# CHANGING BASIS and DICHOTOMIC SEARCH

Brigitte Vallée

GREYC (CNRS and University of Caen)

Joint work with

Julien Clément, Dimitri Darthenay, Loïck Lhote

Journées AleaEnAmSud

Caen, June 6-7 2017

## Plan of the talk

I. Usual dichotomic search

II. Dichotomic search in the context of sources

III. Main result : statement, sketch of the proof

IV. Good sources

V. A more efficient algorithm?

# Usual dichotomic search

# Context

Classical analyses of sorting and searching algorithms deal with "keys" .

What is a "key" ? Not well specified.... Something undecomposable ...

The main operation : comparison between keys.

The cost of the comparison between two keys is a unitary cost.

The complexity of the algorithm is then the total number of key comparisons.

Now, keys are viewed as words – (decomposable) sequence of symbols.

The main operation : comparison between words,

and thus between symbols.

Then, the cost of the comparison between two words depends

on their coincidence = the length of their longest common prefix.

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

Coincidence = 3

Number of symbol comparisons = 4

The complexity of the algorithm is now the total number of symbol comparisons

## Average-case analysis of sorting and searching algorithms.

Then, the average-case analysis of sorting or searching algorithms studies the mean number of symbol comparisons performed by the algorithm.

Various sorting and searching algorithms are already analyzed.
Here, we give the dominant term of their average-case complexity:

| Algorithm | Mean number of key comparisons | Mean number of symbol comparisons |
|---|---|---|
| Quicksort | $n \log n$ | $(1/\mu) \cdot n \log^2 n$ |
| Insertion sort | $(1/4) \cdot n^2$ | $(1/4) \cdot \mathrm{E}[\gamma]\, n^2$ |

It involves characteristics of the source, the entropy $\mu$ or the coincidence $\gamma$.

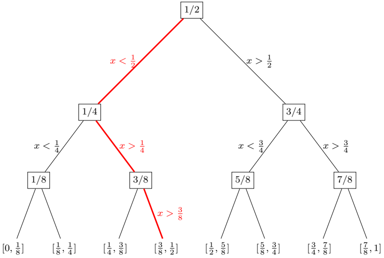
We consider here the Dichotomic Selection Algorithm,
and we wish to analyze it inside this more realistic framework.

# The (classical) dichotomic selection algorithm

An integer $L$ for the depth;
A binary tree of depth $L$ built on $\{v_k := k \cdot 2^{-L} \mid k \in [1..2^L - 1]\}$.
(The nodes $v_k$ are placed in the symmetric order in the tree)

Given $x \in [0, 1[$, find the interval $[v_k, v_{k+1}[$ that contains $x$.

Example with $L = 3, x = 2/5$




`Dicho`$(x, b, e)$.

***Input.*** A real $x \in [v_b, v_e]$;
***Output.*** The index $k$ s.t $x \in [v_k, v_{k+1}[$

**If** $b + 1 = e$ **then** return $b$;
$m := \lfloor (b + e)/2 \rfloor$;
**If** $x < v_m$
  **then** return `Dicho`$(x, b, m)$
  **else** return `Dicho`$(x, m, e)$.

With a tree of depth $L$, the number $K_L$ of key comparisons is (always) $K_L = L$.

# Dichotomic search in the context of sources

# Modeling with sources (I)

Now, a source produces words that encode the keys.

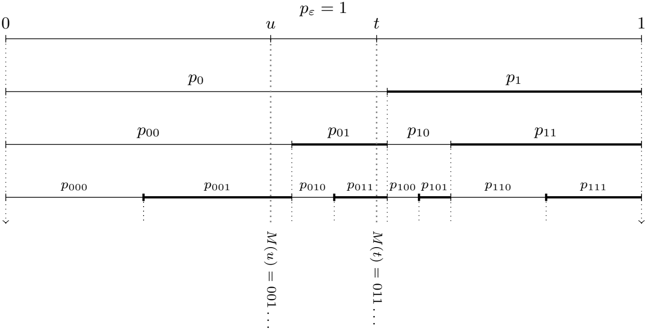
A source on the alphabet $\Sigma$ is defined by its set of fundamental probabilities

$$\{p_w \mid w \in \Sigma^\star\}, \qquad \text{where} \quad p_w := \Pr[\text{ a word begins with the prefix } w]$$

With an order on $\Sigma$, this defines the fundamental intervals $I_w$ of $w$

$$I_w := \{x \mid M(x) \text{ begins with } w\}.$$

and the tree of the ends of the fundamental intervals (TEFI)



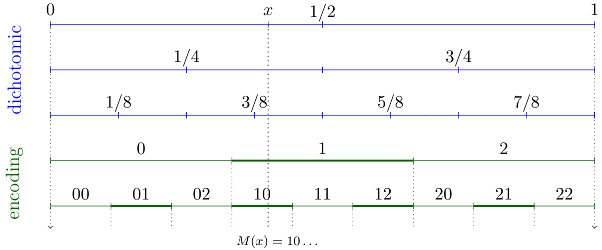

The process associates with $x$ an infinite word $M(x) \in \Sigma^{\mathbb{N}}$.

# Modeling with sources (II)

Here, we deal with two sources

- the dichotomic source $\mathcal{B}$ produces the nodes used for the dichotomy.
- the encoding source $\mathcal{M}$ produces the words that encode the keys

We have two types of fundamental intervals : the blue and the green ones.




Here, two regular sources : the blue one is binary and the green one is ternary.

# The dichotomic algorithm dealing with sources : Dicho-Source

The binary (dichotomic) source $\mathcal{B}$; an integer $L$ for the depth;

– the fundamental intervals of $\mathcal{B}$ of depth $\leq L$;

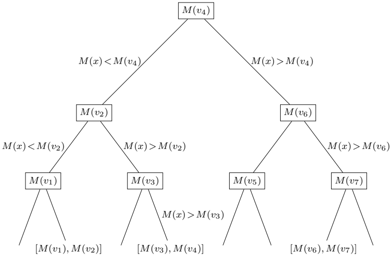
– the TEFI tree of their ends $v_k$ (in the symmetric order);

The encoding source $\mathcal{M}$ encodes the $v_k$ with the words $M(v_k)$;

We deal with the TEFI tree of the $M(v_k)$.

Given the coding $M(x)$ of $x \in [0, 1[$,

find the interval $[M(v_k), (M(v_{k+1})[$ that contains $M(x)$.




`Dicho-Source`$(M(x), b, e)$.

***Input.*** A word $M(x) \in [M(v_b), M(v_e)]$;

***Output.*** Index $k$ s.t $M(x) \in [M(v_k), M(v_{k+1})[$;

**If** $b + 1 = e$ **then** return $b$;

$m := \lfloor (b + e)/2 \rfloor$;

**If** $M(x) < M(v_m)$

**then** return `Dicho-Source`$(M(x), b, m)$

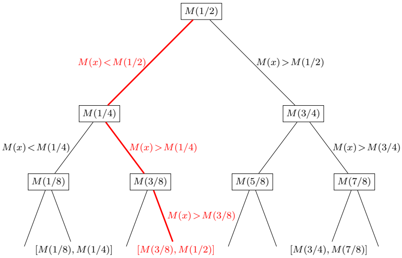
**else** return `Dicho-Source`$(M(x), m, e)$.

# An example of the execution of Dicho-Source

The dichotomic source is the regular binary source.

The encoding source is the regular ternary source

Execution with $x = 2/5$; $M(x) = 1012\ldots$




First step : $M(x)$ compared to $M(1/2)$ :
$M(x) = 10\ldots, \quad M(1/2) = 11\ldots$
Needs $S = 2$ proves $M(x) < M(1/2)$

Second step: $M(x)$ compared to $M(1/4)$:
$M(x) = 10\ldots, \quad M(1/4) = 01\ldots$
Needs $S = 1$ proves $M(x) > M(1/4)$

Third step: $M(x)$ compared to $M(3/8)$:
$M(x) = 1012\ldots, \quad M(3/8) = 1010$
Needs $S = 4$ proves $M(x) > M(3/8)$

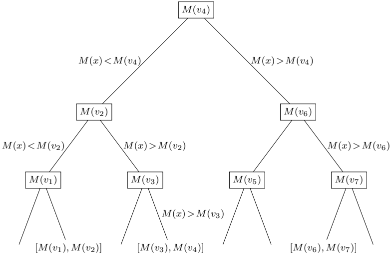
Finally $M(x) \in ]M(3/8), M(1/4)[$ with $S = 7$.

# Another point of view on the Dicho-Source algorithm

The binary (dichotomic) source $\mathcal{B}$; an integer $L$ for the depth;
the fundamental intervals of $\mathcal{B}$ of depth $\leq L$;
the tree of their ends $v_k$ (in the symmetric order);

The encoding source $\mathcal{M}$ encodes the $v_k$ with the TEFI of words $M(v_k)$;

Given the coding $M(x)$ of $x \in [0, 1[$,
find the interval $[M(v_k), (M(v_{k+1})[$ that contains $M(x)$.

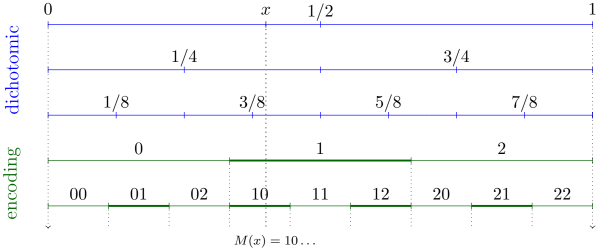



This means :

Compute the beginning of the word $B(x)$, namely the prefix of length $L$ of $B(x)$:

From $M(x)$, it computes $B(x)$

The "basis" changes !



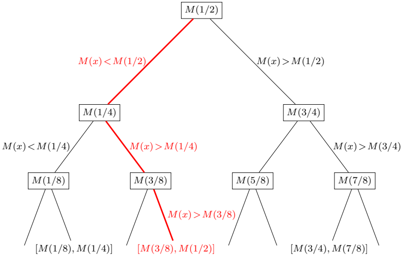
0
x
1/2
1
dichotomic
1/4
3/4
1/8
3/8
5/8
7/8
encoding
0
1
2
00
01
02
10
11
12
20
21
22
$M(x) = 10\ldots$

# An example of the execution of Dicho-Source

The dichotomic source is the regular binary source.

The encoding source is the regular ternary source

Execution with $x = 2/5$; the actual input is $M(x) = 1012\ldots$




First step : $M(x)$ compared to $M(1/2)$ :
$M(x) = 10\ldots, \quad M(1/2) = 11\ldots$
Needs $S = 2$ proves $M(x) < M(1/2) \to L = 0$

Second step: $M(x)$ compared to $M(1/4)$:
$M(x) = 10\ldots, \quad M(1/4) = 01\ldots$
Needs $S = 1$ proves $M(x) > M(1/4) \to L = 1$

Third step: $M(x)$ compared to $M(3/8)$:
$M(x) = 1012\ldots, \quad M(3/8) = 1010$
Needs $S = 4$ proves $M(x) > M(3/8) \to L = 1$

Finally $M(x) \in ]M(3/8), M(1/4)[$ with $S = 7$.

From $M(x) = 1012\ldots$ we have computed $B(x) = 011\ldots.$

# Our result

## Our main result

Consider two general entropic sources,

- a binary source $\mathcal{B}$, with entropy $\beta$, that implements dichotomy,
- an encoding source $\mathcal{M}$, with entropy $\mu$, that produces the words,

Consider the Source-Dichotomic Algorithm $\texttt{Dicho}(\mathcal{M}, \mathcal{B}, L)$ that

- deals with a random word produced by the encoding source $\mathcal{M}$
- uses the TEFI of the $\mathcal{M}$–coding of the $\mathcal{B}_L$ partition of the source $\mathcal{B}$

Then, if the sources are "good",
the mean number of symbol comparisons performed by Dicho is

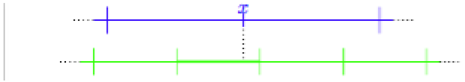
$$S_L = \frac{L^2}{2}\frac{\beta}{\mu} + O(L^\alpha) \quad \text{for } L \to \infty \quad (\alpha < 2 \text{ depends on the source..})$$

## What can be expected for basis changing ?

Two sources: The (input) source $\mathcal{M}$ and the (output) source $\mathcal{B}$;
Assume here (just for this slide) the two sources to be “completely known”.

Question: How many digits [say $k$] do we need on $M(x)$
to compute some digits [say $p$] of $B(x)$?

Answer : Compare the length of the fundamental intervals that contain $x$.



**Shannon-MacMillan-Breimann Theorem** : For two good sources,
the input source $\mathcal{M}$ with entropy $\mu$, the output source $\mathcal{B}$ with entropy $\beta$,
one has a.e $\quad |\log \ell_k(x)| \sim k \cdot \mu \qquad |\log b_p(x)| \sim p \cdot \beta$

$\Longrightarrow$ An information theory lower bound for any “basis changing” algorithm

$$\text{Number of symbols used } k \geq p \, \frac{\beta}{\mu}$$

# The main formula of the proof.

For $v$ at depth $p$,

$(a)$ the $(k+1)$-th symbols of $M(x)$ and $M(v)$ are compared $\Longleftrightarrow$ $x$ both belongs to the two fundamental intervals $B_p(v)$ and $I_k(v)$.

$(b)$ $\Pr[$the $(k+1)$-th symbols of $M(x)$ and $M(v)$ are compared$] = |B_p(v) \cap I_k(v)|$

The final formula $S_L$ is obtained by summing over

- the depth $p \leq L$ of the dichotomic tree
- the nodes $v$ at depth $p$ in the dichotomic tree $(v \in \mathcal{V}_p)$
- the depth $k$ of the encoding source

$$S_L = \sum_{p \leq L} S(p), \qquad S(p) = \sum_{k \geq 0} \sum_{v \in \mathcal{V}_p} |B_p(v) \cap I_k(v)|$$

or

$$S(p) = \sum_{v \in \mathcal{V}_p} \sum_{\substack{w \in \Sigma^\star \\ v \in I_w}} |B_p(v) \cap I_w|$$

## The length of the intersection $B_p(v) \cap I_k(v)$

Depends on – two depths $p$ for source $\mathcal{B}$, $k$ for source $\mathcal{M}$,
– nodes $v$ (the ends of the intervals for $\mathcal{B}$).

Two points of view on fundamental intervals or probabilities, via

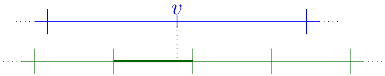
- the prefixes:
  $I_w := \{x \mid M(x) \text{ begins with the prefix } w\}, \quad p_w := |I_w|$
- the coincidence:
  $I_k(y) = \{x \mid \gamma(M(x), M(y) \geq k\}, \quad \ell_k(v) := |I_k(v)|$

## The length of the intersection $B_p(v) \cap I_k(v)$ (I)

Case when $\mathcal{B}$ is regular: Easier as $|B_p(v)| = 2^{-p}$ does not depend on $v$.

There are two cases for the triple $(k, v)$

(1) when $\ell_k(v) \leq 2^{-p}$, then $|B_p(v) \cap I_k(v)| \leq 2^{-p}$



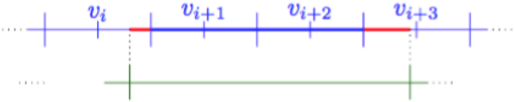

(2) when $\ell_k(v) > 2^{-p}$,

we consider the $v$'s that belong to a given interval $I_w$;

we let $B_p^w := \bigcup_{v \in I_w} B_p(v)$,

then: $I_w \setminus [B_p(v_-) \cup B_p(v_+)] \subset I_w \cap B_p^w \subset I_w$

$v_-$ and $v_+$ are the first nodes $v$ not to belong to $I_w$.




The case $(2)$ will provide the dominant term

# The length of the intersection $B_p(v) \cap I_k(v)$ (II)

Case when $\mathcal{B}$ is regular:

(1) Case $\ell_k(v) \leq 2^{-p}$. We use the geometric decreasing of $k \mapsto \ell_k(v)$.

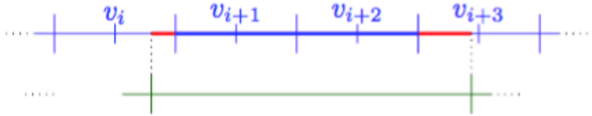
$$S_1(p) \leq 2^p \cdot \frac{2^{-p}}{1-a} \qquad \text{for some } a < 1$$

(2) Case $\ell_k(v) > 2^{-p}$.

We consider for $x = 2^{-p}$ the two sums $A(x)$ and $A(x) - xB(x)$ with

$$A(x) := \sum_{w \in \Sigma^\star} p_w [\![ p_w > x ]\!] \qquad B(x) := \sum_{w \in \Sigma^\star} [\![ p_w > x ]\!] .$$

and the estimates hold $\quad A(2^{-p}) - 2^{-p} B(2^{-p}) \leq S_2(p) \leq A(2^{-p})$




And now, when $\mathcal{B}$ is no longer regular ?

We use the "good distribution" of the lengths $|B_p(v)|$.

## What is here expected for a "good" source ? (I)

Two points of view on fundamental intervals/ probabilities, via

- the prefixes: $I_w := \{x \mid M(x)$ begins with the prefix $w\}, \;\; p_w := |I_w|$
- the coincidence: $I_k(y) = \{x \mid \gamma(M(x), M(y) \geq k\}, \;\; \ell_k(v) := |I_k(v)|$

Three main properties satisfied by the set of fundamental probabilities:

$(a)$ The set of $w \mapsto p_w$ or $y \mapsto \ell_k(y)$ is geometrically decreasing.

$$\exists a < 1, \forall v \in [0,1], \forall k \geq 0, \qquad \ell_{k+1}(y) \leq a\, \ell_k(y)\,.$$

$(b)$ A strong version of the SMMB Theorem (a central limit theorem).

The sequence $(y \mapsto \log b_p(y))$ asymptotically follows a Gaussian law, with a mean value $\mathbb{E}[\log b_p] \sim -p\beta$ (entropy $\beta$) and a variance $\mathbb{V}[\log b_p] \sim \sigma^2 p$,

$$\Pr\left[y \in \mathcal{I} \mid \frac{\log b_p(y) - p\beta}{\sigma\sqrt{p}} \leq A\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{A} e^{-x^2/2} dx + O\left(\frac{1}{\sqrt{k}}\right).$$

## What is here expected for a "good" source ? (II)

$(c)$ There is an estimate for the probability of the "most probable prefixes".

Consider $A(x) := \sum_{w \in \Sigma^\star} p_w [\![ p_w > x ]\!]$ $B(x) := \sum_{w \in \Sigma^\star} [\![ p_w > x ]\!]$.

and also $\widehat{A}(x) := A(x) - xB(x) = \int_x^1 B(t)dt$

Then, the two functions satisfy for $(x \to 0)$,

$(i)$ For a periodic source: $A(x) \sim \widehat{A}(x) = (1/\mu) \cdot |\log x| + P(\log x) + O(x^{-\gamma})$

$(ii)$ For an aperiodic source: $A(x) \sim \widehat{A}(x) \sim (1/\mu) \cdot |\log x|$

$(iii)$ There are aperiodic cases where we get information on the remainder term.

## Return to the analysis of Dicho-Source for good sources: main principles.

A "good" binary source of entropy $\beta$, a real parameter $\theta \in ]1/2, 1[$

With the Gaussian law for the lengths $\log b_p$,

with $D_p := \exp(-p\beta - p^\theta)$, $\qquad C_p := \exp(-p\beta + p^\theta)$

the following holds: $$\sum_{v \in \mathcal{V}_p} b_p(v) [\![ b_p(v) \notin [D_p, C_p] ]\!] = O(p^{-1/2})$$

(1) when $\ell_k(v) \le b_p(v)$, then [Geom. decreasing]

$$S_1(p) \le 1/(1-a) \cdot \sum_{v \in \mathcal{V}_p} b_p(v) \qquad \Longrightarrow \qquad S_1(p) \le 1/(1-a)$$

(2) when $\ell_k(v) > b_p(v)$ and $b_p(v) \in [D_p, C_p]$ then

[Most Prob. prefixes]+ [Gaussian Law]

$$A(D_p) - 2C_p\, B(D_p) \le S_2(p) \le A(D_p) \qquad \Longrightarrow \qquad S_2(p) \sim \frac{1}{\mu} |\log D_p|$$

(3) when $\ell_k(v) > b_p(v)$ and $b_p(v) \notin [D_p, C_p]$ then [Gaussian law]

$$S_3(p) \le K p \sum_{v} b_p(v) [\![ b_p(v) \notin [D_p, C_p] ]\!] \qquad \Longrightarrow \qquad S_3(p) \le K p^{1/2}.$$

The case (2) provides the dominant term $(\beta/\mu)\, p$

# More on Good Sources

## Sufficient conditions for a source to be Good

## What is here a “good” source ? (III)

As proven in [V01], there are close connections between

– probabilistic properties $(b), (c)$

– analytic properties of the Dirichlet generating function of the source,

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

$(b)$ : Via the Quasi-Power Theorem, there is a relation between

– a quasi-powers property for $\Lambda_k(s)$ (for $s$ close to 1)

– asymptotic Gaussian laws for $\ell_k(y)$

$(c)$ : Via the Mellin transform, there is a relation between

– the asymptotics for $A(x)$ and $\widehat{A}(x)$ (for $x \to 0$)

– the position of singularities of $s \mapsto \Lambda(s)$

that are close to the vertical line $\Re s = 1$

## Some instances for $\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s$

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 $\mathbf{R}$ of initial probabilities $(r_i)$

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

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

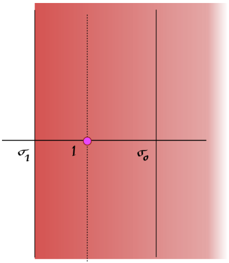
A general source, with its (pruned) transition matrix $\mathbf{P}_s$,

$$\Lambda(s) = {}^t\mathbf{E} \cdot (I - \mathbf{P}_s)^{-1}[\mathbf{1}] \quad \text{with} \quad {}^t\mathbf{E} := (1, 0, 0 \ldots)$$

For $(c)$, importance of a tameness region $\mathcal{R}$ for $\Lambda(s)$ near $\Re s = 1$ where.

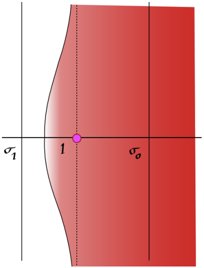
- $\Lambda(s)$ has a unique singularity: this is a simple pole located at $s = 1$
- $\Lambda(s)$ is of polynomial growth.

Possible tameness regions for a simple source (memoryless or Markov)



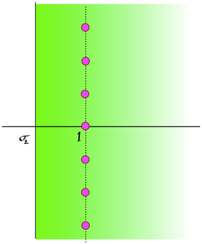

Situation 1

Vertical strip




Situation 2

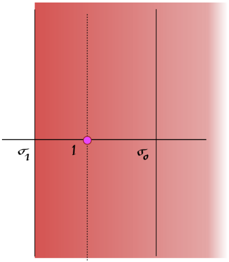
Hyperbolic region




Situation 3

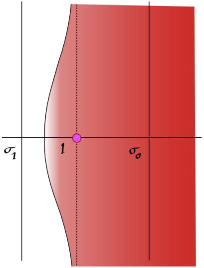
Vertical strip with holes

## Possible tameness regions for a simple source



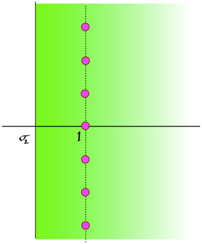

Situation 1

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 $(p_1, p_2, \ldots, p_r)$

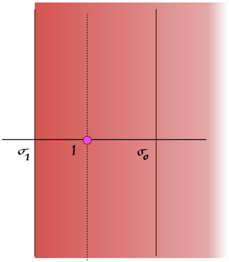
– S1 is impossible

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

– S2 occurs if there exists a ratio $\log p_i / \log p_j$

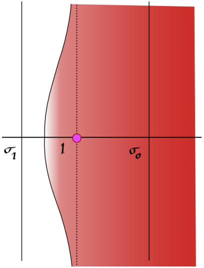
which is “diophantine” [badly approximable by rationals]

# Possible tameness regions for a simple source



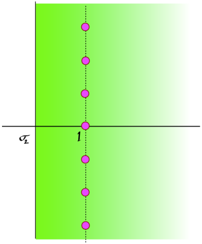

Situation 1

Vertical strip




Situation 2

Hyperbolic region




Situation 3

Vertical strip with holes

A close relation between the “shape” of a tameness region for $\Lambda(s)$

and the estimates of $\left[A(x) - \frac{1}{\mu}|\log x|\right], \quad \left[\widehat{A}(x) - \frac{1}{\mu}|\log x|\right], \quad (x \to 0)$

with $\quad A(x) := \sum_{w \in \Sigma^\star} p_w [\![p_w > x]\!] \qquad \widehat{A}(x) := \sum_{w \in \Sigma^\star} (p_w - x)[\![p_w > x]\!]\,.$

# A more efficient version of the algorithm

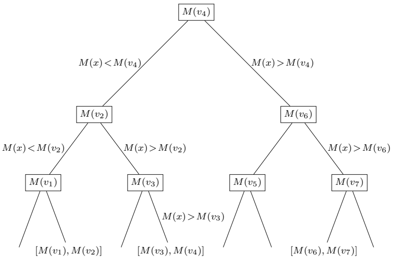
## A new version of the algorithm Dicho-Source?

Our version of the algorithm is very naive:

- It does not use the knowledge from previous comparisons between $M(x)$, $M(v_b)$, and $M(v_e)$
- it performs the comparison of words $M(x)$ and $M(v_m)$ from scratch

A possible improvement :

- Memorize at each step the two coincidences $\gamma(M(x), M(v_b))$ and $\gamma(M(x), M(v_e))$ .
- At the next step, begin the comparison of $M(x)$ and $M(v_m)$ at depth $\min(\gamma(M(x), M(v_b)), \gamma(M(x), M(v_e)))$ .




`Dicho-Source`$(M(x), b, e)$.

***Input.*** A word $M(x) \in [M(v_b), M(v_e)]$;
***Output.*** Index $k$ s.t $M(x) \in [M(v_k), M(v_{k+1})[$;

**If** $b + 1 = e$ **then** return $b$;
$m := \lfloor (b+e)/2 \rfloor$;
**If** $M(x) < M(v_m)$
  **then** return `Dicho-Source`$(M(x), b, m)$
  **else** return `Dicho-Source`$(M(x), m, e)$.

## Analysis of the new version of Dicho-Source where

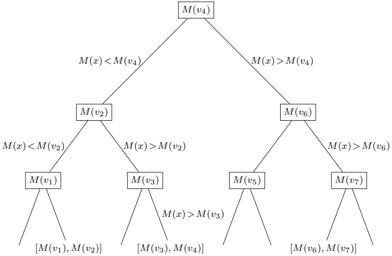
– we memorize at each step the two coincidences

$$\gamma_b := \gamma(M(x), M(v_b)) \text{ and } \gamma_e := \gamma(M(x), M(v_e))$$
$$\pi(b, e) := \min(\gamma_b, \gamma_e)$$

– at the next step,

we begin the comparison of $M(x)$ and $M(v_m)$ at depth $\pi(b, e)$

we recompute the new $\pi(b, e) = \min(\gamma_b, \gamma_m)$ or $\pi(b, e) = \min(\gamma_m, \gamma_e)$




Question: Estimate the difference between $\gamma_m$ and $\min(\gamma_e, \gamma_b)$

We aim to analyse this efficient version of the algorithm,
and estimate the mean number $\widehat{S}_L$ of symbol comparisons performed.
Is the complexity of the algorithm closer to the lower bound $\frac{\beta}{\mu} L$.

## Evolution of parameter $\pi(b,e) := \min(\gamma_b, \gamma_e)$.

For instance, in the previous execution with $M(x) = 1012...$

First step : $M(x)$ compared to $M(1/2)$

We begin at $\gamma_0 = 1, \quad \gamma_8 = 0, \quad \pi(0,8) = 0$

$M(x) = 10\ldots, \quad M(1/2) = 11\ldots \quad \to \gamma_4 = 1$

Needs $S = 2$ proves $M(x) < M(1/2) \to L = 0$

Now $\pi(0,4) = 0$

Second step: $M(x)$ compared to $M(1/4)$:

We begin at $\pi(0,4) = 0$

$M(x) = 10\ldots, \quad M(1/4) = 01\ldots \quad \to \gamma_2 = 0$

Needs $S = 1$ proves $M(x) > M(1/4) \to L = 1$

Now $\pi(2,4) = 0$

Third step: $M(x)$ compared to $M(3/8)$;

We begin at $\pi(2,4) = 0$.

$M(x) = 1012\ldots, \quad M(3/8) = 1010 \quad \to \gamma_3 = 3$

Needs $S = 4$ proves $M(x) > M(3/8) \to L = 1$

Now $\pi(3,4) = \min(\gamma_3, \gamma_4) = 1$

Finally $M(x) \in ]M(3/8), M(1/4)[$ with $S = 7$ ;