Digital Trees and Memoryless Sources: from Arithmetics to Analysis

Philippe Flajolet,
Mathieu Roux,
Brigitte Vallée

AofA 2010, Wien

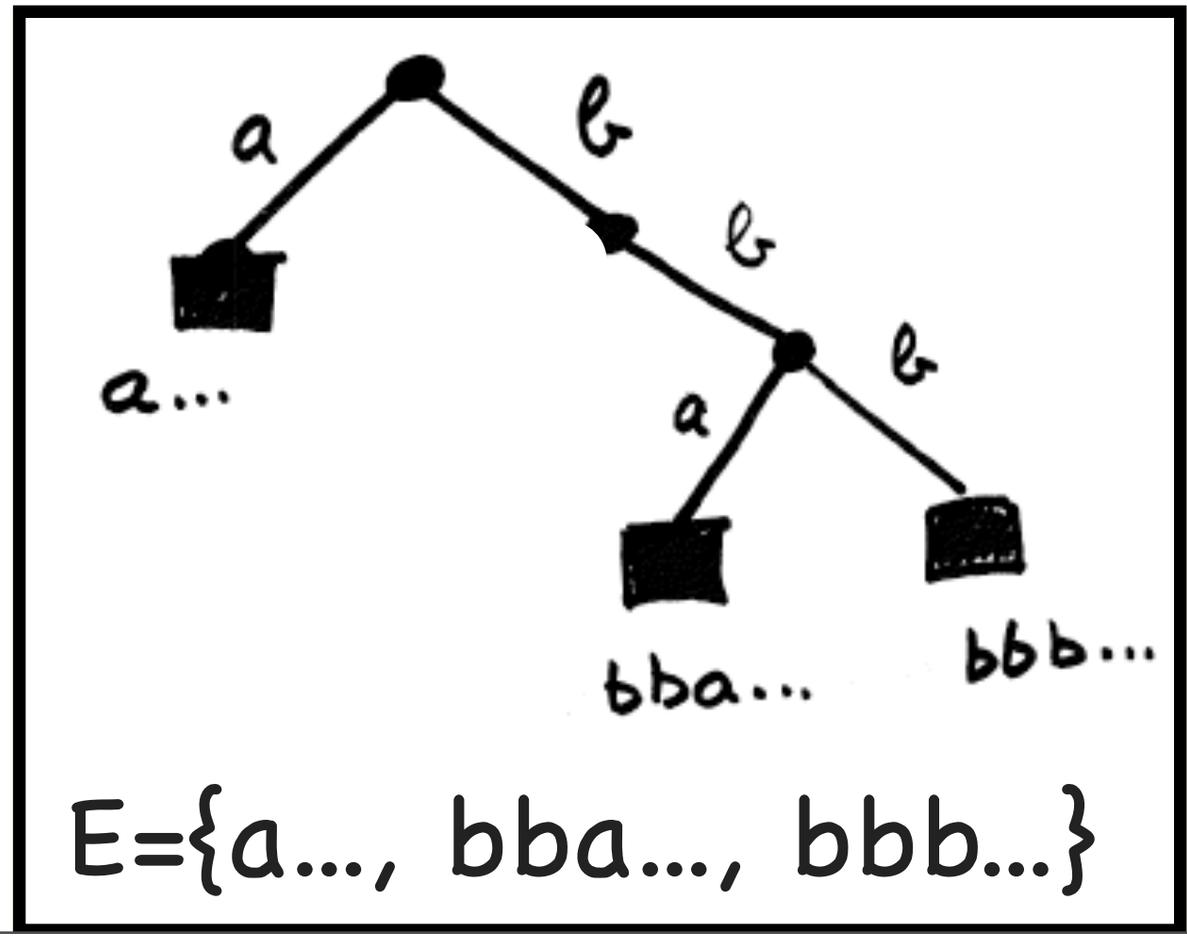
What is a digital tree, aka "TRIE"?

= a data structure for dynamic dictionaries

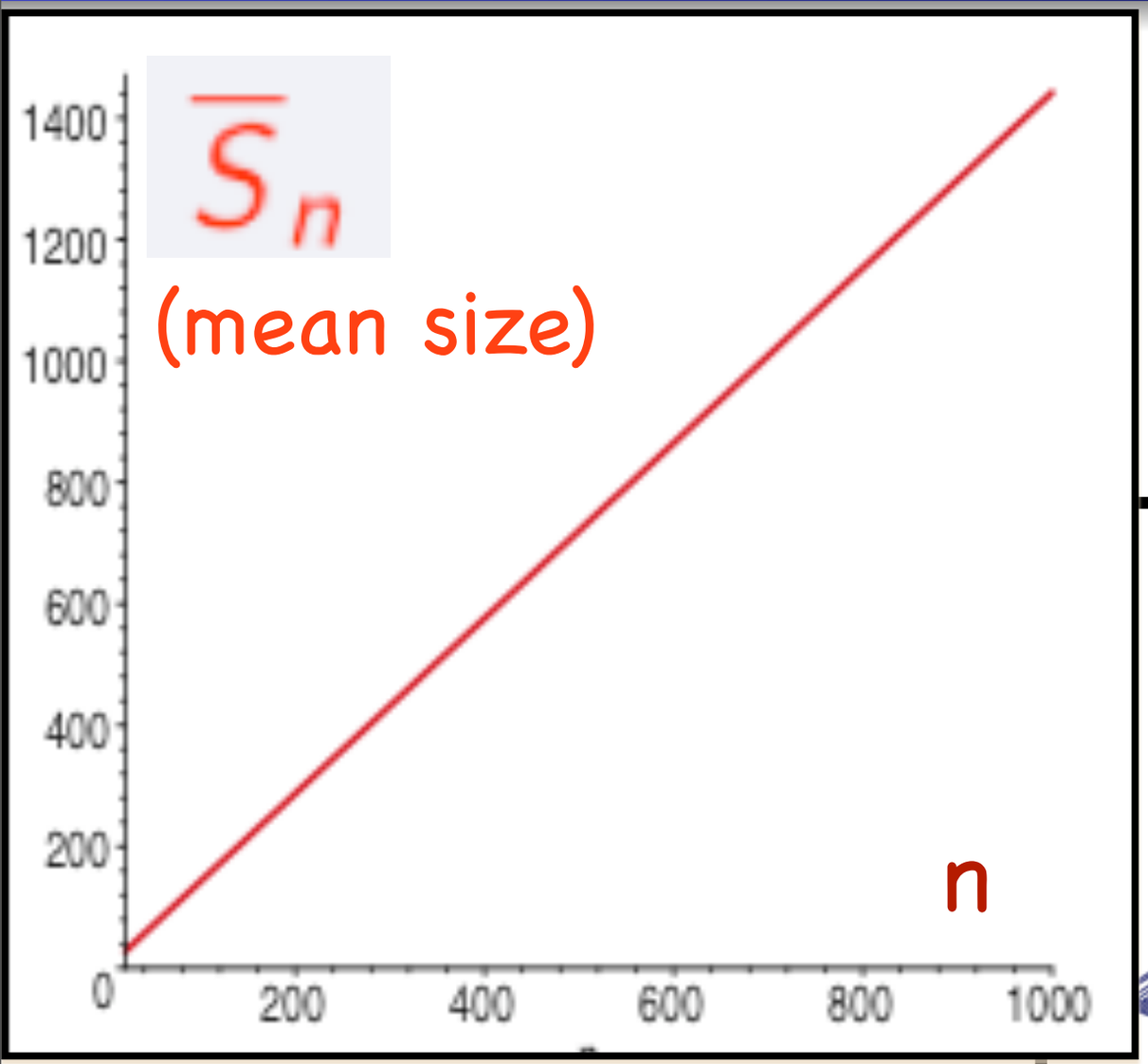
A = Finite alphabet
W = infinite sequences
E : W^n -> tree

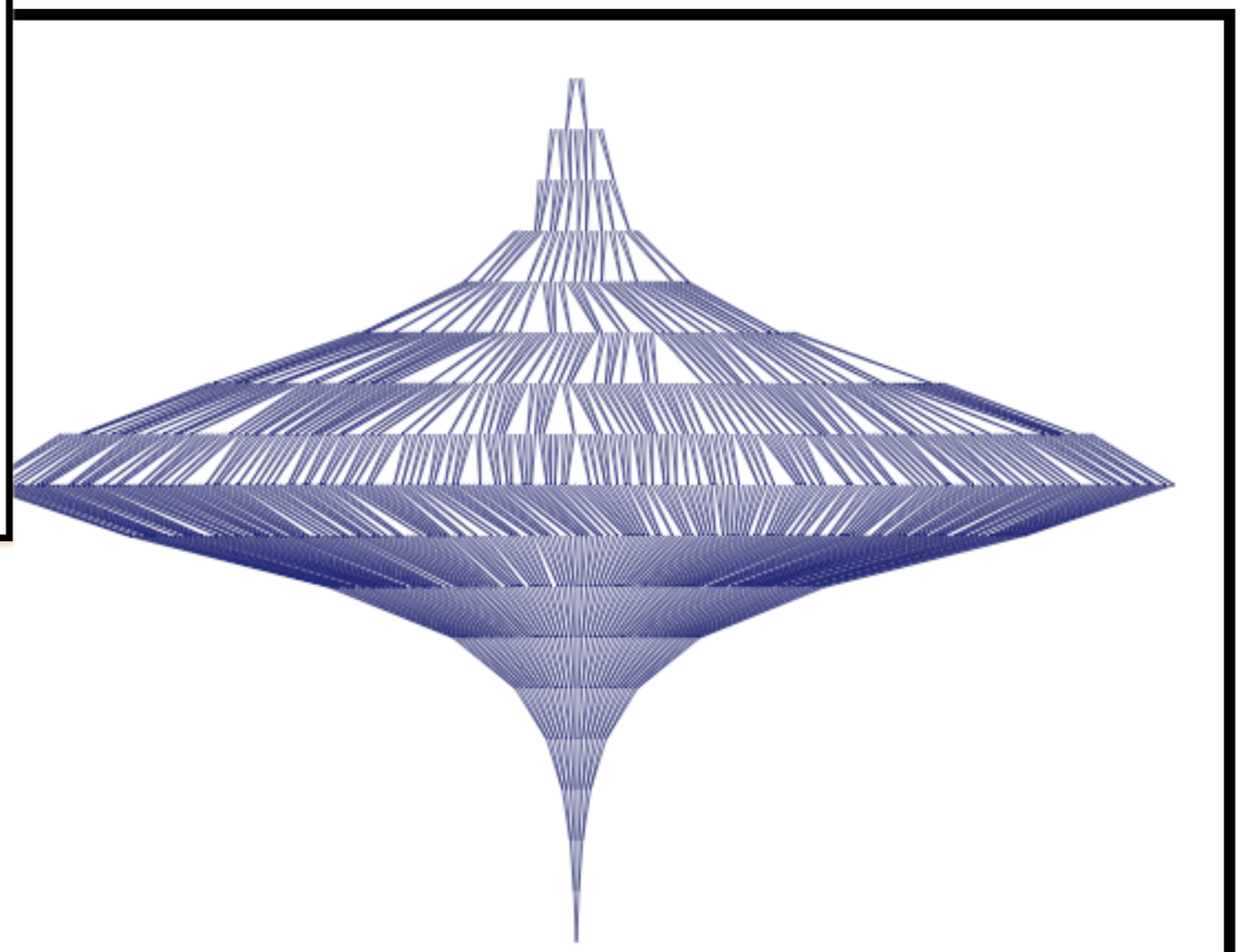
- TOP-DOWN construction: Set $\mathcal{E}$ is split into $\mathcal{E}_a$,...,$\mathcal{E}_z$, according to initial letter; continue with next letter; stop when elements are separated.

- INCREMENTAL construction: start with the empty tree and insert elements of E one after the other.



What does a trie look like?



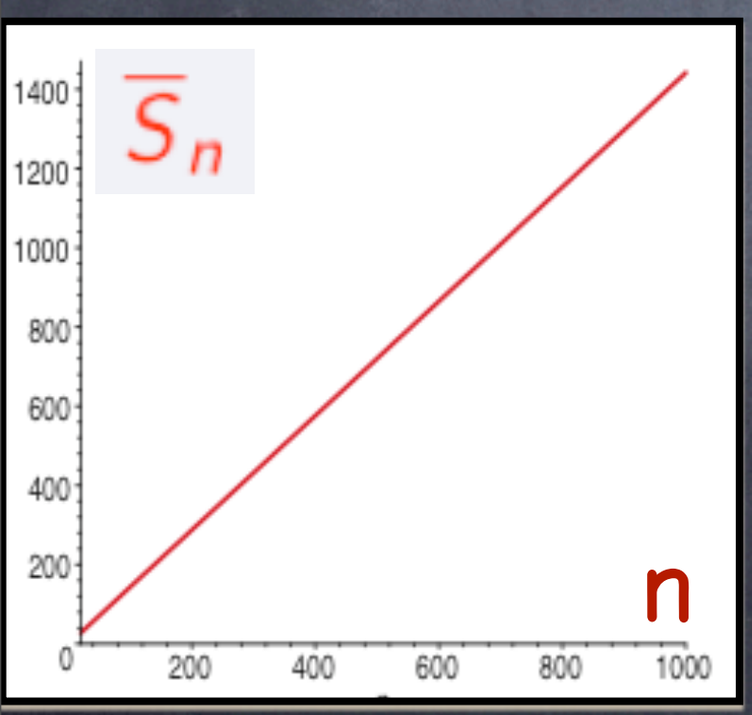


here: uniform data

A random trie on n=500 uniform binary sequences;
size =741 internal nodes; height=18

What does a trie look like?

- Expected size seems to be **asymptotically linear**.
- Convergence to asymptotic regime seems to be **fast**.



But...Things are not quite as they seem!

Probabilistic model: **Memoryless sources**

- A **finite alphabet** $\mathcal{A} = \{a_1, \ldots, a_r\}$.
- **Letters drawn independently** to form words from $\mathcal{W} = \mathcal{A}^\infty$:

$$\mathbb{P}(a_j) = p_j.$$

- **Words drawn independently**: model is $\mathcal{W}^n$.

- Want fixed number, n items, to build the trie.
- Often use $N = \mathcal{P}\text{oisson}(x)$ items:

$$\mathbb{P}(N = n) = e^{-x}\frac{x^n}{n!}.$$

Expect ($\pm$elementarily) $\mathcal{P}(x) \approx$ fixed-n, when $x \approx n$.

Poisson

Memoryless sources (Bernoulli)

- 1965: Knuth & De Bruijn analyse binary tries, with Pr(0)=Pr(1)=1/2, showing **oscillations**.

- 1973: Knuth discusses **biased bit models**, including golden-section case [Ex 5.2.2-53]

- 1986: Fayolle-F-Hofri exhibit **periodicity criterion**, extended by, e.g., Schachinger [2000]; Jacquet-Szpankowski-Tang [2001]

- 1990-2000: **Convergence to asymptotic regime** often wrongly assumed to be fast. Caveats by Schachinger (~2000).

- 2010; this paper: convergence to asymptotic regime is **very slow** and depends on **fine arithmetic properties of probabilistic model**.

The periodic case

Definition

The probability vector $(p_1, \ldots, p_r)$ is periodic if

— *all* ratios $\dfrac{\log p_j}{\log p_k}$ are rational. (E.g., $\dfrac{\log p_2}{\log p_1} \in \mathbb{Q}$; binary alph.)

Theorem (Periodic sources; folklore)

$$H = -p_1 \log p_1 - \cdots - p_r \log p_r$$

Expected size $\overline{S_n}$ *is, with* Φ *a smooth periodic function:*

$$\overline{S}_n = \frac{n}{H} + n\Phi(\log n) + O(n^{1-A}), \qquad A > 0.$$

$\Longrightarrow$ Oscillations $(O(n))$, plus good error term.

- These cases are **exceptional**: the p_j are *algebraic numbers*. Such families are a denumerable set; hence have measure 0.

The aperiodic case (main result)

Definition

The probability vector $(p_1, \ldots, p_r)$ is aperiodic if

— at least one ratio $\dfrac{\log p_j}{\log p_k}$ is irrational. (E.g., $\dfrac{\log p_2}{\log p_1} \notin \mathbb{Q}$; binary a.)

Theorem (Aperiodic sources; this paper)

$$H = -p_1 \log p_1 - \cdots - p_r \log p_r$$

Expected size $\overline{S_n}$ *is, for* **"diophantine sources"** *(generic case)*

$$\overline{S}_n = \frac{n}{H} + O\left(n \exp(-\sqrt[\theta]{\log n}) \right), \qquad \theta > 1.$$

This is better than $n/(\log n)^a$*, any* a*; much worse than* $n^{1-\epsilon}$*, any* ϵ*.*

- *For remaining* **"Liouvillean sources"** *(rare), error term can come arbitrarily close to* $o(n)$*.*

$\Longrightarrow$ No oscillation, but poor error term.

- This case is **generic**: it has has measure 1.

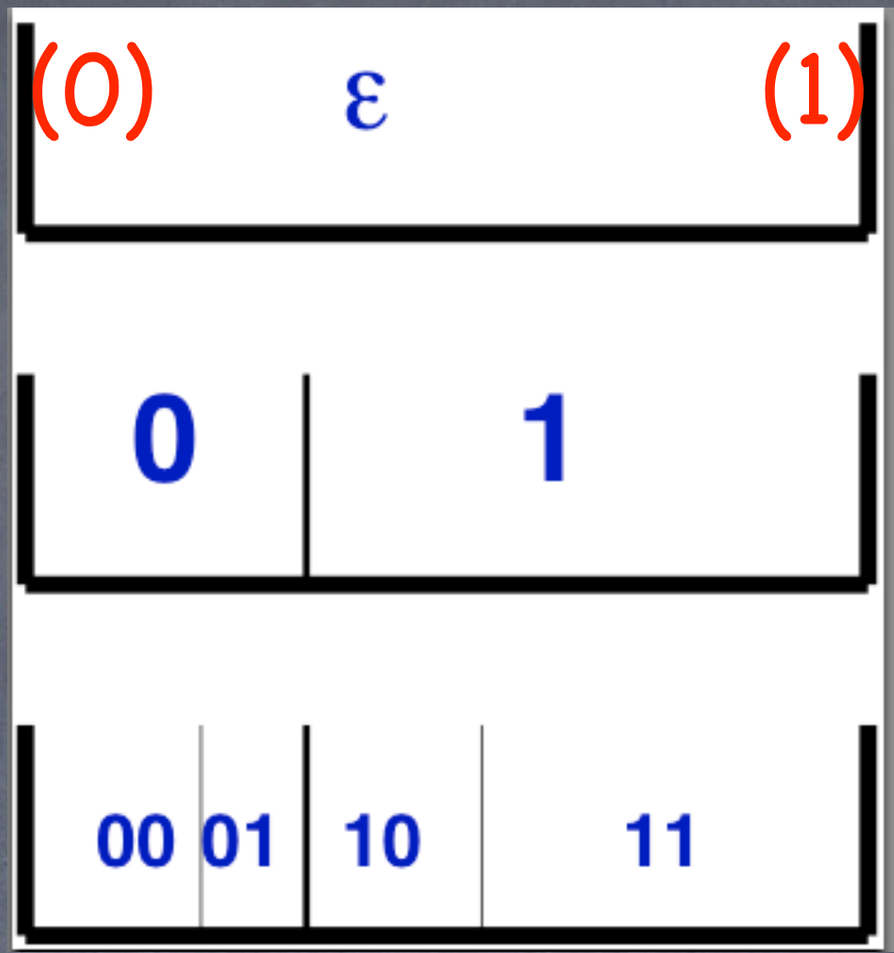
1. Basics

- Fundamental intervals
- + Mellin
- = Formal analysis

[Vallée 1997++]

- View source model in terms of fundamental intervals:

$w \to p_w$



- Size = Number of places occupied by at least two prefixes

- Mellinize ->...

$$\begin{cases} E_{\mathcal{P}(x)}[\text{Size}] = \sum_{w \in \mathcal{A}^\star} g(p_w x) \\ g(x) = 1 - (1+x)e^{-x}. \end{cases}$$

$$g(x) \equiv \mathbb{P}_{\mathcal{P}(x)}(Z \geq 2)$$

The Mellin transform

$$f(x) \overset{\mathcal{M}}{\rightsquigarrow} f^\star(s) := \int_0^\infty f(x) x^{s-1}\, dx$$

(It exists in strips of $\mathbb{C}$ determined by growth of $f(x)$ at $0, +\infty$.)

Property 1. Factors *harmonic sums*:

$$\sum_{(\lambda,\mu)} \lambda f(\mu x) \overset{\mathcal{M}}{\rightsquigarrow} \left(\sum_{(\lambda,\mu)} \lambda \mu^{-s} \right) \cdot f^\star(x).$$

Property 2. Maps asymptotics of f on singularities of $f^\star$:

$$f^\star \approx \frac{1}{(s - s_0)^m} \quad \Longrightarrow \quad f(x) \approx x^{-s_0} (\log x)^{m-1}.$$

Proof of $\mathbf{P_2}$ is from Mellin inversion + residues

Singularities?

Lambda(s)

$$\begin{cases} E_{\mathcal{P}(x)}[\text{Size}] = \sum_{w \in \mathcal{A}^\star} g(p_w x) \\ g(x) = 1 - (1+x)e^{-x}. \end{cases}$$

Harmonic sum!

$$\begin{cases} S^\star(s) = -(s+1)\Gamma(s)\Lambda(s) \\ \Lambda(s) := \sum_w p_w^{-s} \end{cases}$$

Singularities?

- Geometry of the poles of

$$\Lambda(s) = \frac{1}{1 - p_1^s - \cdots - p_r^s}$$

2. Geometry of poles

- Poles are associated with **simultaneous approximations** to **logs of probabilities**
- Distinguish:
  -- Diophantine = badly approximable (generic);
  -- Liouvillean = unusally well approximable (rare)

Poles of $\Lambda(s)$ near $\Re(s) = 1$

$$\Lambda(s) = \frac{1}{1 - p_1^{-s} - \cdots - p_r^{-s}}$$

- Look for s: $\boxed{p_1^s + p_2^s = 1}$, $s = \sigma + it$.

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

Implies $p_1^{it} \approx 1$ *and* $p_2^{it} \approx 1$; i.e., $t \approx \dfrac{2\pi}{\log p_1} q_1$ *and* $t \approx \dfrac{2\pi}{\log p_2} q_2$.

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

Pole of $\Lambda(s)$ $\Longrightarrow$ **"good" rational approximation** to $\dfrac{\log p_2}{\log p_1}$.

For general $(p_1, \ldots, p_r)$, must have a **common denominator** q_1:

$\forall j : q_1 \dfrac{\log p_j}{\log p_1}$ is a near-integer.

Poles of $\Lambda(s)$ near $\Re(s) = 1$

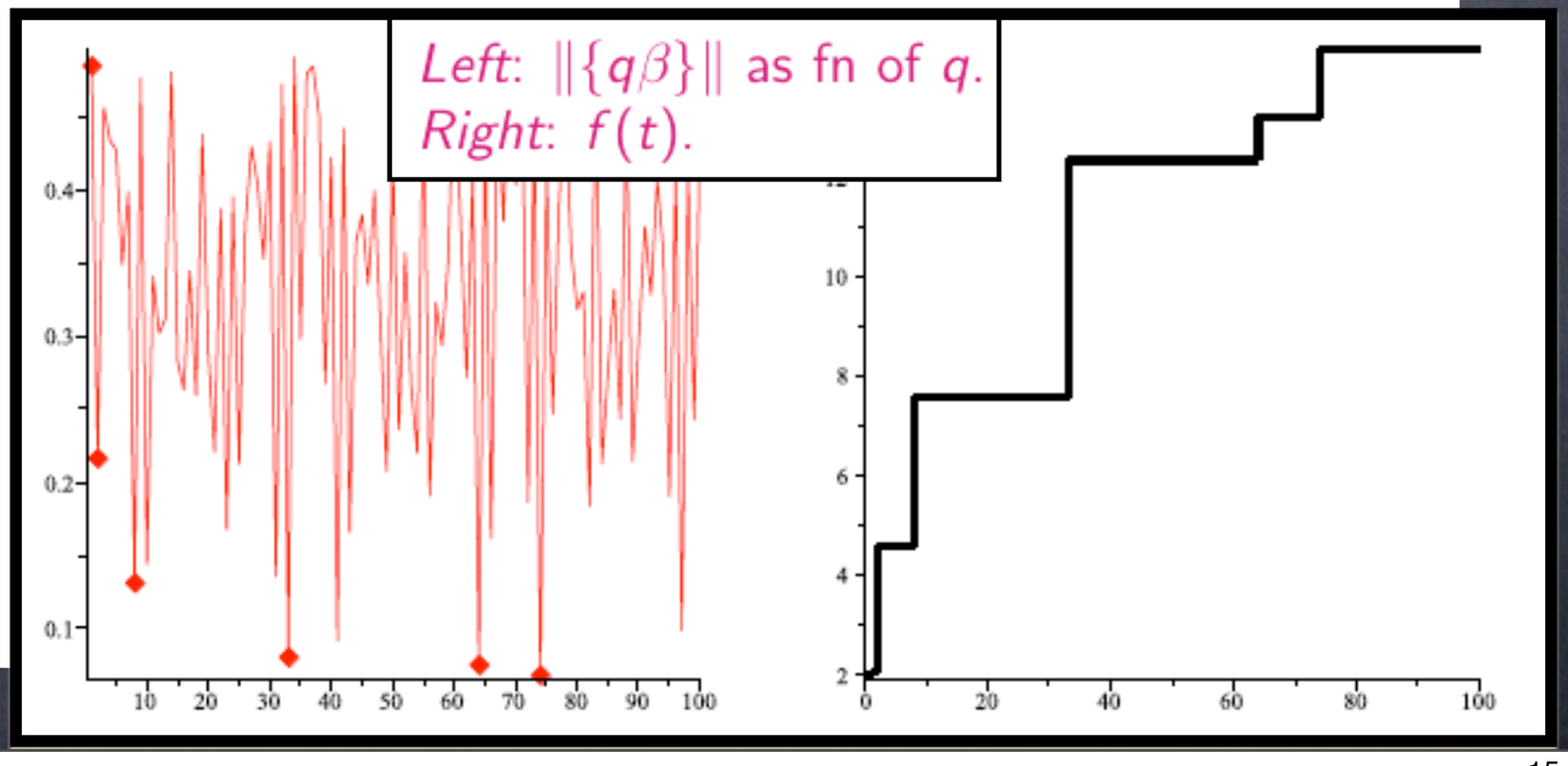
$\beta = (\beta_1, \ldots, \beta_r) \in \mathbb{R}^r$; fix a norm $\|\cdot\|$ on $\mathbb{R}^r$.

$\{x\}$ = centred fractional part; $\|\{\beta\}\|$ is distance to nearest integer lattice point.

*Look at "***record***" approximants; measure* **quality** *by* $f(t)$.

Definition

- Q is a **Best Simultaneous Approximant Denominator** (BSAD), if $\|\{Q\beta\}\| < \|\{q\beta\}\|$, for all $q < Q$.
- $f(t)$, the **approximation function**, is staircase and $f(t) = \frac{1}{\|\{Q^-\beta\}\|}$. , if Q^-, Q^+ are the BSADs that frame t. Thus:



Basic trichotomy

For a probability vector $(p_1, \ldots, p_r)$:

- **Periodic sources** (All ratios of logs are in $\mathbb{Q}$)
- **Aperiodic sources** (Some ratios $\notin \mathbb{Q}$):
  - **Diophantine**: approximation function $f(t)$ is **polynomial**; optimal exponent is known as *irrationality measure*;
  - **Liouvillean**: approximation function $f(t)$ is **superpolynomial**.

— Scalars $\pi, e, \tan(1), \sqrt[3]{2}, \zeta(3), \log 5, \ldots$ are Diophantine.
Logs of rational and algebraic numbers are Diophantine.
Also numbers with bounded continued fraction quotients, ...

— Numbers with very fast-converging sums, e.g., $\sum 2^{-2^n}$, are Liouvillean.

Theorem

If $(p_1, \ldots, p_r)$ *is Diophantine, zeros are well-separated from* $\Re(s)$*:*

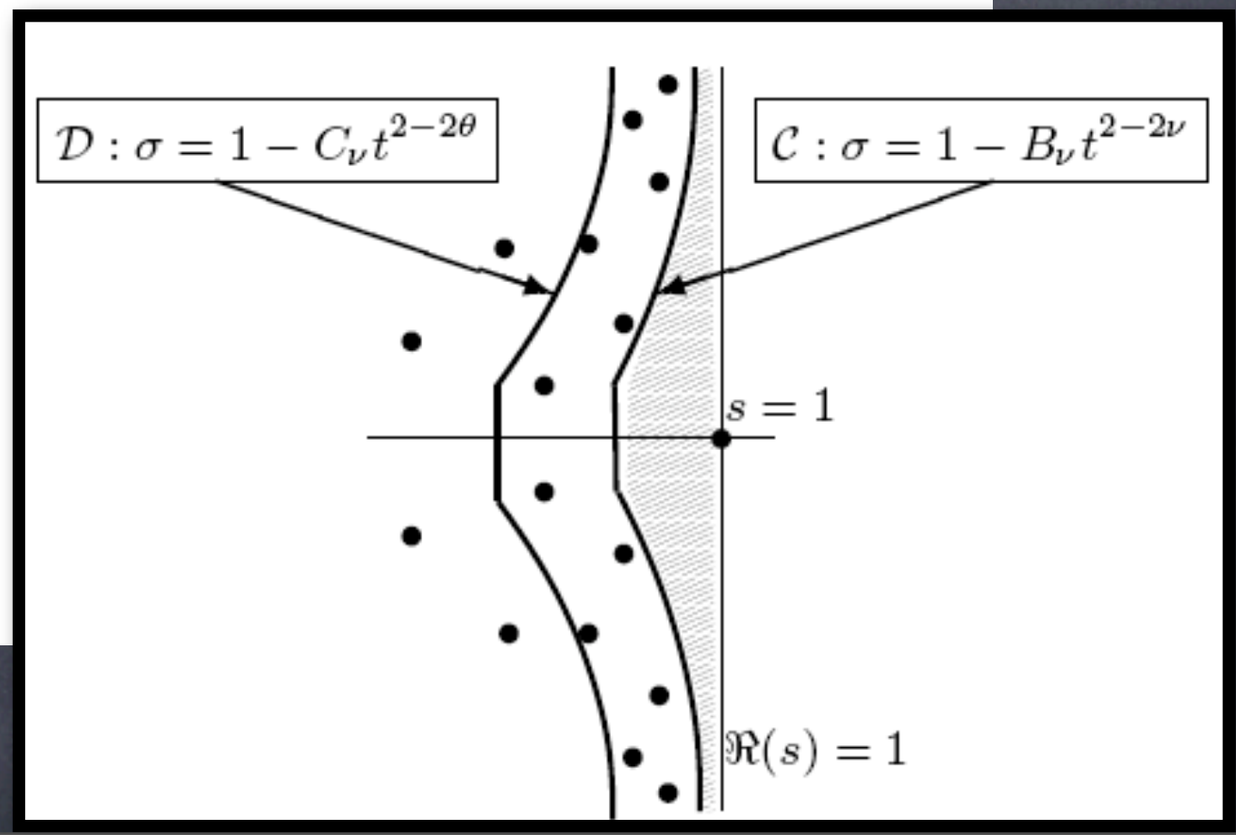
- **All** *zeros are to the left of a pseudo-hyperbola;*
- **Infinitely many zeros** *are to the right of a pseudo-hyperbola.*

Theorem

If $(p_1, \ldots, p_r)$ *is Liouvillean, zeros come closer to* $\Re(s) = 1$*:*

- **All** *zeros are to the left of a curve* $1 - 1/F_-(t)$*;*
- **Infinitely many zeros** *are to the right of* $-1 + 1/F_+(t)$*.*

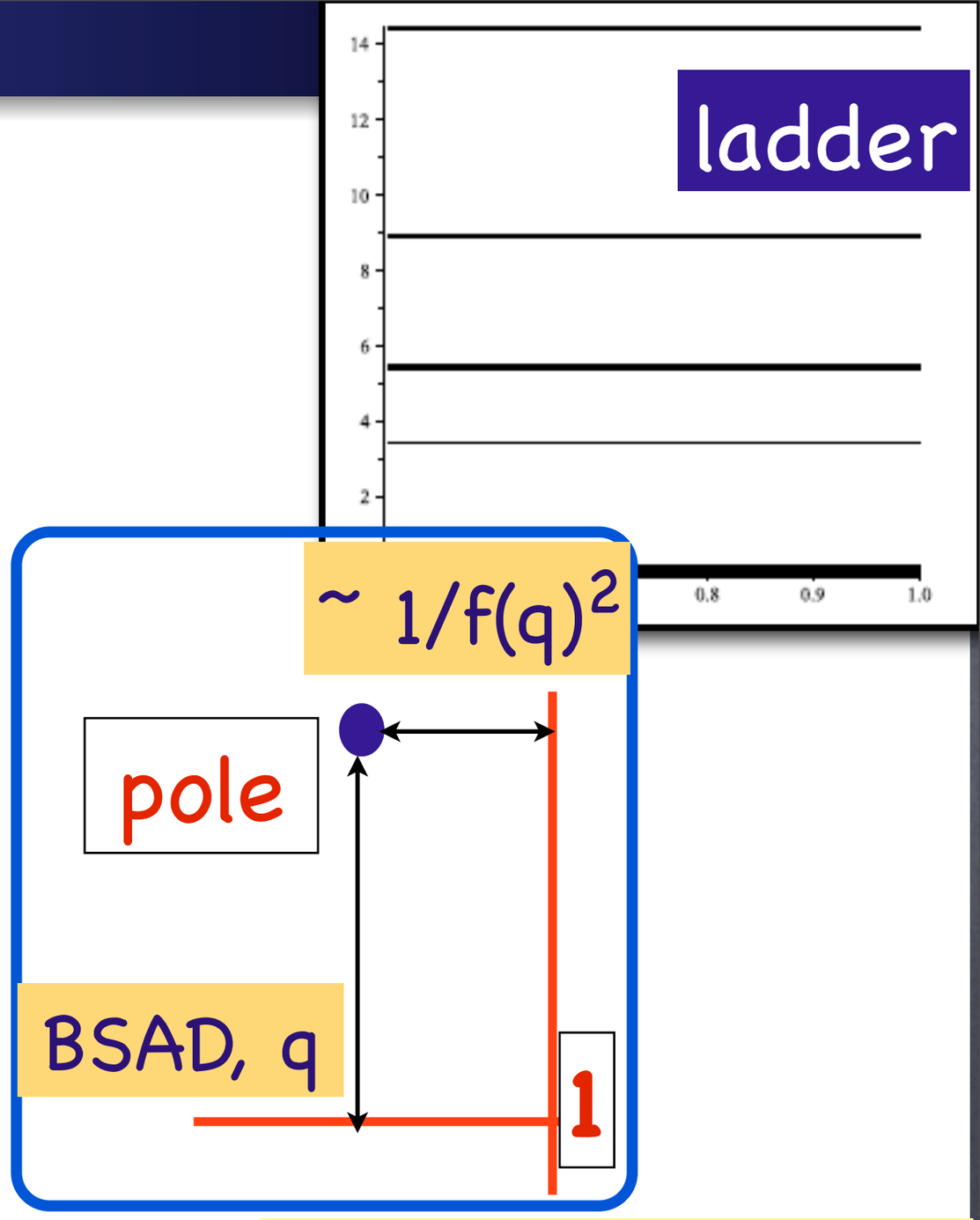
$F_-(t), F_+(t)$ *are dictated by approximation functions of* $(\log p_j)/(\log p_k)$*.*



Proofs

• **Pole** of $\Lambda(s)$ $\Longrightarrow$ **"good" rational approximation** to $(\log p_j)(\log p_k)$.

— Follow sketch above and develop properties of **"ladders"**.



• **"Good" rational approximation** to $(\log p_j)/(\log p_k)$ $\Longrightarrow$ **Pole** of $\Lambda(s)$.

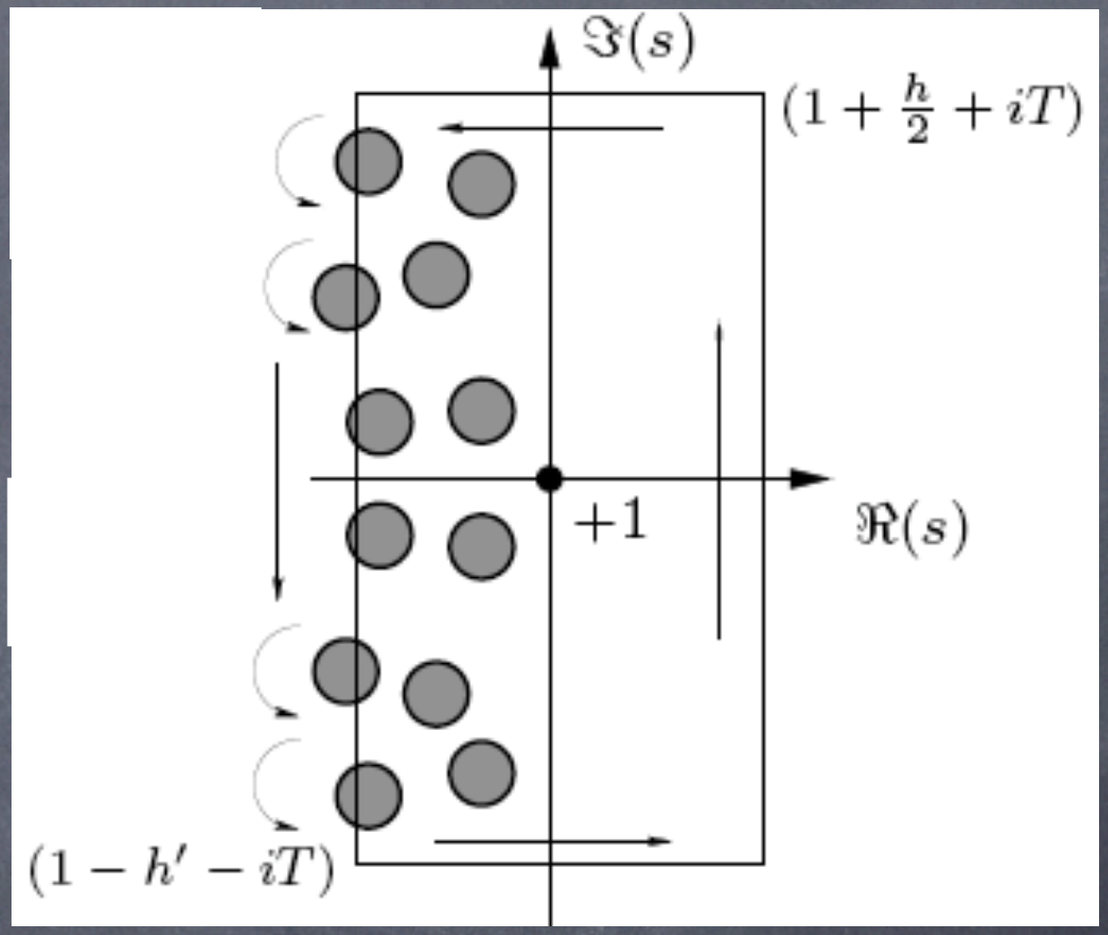
— use analytic, multivariate **Implicit Function Theorem**, $\Re(s) \approx 1$; $u_j \approx 0$:

$$1 - p_1^s p_1^{iu_1} - \cdots p_r^s p_r^{iu_r} = 0.$$

++ Lapidus, van Frankenhuijsen

3. Inverse Mellin analysis

- Make use of integration contour that avoids poles
- Estimate global contribs: pole-free region matters
- Poles are well-separated



4. Tries and QuickSort

- Applies to size of tries
- & almost anything that contains **Lambda(s).**

$$\Lambda(s) = \frac{1}{1 - p_1^s - \cdots - p_r^s}$$

- **Diophantine** => error terms are exp-of-root-of-log
- **Liouvillean** => error terms are o(n) and very close to O(n)

Theorem

Consider aperiodic **Diophantine probabilities** *with irrationality exponent* μ.

$$\begin{cases} \textit{trie size;} & \overline{S}_n = \dfrac{n}{H} + \mathbf{n\Phi(n)} \\ \textit{trie pathlength:} & \overline{S}_n = \dfrac{1}{H} n \log n + Cn + \mathbf{n\Phi(n)} \\ \textit{symbol-cost, Quicksort:} & \overline{S}_n = \dfrac{2}{H} n \log^2 n + Cn \log n + C' n + \mathbf{n\Phi(n)}, \end{cases}$$

where **error term** *is, for any* $\theta > \mu$*:*

$$\mathbf{\Phi(x)} = O\left(\exp\left[-(\log n)^{1/\theta} \right] \right),$$

Makes precise or improves on results of
Clément, Fill, Flajolet, Jacquet, Janson, Szpankowski, Vallée,...

Source models

- **memoryless**
  - periodic: good error terms
  - aperiodic: generally (very) bad error terms (us!)
    - Diophantine versus Liouvillean
- **Markov**; cf Szpa+Jacquet+Tang: similar (?)
- **dynamical**: Vallée + Cl-F-Vallée; cf Dolgopyat, B-V.
- **general**: à la Vallée-Clément-Fill-F.

Numerics

- (Proved for Poisson; transfers to fixed-size)
- Initial oscillations often not seen **numerically,** **for small n;** but they matter **asymptotically**