,

with the same distribution defined by $p_i := \Pr[X_k = i]$ $(i \in \Sigma)$

$$\lambda(s) := p_1^s + p_2^s + \ldots + p_r^s$$

Markov chain: The only dependence is between consecutive X_k's

defined by the transition matrix $p_{i|j} := \Pr[X_{k+1} = i \,|X_k = j]$

$\lambda(s) :=$ the dominant eigenvalue of the Dirichlet matrix $\mathbf{P}_s := (p_{i|j}^s)$

Simple sources

A source:= a random mechanism which emits symbols from alphabet Σ,
The time is discrete, $X_i :=$ the symbol emitted at time $t = i$
The source is defined by the sequence $(X_0, X_1, \ldots, X_k, X_{k+1} \ldots)$

Simple sources: sources with weak correlations between successive symbols

Memoryless source : The variables X_k are independent,
with the same distribution defined by $p_i := \Pr[X_k = i]$ $(i \in \Sigma)$

$$\lambda(s) := p_1^s + p_2^s + \ldots + p_r^s$$

Markov chain: The only dependence is between consecutive X_k's
defined by the transition matrix $p_{i|j} := \Pr[X_{k+1} = i \,|X_k = j]$
$\lambda(s) :=$ the dominant eigenvalue of the Dirichlet matrix $\mathbf{P}_s := (p_{i|j}^s)$

In both cases, $\lambda(s)$ is called the dominant eigenvalue of the source.
The position of the set $\mathcal{Z} := \{s, \lambda(s) = 1\}$ wrt the vertical line $\Re s = 1$
is essential. It is related to arithmetical properties of probabilities.

(II) General sources and new results.

(II) General sources and new results.

Find a natural extension of the transition matrix $\mathbf{P}_s$

(II) General sources and new results.

Find a natural extension of the transition matrix $\mathbf{P}_s$

Deal with it for extending the results for dst's to a general source

General sources

For applications, importance to deal with a general source $\mathcal{S}$

General sources

For applications, importance to deal with a general source $\mathcal{S}$

A general source $\mathcal{S}$ is completely defined by its fundamental probabilities

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

General sources

For applications, importance to deal with a general source $\mathcal{S}$

A general source $\mathcal{S}$ is completely defined by its fundamental probabilities

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

The source $\mathcal{S}$ defines a sequence of shifted sources $\mathcal{S}_{(u)}$ (for $u \in \Sigma^\star$)

For $u \in \Sigma^\star$ with $p_u \neq 0$, the source $\mathcal{S}_{(u)} = \mathcal{S}|_u$ is a shifted source

- which gathers all the words of $\mathcal{S}$ which begin with $u \in \Sigma^\star$,
- from which the prefix u is removed.

General sources

For applications, importance to deal with a general source $\mathcal{S}$

A general source $\mathcal{S}$ is completely defined by its fundamental probabilities

$p_w :=$ the probability that a word of $\mathcal{S}$ begins with the prefix $w \in \Sigma^\star$

The source $\mathcal{S}$ defines a sequence of shifted sources $\mathcal{S}_{(u)}$ (for $u \in \Sigma^\star$)

For $u \in \Sigma^\star$ with $p_u \neq 0$, the source $\mathcal{S}_{(u)} = \mathcal{S}|_u$ is a shifted source

– which gathers all the words of $\mathcal{S}$ which begin with $u \in \Sigma^\star$,

– from which the prefix u is removed.

When w is any finite sur-fix of u such that $w = u \cdot v$

the conditional probabilities $p_w/p_u = p_{(u.v)|u}$, denoted as $q_{v|u}$

are the fundamental probabilities of the source $S_{(u)}$.

A source is smooth if the probabilities $q_{i|w}$ satisfy

$$\exists p < 1, \qquad 0 < q_{i|w} \leq p, \qquad \forall (i, w) \in \Sigma \times \Sigma^\star$$

The graph and the basic operator of a source $\mathcal{S}$

The graph and the basic operator of a source $\mathcal{S}$

The graph of the source is defined by

– its vertices: the shifted sources $\mathcal{S}_{(w)}$,

– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.

Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.

The graph and the basic operator of a source $\mathcal{S}$

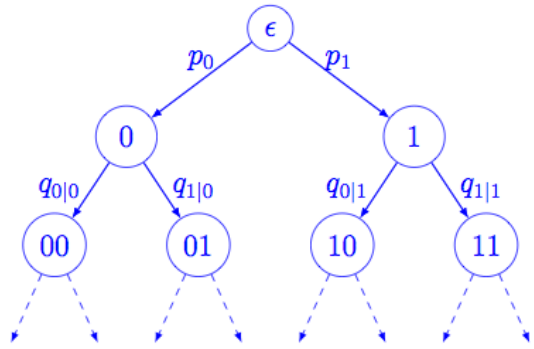
The graph of the source is defined by

– its vertices: the shifted sources $\mathcal{S}_{(w)}$,

– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.

Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.



The graph and the basic operator of a source $\mathcal{S}$

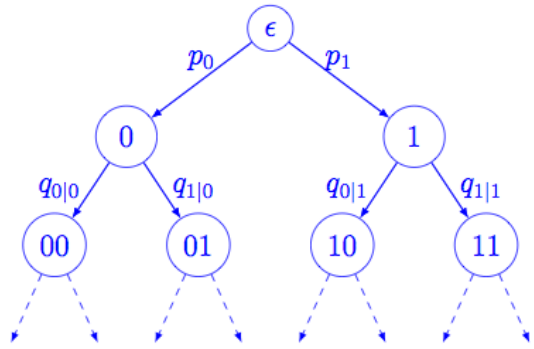
The graph of the source is defined by

– its vertices: the shifted sources $\mathcal{S}_{(w)}$,

– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.

Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.



The matrix $\mathbf{P}_s$ with coefficients $q^s_{i|w}$ at $(w, w\cdot i)$

is called the basic operator of the source. – the main object here.

The graph and the basic operator of a source $\mathcal{S}$

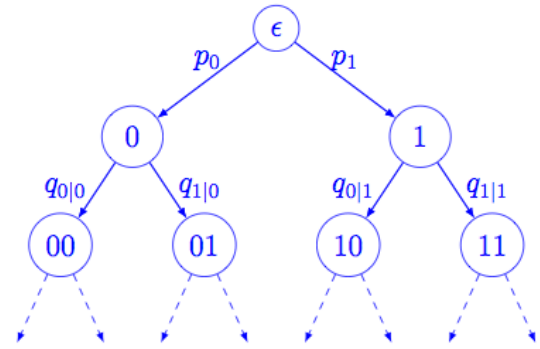
The graph of the source is defined by

– its vertices: the shifted sources $\mathcal{S}_{(w)}$,

– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.

Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.



The matrix $\mathbf{P}_s$ with coefficients $q^s_{i|w}$ at $(w, w\cdot i)$

is called the basic operator of the source. – the main object here.

Sometimes, the graph is redundant and can be pruned via the equivalence relation

$\mathcal{S}_{(u)} \equiv \mathcal{S}_{(v)} \iff \mathcal{S}_{(u)}$ and $\mathcal{S}_{(v)}$ have the same fundamental probabilities

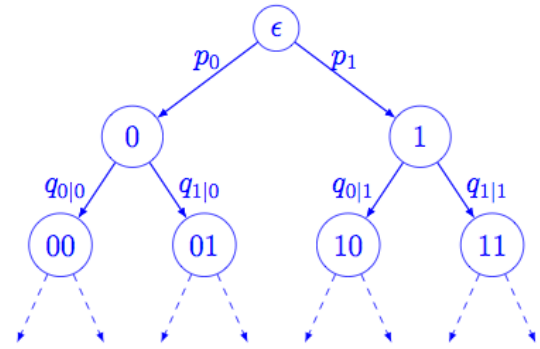
For a simple source, the pruned matrix is a finite matrix,

closely related to the "classical" transition matrix $\mathbf{P}_s$

The graph and the basic operator of a source $\mathcal{S}$

The graph of the source is defined by
– its vertices: the shifted sources $\mathcal{S}_{(w)}$,
– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.
Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.



The matrix $\mathbf{P}_s$ with coefficients $q^s_{i|w}$ at $(w, w\cdot i)$
is called the basic operator of the source. – the main object here.

Sometimes, the graph is redundant and can be pruned via the equivalence relation

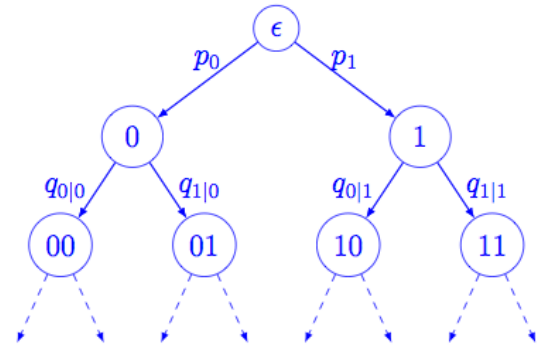
$\mathcal{S}_{(u)} \equiv \mathcal{S}_{(v)} \iff \mathcal{S}_{(u)}$ and $\mathcal{S}_{(v)}$ have the same fundamental probabilities

For a simple source, the pruned matrix is a finite matrix,
closely related to the "classical" transition matrix $\mathbf{P}_s$

The graph and the basic operator of a source $\mathcal{S}$

The graph of the source is defined by
– its vertices: the shifted sources $\mathcal{S}_{(w)}$,
– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.
Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.

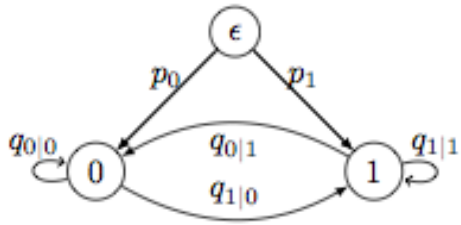


The matrix $\mathbf{P}_s$ with coefficients $q^s_{i|w}$ at $(w, w\cdot i)$
is called the basic operator of the source. – the main object here.

Sometimes, the graph is redundant and can be pruned via the equivalence relation

$\mathcal{S}_{(u)} \equiv \mathcal{S}_{(v)} \iff \mathcal{S}_{(u)}$ and $\mathcal{S}_{(v)}$ have the same fundamental probabilities

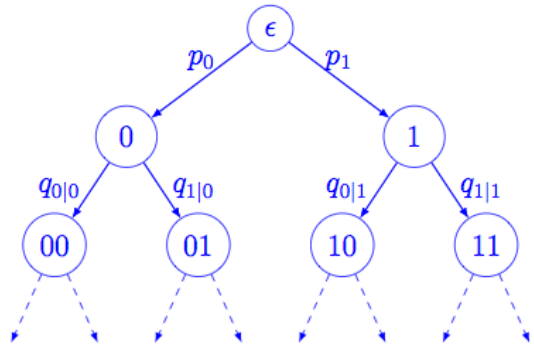
For a simple source, the pruned matrix is a finite matrix,
closely related to the "classical" transition matrix $\mathbf{P}_s$



The graph and the basic operator of a source $\mathcal{S}$

The graph of the source is defined by
– its vertices: the shifted sources $\mathcal{S}_{(w)}$,
– its edge from $\mathcal{S}_{(w)}$ to $\mathcal{S}_{(w\cdot i)}$ is weighted with $q_{i|w}$.

Its matrix $\mathbf{P}$ has its rows and columns indexed by $\Sigma^\star$.
Its coefficient at $(w, w\cdot i)$ equals $q_{i|w}$.

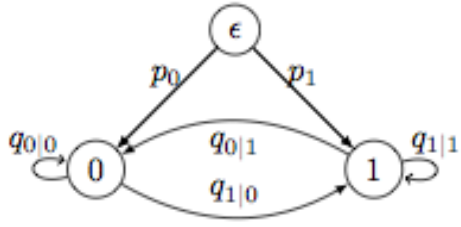


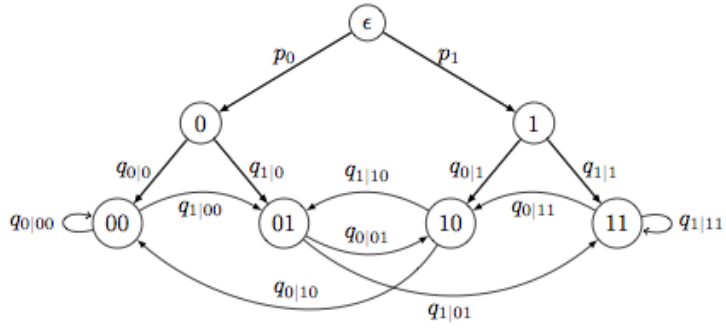
The matrix $\mathbf{P}_s$ with coefficients $q^s_{i|w}$ at $(w, w\cdot i)$
is called the basic operator of the source. – the main object here.

Sometimes, the graph is redundant and can be pruned via the equivalence relation

$\mathcal{S}_{(u)} \equiv \mathcal{S}_{(v)} \iff \mathcal{S}_{(u)}$ and $\mathcal{S}_{(v)}$ have the same fundamental probabilities

For a simple source, the pruned matrix is a finite matrix,
closely related to the "classical" transition matrix $\mathbf{P}_s$





General facts for the analysis of a `dst` on a general source.

General facts for the analysis of a `dst` on a general source.

As we will see, for simple sources as for general sources,
the analysis for `dst` is based on the behavior of the infinite product

$$\mathbf{Q}(s,u) := (I - u\mathbf{P}_s)^{-1} \circ (I - u\mathbf{P}_{s+1})^{-1} \circ (I - u\mathbf{P}_{s+2})^{-1} \circ \dots ...$$

The analyses for `tries` only involve the first factor via $(I - u\mathbf{P}_s)^{-1}[1]$.

General facts for the analysis of a dst on a general source.

As we will see, for simple sources as for general sources,
the analysis for dst is based on the behavior of the infinite product

$$\mathbf{Q}(s,u) := (I - u\mathbf{P}_s)^{-1} \circ (I - u\mathbf{P}_{s+1})^{-1} \circ (I - u\mathbf{P}_{s+2})^{-1} \circ \dots ...$$

The analyses for tries only involve the first factor via $(I - u\mathbf{P}_s)^{-1}[1]$.

For $\Re s$ close to 1, the dominant behavior of $\mathbf{Q}(s,u)$
is actually dictated by the first factor $(I - u\mathbf{P}_s)^{-1}$.
This explains the similarity of the results for dst and trie

General facts for the analysis of a dst on a general source.

As we will see, for simple sources as for general sources,
the analysis for dst is based on the behavior of the infinite product

$$\mathbf{Q}(s,u) := (I - u\mathbf{P}_s)^{-1} \circ (I - u\mathbf{P}_{s+1})^{-1} \circ (I - u\mathbf{P}_{s+2})^{-1} \circ \dots ...$$

The analyses for tries only involve the first factor via $(I - u\mathbf{P}_s)^{-1}[1]$.

For $\Re s$ close to 1, the dominant behavior of $\mathbf{Q}(s,u)$
is actually dictated by the first factor $(I - u\mathbf{P}_s)^{-1}$.
This explains the similarity of the results for dst and trie

However, the analysis for dst with a general source is more difficult:
– We need to study the operator $\mathbf{P}_s$ itself, not only $(I - u\mathbf{P}_s)^{-1}[1]$
– For a general source, the operator $\mathbf{P}_s$ is no longer a finite matrix, and the properties of $\mathbf{P}_s$ depend on the functional space on which it acts.

Tameness of a source.

It is defined via the tameness of the operator $\mathbf{P}_s$

Tameness of a source.

It is defined via the tameness of the operator $\mathbf{P}_s$

An operator $\mathbf{P}_s$ is said to be tame if there exists a functional space

for which $\mathbf{P}_s$ satisfies "nice" analytic properties

– it possesses a dominant eigenvalue $\lambda(s)$ for s close to 1.

– $(I - \mathbf{P}_s)^{-1}$ is of polynomial growth when $|\Im s| \to \infty$

on a region $\mathcal{R}$ located on the left of the vertical line $\Re s = 1$.

Tameness of a source.

It is defined via the tameness of the operator $\mathbf{P}_s$

An operator $\mathbf{P}_s$ is said to be tame if there exists a functional space
for which $\mathbf{P}_s$ satisfies "nice" analytic properties

- it possesses a dominant eigenvalue $\lambda(s)$ for s close to 1.
- $(I - \mathbf{P}_s)^{-1}$ is of polynomial growth when $|\Im s| \to \infty$
  on a region $\mathcal{R}$ located on the left of the vertical line $\Re s = 1$.

This gives rise to various notions of tameness for a source. A source is

- tame if its basic operator $\mathbf{P}_s$ is tame.
- super-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1, $|u| = 1$.
- hyper-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1.

Tameness of a source.

It is defined via the tameness of the operator $\mathbf{P}_s$

An operator $\mathbf{P}_s$ is said to be tame if there exists a functional space
for which $\mathbf{P}_s$ satisfies "nice" analytic properties

- it possesses a dominant eigenvalue $\lambda(s)$ for s close to 1.
- $(I - \mathbf{P}_s)^{-1}$ is of polynomial growth when $|\Im s| \to \infty$
  on a region $\mathcal{R}$ located on the left of the vertical line $\Re s = 1$.

This gives rise to various notions of tameness for a source. A source is

- tame if its basic operator $\mathbf{P}_s$ is tame.
- super-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1, $|u| = 1$.
- hyper-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1.

These notions extend in a natural way all the situations already studied
for simple sources where the operator $\mathbf{P}_s$ is a finite matrix

Tameness of a source.

It is defined via the tameness of the operator $\mathbf{P}_s$

An operator $\mathbf{P}_s$ is said to be tame if there exists a functional space
for which $\mathbf{P}_s$ satisfies "nice" analytic properties

- it possesses a dominant eigenvalue $\lambda(s)$ for s close to 1.
- $(I - \mathbf{P}_s)^{-1}$ is of polynomial growth when $|\Im s| \to \infty$
  on a region $\mathcal{R}$ located on the left of the vertical line $\Re s = 1$.

This gives rise to various notions of tameness for a source. A source is

- tame if its basic operator $\mathbf{P}_s$ is tame.
- super-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1, $|u| = 1$.
- hyper-tame if the operator $u\mathbf{P}_s$ is tame for u close to 1.

These notions extend in a natural way all the situations already studied
for simple sources where the operator $\mathbf{P}_s$ is a finite matrix

Most of "good" dynamical sources are tame or even hyper-tame.

Our main results in the case of a general source – (I) Average -case analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be a smooth, stationary, and tame, its basic operator $\mathbf{P}_s$, its dominant eigenvalue $\lambda(s)$.
Then, the entropy of the source $\mathcal{S}$ is equal to $-\lambda'(1)$.

Our main results in the case of a general source – (I) Average -case analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be a smooth, stationary, and tame, its basic operator $\mathbf{P}_s$, its dominant eigenvalue $\lambda(s)$.
Then, the entropy of the source $\mathcal{S}$ is equal to $-\lambda'(1)$.

Consider n words independently emitted by $\mathcal{S}$.

Our main results in the case of a general source – (I) Average -case analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be a smooth, stationary, and tame, its basic operator $\mathbf{P}_s$, its dominant eigenvalue $\lambda(s)$.

Then, the entropy of the source $\mathcal{S}$ is equal to $-\lambda'(1)$.

Consider n words independently emitted by $\mathcal{S}$.

(a) Then, the mean typical depth of the `dst` satisfies

$$\mathbb{E}[d_n] = \frac{1}{h_{\mathcal{S}}} \log n + A_{\mathcal{S}} + \delta_n + R_n.$$

Our main results in the case of a general source – (I) Average -case analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be a smooth, stationary, and tame, its basic operator $\mathbf{P}_s$, its dominant eigenvalue $\lambda(s)$.
Then, the entropy of the source $\mathcal{S}$ is equal to $-\lambda'(1)$.

Consider n words independently emitted by $\mathcal{S}$.

(a) Then, the mean typical depth of the `dst` satisfies

$$\mathbb{E}[d_n] = \frac{1}{h_{\mathcal{S}}} \log n + A_{\mathcal{S}} + \delta_n + R_n.$$

(b) The constant $A_{\mathcal{S}}$ is expressed with the infinite product $\mathbf{Q}_s$.

Our main results in the case of a general source – (I) Average -case analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be a smooth, stationary, and tame, its basic operator $\mathbf{P}_s$, its dominant eigenvalue $\lambda(s)$.
Then, the entropy of the source $\mathcal{S}$ is equal to $-\lambda'(1)$.

Consider n words independently emitted by $\mathcal{S}$.

(a) Then, the mean typical depth of the `dst` satisfies

$$\mathbb{E}[d_n] = \frac{1}{h_{\mathcal{S}}} \log n + A_{\mathcal{S}} + \delta_n + R_n.$$

(b) The constant $A_{\mathcal{S}}$ is expressed with the infinite product $\mathbf{Q}_s$.

(c) The form of δ_n and R_n depend on $\mathbf{P}_s$ and $\lambda(s)$.

- δ_n exists when $\lambda(s)$ is periodic. It is a periodic function of $\log n$.
- $R_n \to 0$. It depends on the shape of the tameness region itself related to diophantine properties of probabilities.

Our main results in the case of a general source – (II) Distributional analysis

Exact extension of previously known results for simple sources.

Our main results in the case of a general source – (II) Distributional analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be smooth, stationary and super-tame,

– which is not similar to a memoryless unbiased source.

– denote by $\lambda(s)$ its dominant eigenvalue .

Our main results in the case of a general source – (II) Distributional analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be smooth, stationary and super-tame,

– which is not similar to a memoryless unbiased source.

– denote by $\lambda(s)$ its dominant eigenvalue .

The typical depth d_n of a `dst` built on n words independently emitted by the source $\mathcal{S}$ follows an asymptotic gaussian law.

Our main results in the case of a general source – (II) Distributional analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be smooth, stationary and super-tame,
– which is not similar to a memoryless unbiased source.
– denote by $\lambda(s)$ its dominant eigenvalue .

The typical depth d_n of a `dst` built on n words independently emitted by the source $\mathcal{S}$ follows an asymptotic gaussian law.

The expectation $\mathbb{E}[d_n]$ and the variance $\mathbb{V}[d_n]$ are both of order $\log n$,

$$\mathbb{E}[d_n] \sim \frac{-1}{\lambda'(1)} \log n, \quad \mathbb{V}[d_n] \sim \frac{\lambda'(1)^2 - \lambda''(1)}{\lambda'(1)^3} \log n$$

The dominant constants are the same as for tries.

Our main results in the case of a general source – (II) Distributional analysis

Exact extension of previously known results for simple sources.

Consider a source $\mathcal{S}$ assumed to be smooth, stationary and super-tame,
– which is not similar to a memoryless unbiased source.
– denote by $\lambda(s)$ its dominant eigenvalue .

The typical depth d_n of a `dst` built on n words independently emitted by the source $\mathcal{S}$ follows an asymptotic gaussian law.

The expectation $\mathbb{E}[d_n]$ and the variance $\mathbb{V}[d_n]$ are both of order $\log n$,

$$\mathbb{E}[d_n] \sim \frac{-1}{\lambda'(1)} \log n, \quad \mathbb{V}[d_n] \sim \frac{\lambda'(1)^2 - \lambda''(1)}{\lambda'(1)^3} \log n$$

The dominant constants are the same as for tries.

When the source is hyper-tame,

the speed of convergence towards the gaussian law is $O(\log n)^{-1/2}$.

More generally, it depends on the shape of the tameness region

(III) Some hints on the methods.

Fimal aim: an asymptotic estimate of $B_n(u) = n![z^n](e^z B(z,u))$

closely related to the characteristic function of the typical depth

(III) Some hints on the methods.

Fimal aim: an asymptotic estimate of $B_n(u) = n![z^n](e^z B(z,u))$
closely related to the characteristic function of the typical depth

– First, Step 1: Derivation of a system of basic equations
satisfied by a family of Poisson generating functions

(III) Some hints on the methods.

Fimal aim: an asymptotic estimate of $B_n(u) = n![z^n](e^z B(z,u))$
closely related to the characteristic function of the typical depth

– First, Step 1: Derivation of a system of basic equations
satisfied by a family of Poisson generating functions

– Then, Steps 2, 3, 4 : Using three main tools : Laplace; Rice, Mellin

We consider a `dst` built on n words independently emitted by a source $\mathcal{S}$.

We consider a `dst` built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

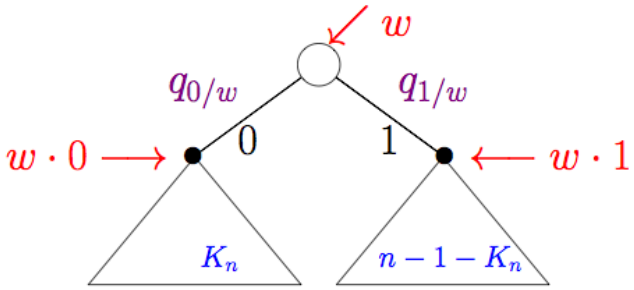
We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$

We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

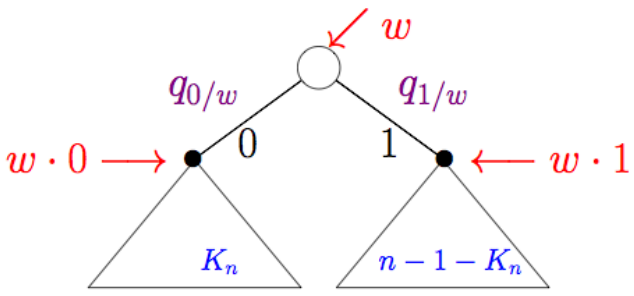
$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$



We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$



For $\Sigma := \{0, 1\}$, the sequence $b_{n,k}^{(w)}$ satisfies

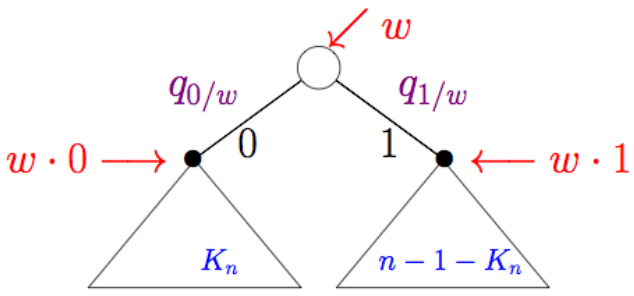
$$b_{n,k}^{(w)} = b_{K_n,k-1}^{(w.0)} + b_{n-1-K_n,k-1}^{(w.1)}$$

The number of nodes $K_n := K_n^{(w)}$ in the left subtree follows a binomial law of parameters $n-1$, $q_{0|w}$.

We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$



For $\Sigma := \{0, 1\}$, the sequence $b_{n,k}^{(w)}$ satisfies

$$b_{n,k}^{(w)} = b_{K_n,k-1}^{(w.0)} + b_{n-1-K_n,k-1}^{(w.1)}$$

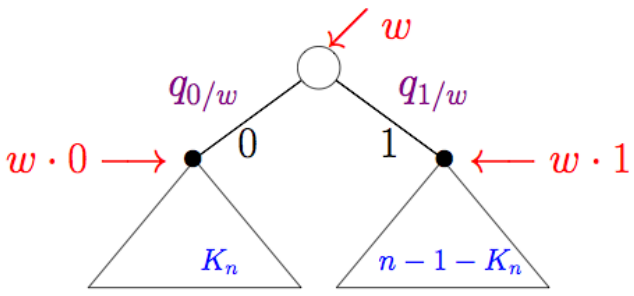
The number of nodes $K_n := K_n^{(w)}$ in the left subtree follows a binomial law of parameters $n-1$, $q_{0|w}$.

A system of equations for the Poisson gf's $B^{(w)}(z,u) := e^{-z} \sum_{n,k \geq 0} B_{n,k}^{(w)} \, u^k \, \frac{z^n}{n!}$

We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$



For $\Sigma := \{0, 1\}$, the sequence $b_{n,k}^{(w)}$ satisfies

$$b_{n,k}^{(w)} = b_{K_n,k-1}^{(w.0)} + b_{n-1-K_n,k-1}^{(w.1)}$$

The number of nodes $K_n := K_n^{(w)}$ in the left subtree follows a binomial law of parameters $n-1$, $q_{0|w}$.

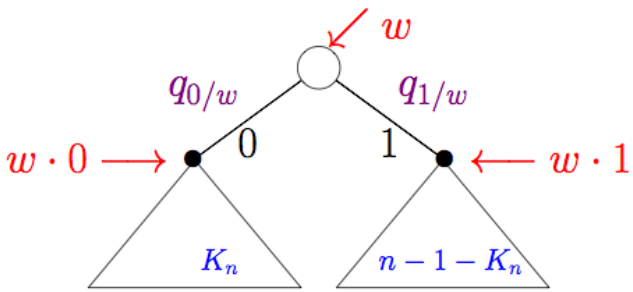
A system of equations for the Poisson gf's $B^{(w)}(z,u) := e^{-z} \sum_{n,k\geq 0} B_{n,k}^{(w)}\, u^k\, \frac{z^n}{n!}$

$$\frac{d}{dz} B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w} z, u)$$

We consider a dst built on n words independently emitted by a source $\mathcal{S}$.

We deal with all the shifted sources $\mathcal{S}_{(w)}$ and all the profiles

$$b_{n,k}^{(w)} = \text{the profile of a dst of size } n \text{ built on } \mathcal{S}_{(w)}$$



For $\Sigma := \{0, 1\}$, the sequence $b_{n,k}^{(w)}$ satisfies

$$b_{n,k}^{(w)} = b_{K_n,k-1}^{(w.0)} + b_{n-1-K_n,k-1}^{(w.1)}$$

The number of nodes $K_n := K_n^{(w)}$ in the left subtree follows a binomial law of parameters $n-1$, $q_{0|w}$.

A system of equations for the Poisson gf's $B^{(w)}(z,u) := e^{-z} \sum_{n,k\geq 0} B_{n,k}^{(w)}\, u^k\, \frac{z^n}{n!}$

$$\frac{d}{dz} B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u \sum_{i\in\Sigma} B^{(w.i)}(q_{i|w} z, u)$$

which involves three main operations

– the derivation d/dz – the change of variables $z \mapsto qz$

– the shift on words $w \mapsto w.i$

In comparison, for tries, the derivation does not occur.

An asymptotic estimate for $B_n(u)$?

Basic equations

Laplace $\Downarrow$

$$\frac{d}{dz}B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u\sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z, u)$$

$$\underline{B}(z,u) = \sum_{v\in\Sigma^\star} \delta(v,u)[e^{-zp_v} - 1 + zp_v] \quad \Longrightarrow \quad B_n(u) = \sum_{\ell=2}^{n}(-1)^\ell \binom{n}{\ell}\Delta(\ell,u)$$

An asymptotic estimate for $B_n(u)$?

Basic equations

Laplace $\Downarrow$

$$\frac{d}{dz}B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u\sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z, u)$$

$$\underline{B}(z,u) = \sum_{v\in\Sigma^\star} \delta(v,u)[e^{-zp_v} - 1 + zp_v] \quad \Longrightarrow \quad B_n(u) = \sum_{\ell=2}^{n}(-1)^\ell \binom{n}{\ell}\Delta(\ell,u)$$

with $\Delta(s,u) = \sum_{v\in\Sigma^\star}\delta(v,u)p_v^s$, $\qquad \delta(v,u) := \frac{1}{p_v}\sum_{w\geq v} u^{|w|}p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1-p_v p_\alpha^{-1}}$

An asymptotic estimate for $B_n(u)$?

Basic equations

Laplace $\Downarrow$

$$\frac{d}{dz}B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u\sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z, u)$$

$$\underline{B}(z,u) = \sum_{v\in\Sigma^\star} \delta(v,u)[e^{-zp_v} - 1 + zp_v] \quad \Longrightarrow \quad B_n(u) = \sum_{\ell=2}^{n}(-1)^\ell \binom{n}{\ell}\Delta(\ell,u)$$

with $\Delta(s,u) = \sum_{v\in\Sigma^\star}\delta(v,u)p_v^s$, $\qquad \delta(v,u) := \dfrac{1}{p_v}\displaystyle\sum_{w\geq v} u^{|w|}p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1-p_v p_\alpha^{-1}}$

Mellin $\Downarrow$ $\qquad$ $\Downarrow$ Rice

An asymptotic estimate for $B_n(u)$?

Basic equations

Laplace ⇓

$$\frac{d}{dz}B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u\sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z, u)$$

$$\underline{B}(z,u) = \sum_{v\in\Sigma^\star} \delta(v,u)[e^{-zp_v} - 1 + zp_v] \quad \Longrightarrow \quad B_n(u) = \sum_{\ell=2}^{n} (-1)^\ell \binom{n}{\ell} \Delta(\ell, u)$$

with $\Delta(s,u) = \sum_{v\in\Sigma^\star} \delta(v,u) p_v^s$, $\quad \delta(v,u) := \frac{1}{p_v} \sum_{w\geq v} u^{|w|} p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1 - p_v p_\alpha^{-1}}$

Mellin ⇓ ⇓ Rice

$$\Delta(s,u) = {}^t\mathbf{E}\ \mathbf{Q}_{s,u}\ \mathbf{Q}_{2,u}^{-1}\mathbf{1}$$
$$\mathbf{Q}(s,u) := (I - u\mathbf{P}_s)^{-1} \cdot (I - u\mathbf{P}_{s+1})^{-1} \dots$$
$${}^t\mathbf{E} = (1,0,0\dots) \quad {}^t\mathbf{1} = (1,1,\dots)$$

An asymptotic estimate for $B_n(u)$?

Basic equations

Laplace ⇓

$$\frac{d}{dz}B^{(w)}(z,u) + B^{(w)}(z,u) = 1 + u\sum_{i\in\Sigma} B^{(w.i)}(q_{i|w}z, u)$$

$$\underline{B}(z,u) = \sum_{v\in\Sigma^\star} \delta(v,u)[e^{-zp_v} - 1 + zp_v] \quad \Longrightarrow \quad B_n(u) = \sum_{\ell=2}^{n}(-1)^\ell \binom{n}{\ell}\Delta(\ell,u)$$

with $\Delta(s,u) = \sum_{v\in\Sigma^\star}\delta(v,u)p_v^s$, $\quad \delta(v,u) := \frac{1}{p_v}\sum_{w\geq v} u^{|w|}p_w \prod_{\substack{\alpha\leq w,\\ \alpha\neq v}} \frac{1}{1-p_v p_\alpha^{-1}}$

Mellin ⇓ ⇓ Rice

$\Delta(s,u) = {}^t\mathbf{E}\ \mathbf{Q}_{s,u}\ \mathbf{Q}_{2,u}^{-1}\mathbf{1}$

$\mathbf{Q}(s,u) := (I - u\mathbf{P}_s)^{-1}\cdot(I - u\mathbf{P}_{s+1})^{-1}\ldots$

${}^t\mathbf{E} = (1,0,0\ldots) \quad {}^t\mathbf{1} = (1,1,\ldots)$

$\Longrightarrow$

Needed: Nice properties of $s \mapsto (I - u\mathbf{P}_s)^{-1}$ on the left of $\Re s = 1$.

Many thanks for your attention