Average internal profiles, from tries to suffix-trees

Pierre Nicodème

CNRS - LIX, École polytechnique

and Project Algorithms, INRIA

Approximate pattern matching and the Jokinen-Ukkonen lemma

Def: q-gram any word of fixed size q

Edit operations over strings

- substitution ($l_1 \to l_2$) $aabdd \to aadcc$
- insertion ($| \to l$) $aa|dd \to aaecc$
- suppression ($l \to |$) $aaedd \to aa|cc$

Edit distance $\delta(S_1, S_2)$ between two strings S_1 and S_2

- minimum number of edit operations transforming S_1 into S_2

Jokinen-Ukkonen 1991 (loose version)

if $|S_1| = m$ and $\delta(S_1, S_2) \leq k$, then at least $m + 1 - (k+1)q$ of the $m - q + 1$ k-words of S_1 occur in S_2

Example

$$S_1 = aaabaaab$$

$$S_2 = aaacaaaa$$

$m = 8, \quad \delta(S_1, S_2) = 2$

$$2 - \text{grams}(S_1) = \{\{aa, aa, ab, ba, aa, aa, ab\}\}$$

$$Q_{S_1,S_2} = 2 - \text{grams}(S_1) \text{ present in } S_2 = \{\{aa, aa, aa, aa\}\}$$

Jokinen-Ukkonen

$$|Q_{S_1,S_2}| \geq m + 1 - (\delta + 1)q$$

$$4 \geq 8 + 1 - (2 + 1)2 = 3$$

Beware of the asymmetry: $|Q_{S_2,S_1}| = 5$

Application

When searching a pattern with errors in a text, slide over the text a window of same size as the pattern and discard windows which do not contain enough k-words of the pattern

Long term aim of this work

We would like to find a statistical indicator based on common k-words to two sequences to infer sequence similarity.

Short term aim

Repeated k-words in one sequence and common k-words to two sequences share many statistical common features.

$\Longrightarrow$ we analyse here the statistics of repeated k-words in a random sequence

Probabilistic model

We consider random strings generated by a Bernoulli model.

$$\Pr(X = 1) = p, \qquad \Pr(X = 0) = q, \qquad q \leq p$$

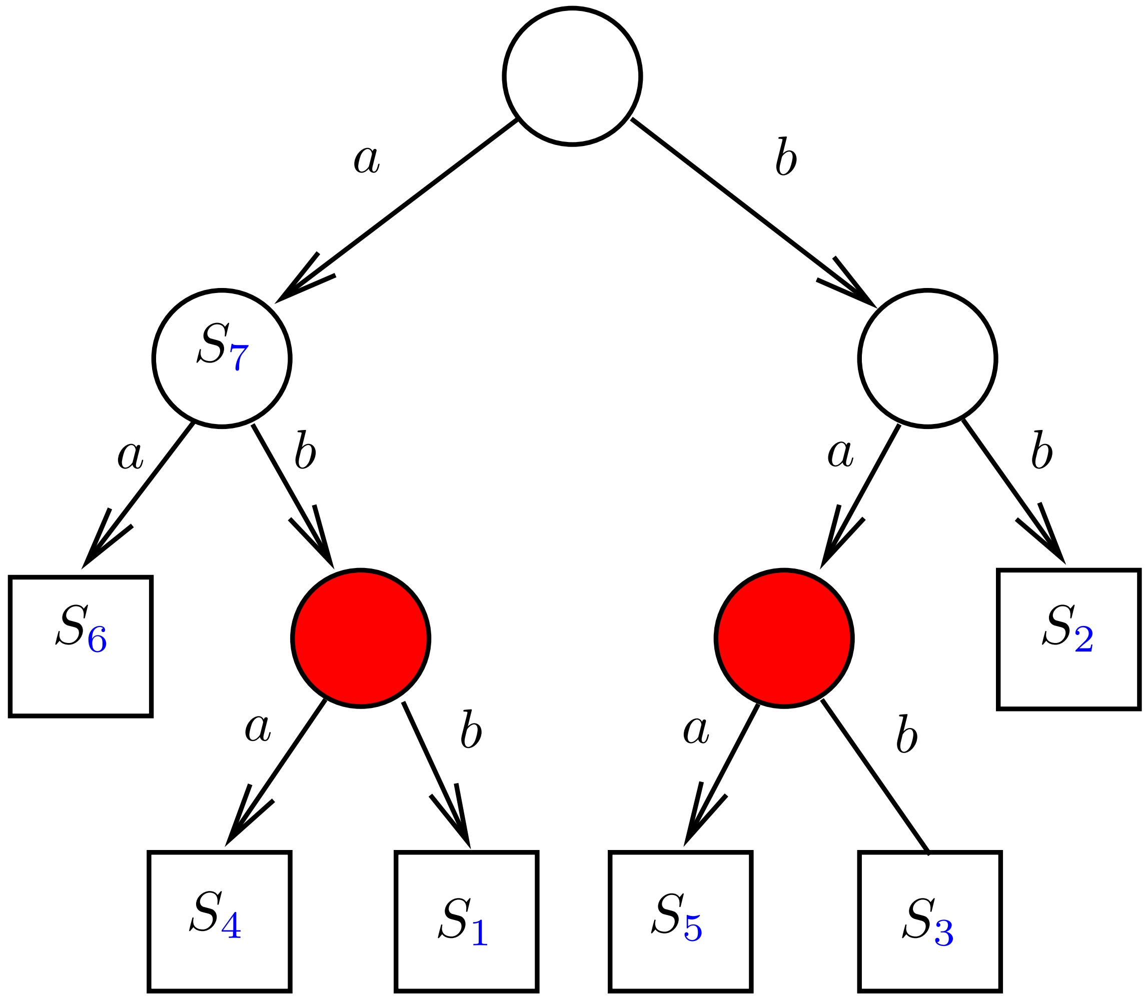
Repeated k-words in a sequence: an example

number of k-words occurring at least twice, without counting multiplicities

$$S = aaaabaaaabbb, \qquad k = 2$$

$$R_{\text{repeated}} = \{aa, ab, bb\} \qquad |R| = 3$$

Tries and Suffix-trees



trie built with keys

$aa.., aba.., abb.., baa.., bab.., bb..$

also suffix-tree built

over sequence $S = abbabaa$

$S = \underset{1234567}{abbabaa} \qquad k = 2$

$S_{\text{repeated}} = \{ab, ba\}$

$|S_{\text{repeated}}| = 2 =$ number of internal nodes at depth k

A heuristic approach

Suffix-tree - Dependent model

abbaaaababb ... aaabbbaaaab

abbaaaaba ...
bbaaaaba ...
baaaaba ...
aaaaaba ...
...

Trie - Independent model

aab
bba
aaa
bab
...

sequence length $l = n + q - 1$ $\Rightarrow$ n k-words

1. analyse the independent trie model
2. compare with the dependent suffix-tree model

Previous work

– Szpankowski and Jacquet - 1994

  asymptotically, the distributions of path lengths of suffix-trees and of tries of same size are equal

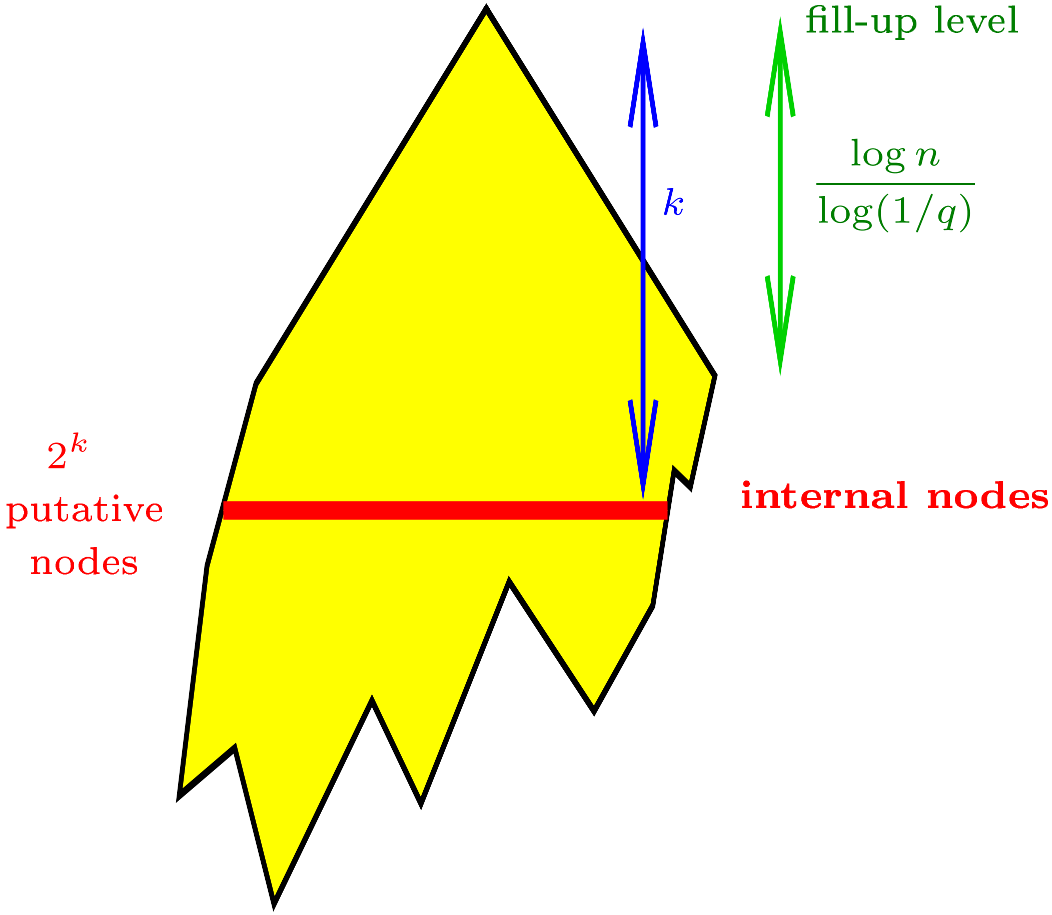
– J. Fayolle - 2002, 2004

  more precise result for the expectation

– Park and Szpankowski - 2005

  asymptotic profile of tries; expectation, standard deviation, distribution

Trie - fill-up level



Profiles expectation - Plan of the talk

I) Non-asymptotic analysis

1. use an urn model with 2^k terms to count the number of nodes at depth k.
2. compute the expectation for the trie as a sum of 2^k terms.
3. do simulations for the suffix-tree and compare

II) Trie Asymptotic analysis

– get asymptotic expectation for the trie by Mellin transform

III) Suffix-tree Asymptotic analysis

– bound asymptotically the difference of expectations of the trie and the suffix-tree

Repeated k-words

Equivalent problems

Input: an alphabet Σ ($|\Sigma| = s$), an integer k, a random sequence S of size $n + k - 1$

Dependent model

1. number of repeated k-words
2. number of internal nodes at depth k of the suffix-tree build on S
3. number of self-intersections of a random walk of length n over the de Bruijn graph $B(s, k)$

Independent model

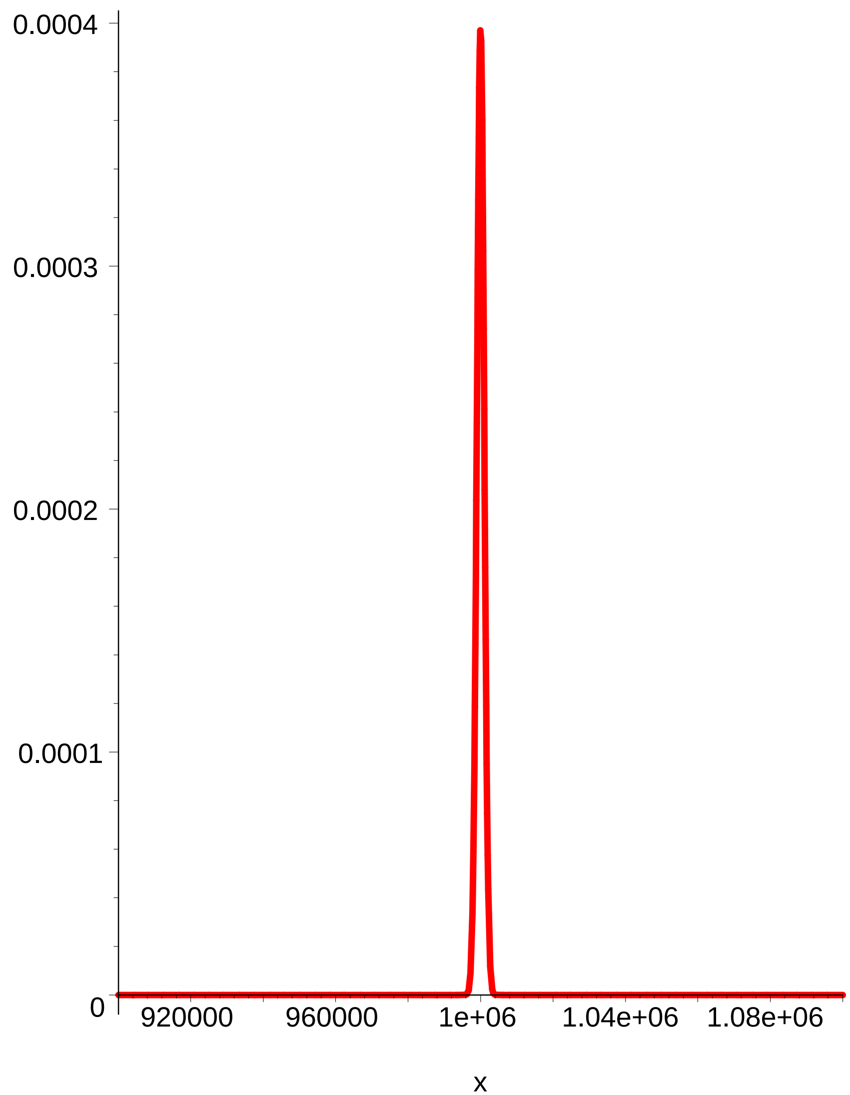
1. number of repeated k-words
2. number of internal nodes at depth k of a trie build with n random keys over Σ
3. number of self-intersections of a random walk of length n over a complete graph $K(s^k)$
4. number of urns containing more than one ball in a system of s^k urns in which n balls are thrown

Part I

Non-asymptotic analysis

Poissonization

do not consider exactly n objects (balls in the urns), but a random number of objects following a Poisson distribution $\mathcal{P}_\nu$ of parameter ν.



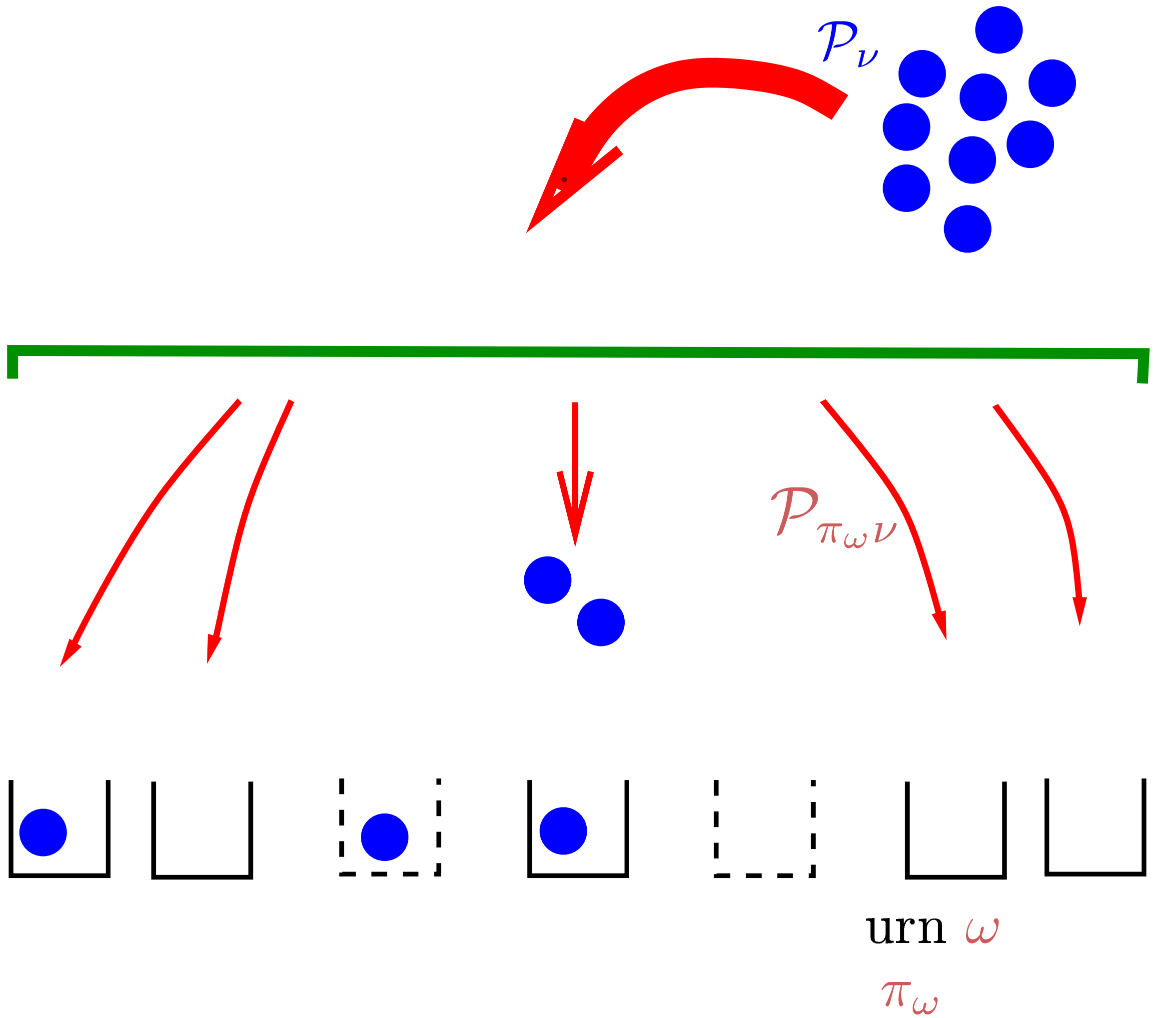
Plot of $\mathcal{P}_{1000000}$ $\qquad \sigma = \sqrt{1000000} = 1000$

Poissonization

internal nodes at depth k in a binary trie $\equiv$ system of 2^k urns

internal node $\omega \equiv$ urn ω contains at least 2 balls

do not throw exactly n balls in the urns, but throw a random number of balls following a Poisson distribution $\mathcal{P}_\nu$ of parameter ν.



The urns behave independently of each other

Poissonization - Depoissonization

$f_1(u), f_2(u), \ldots, f_n(u), \ldots$

Poisson transform of the sequence

$$\Phi(\nu, u) = \sum_{n \geq 0} f_n(u) \frac{\nu^n}{n!} e^{-\nu}$$

Algebraic easy depoissonization (when it works)

$$f_n(u) = [\nu^n] n! e^{\nu} \Phi(\nu, u)$$

Poisson model - Bivariate generating function

$\mathcal{P}_\nu$ balls in the system $\Rightarrow$ $\mathcal{P}_{\pi_\omega \nu}$ balls in urn ω

$$Y_\omega = \begin{cases} 1 & \text{if at least two balls in urn } \omega \\ 0 & \text{elsewhere} \end{cases}$$

$Z = \sum_{|\omega|=k} Y_\omega$

number of internal nodes

$$Y_\omega(u) = e^{-\pi_\omega \nu} \left(1 + \pi_\omega \nu + u \left(\frac{(\pi_\omega \nu)^2}{2!} + \frac{(\pi_\omega \nu)^3}{3!} + \dots \right) \right)$$

$$Z(u) = \prod_{|\omega|=k} Y_\omega(u) = \prod_{|\omega|=k} (1 + \pi_\omega \nu) e^{-\pi_\omega \nu} + u \left(1 - (1 + \pi_\omega \nu) e^{-\pi_\omega \nu} \right)$$

$$\mathbf{E}(Z) = i_{k,\mathcal{P}}(\nu) = \left. \frac{\partial Z(u)}{\partial u} \right|_{u=1} = \sum_{|\omega|=k} 1 - (1 + \pi_\omega \nu) e^{-\pi_\omega \nu}$$

Fixed n model - Trie expectation

Trie with number of keys following a Poisson model of parameter ν

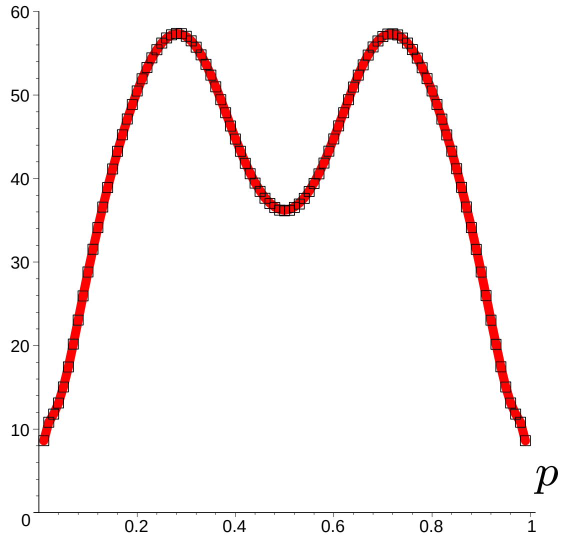
$$\mathbf{E}(Z_\nu) = i_{k,\mathcal{P}}(\nu) = \left.\frac{\partial Z(u)}{\partial u}\right|_{u=1} = \sum_{|\omega|=k} 1 - (1 + \pi_\omega \nu) e^{-\pi_\omega \nu}$$

"fixed model", exactly n keys

$$\mathbf{E}(Z_n) = [\nu^n] n! e^\nu \mathbf{E}(Z_\nu) = \sum_{|\omega|=k} 1 - (1 - \pi_\omega)^n - n\pi_\omega (1 - \pi_\omega)^{n-1}$$

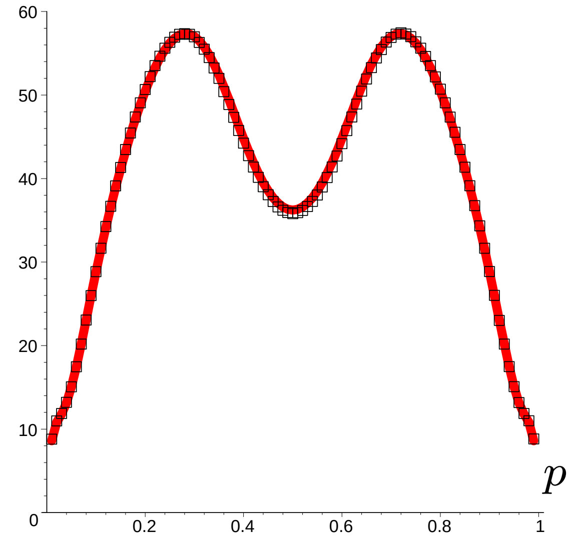
Experimental comparisons

$\mathbf{E}$(internal nodes at depth k)



trie

$\mathbf{E}$(internal nodes at depth k)



suffix-tree

$$\Sigma = \{0,1\} \quad p = \Pr(1) = 1 - \Pr(0) \quad \omega \in \Sigma^k \quad \pi_\omega = \Pr(\omega)$$

$n = 300$ keys $\quad k = 10$

red curves: theoretical curve $R(p)$ for the trie

dots: simulations

$$R(p) = \sum_{|\omega|=k} 1 - (1-\pi_\omega)^n - n\pi_\omega(1-\pi_\omega)^{n-1}$$

Cost of summations

$\Sigma = \{1, 2, 3, 4\}$, $s = |\Sigma|$ $\qquad m = s^q$

group urns by families of urns with equal probability

$|w| = q$, $\quad |w_i| = q_i$ number letters equal to i,

$q = q_1 + q_2 + q_3 + q_4$

population of $(q_1, q_2, q_3, q_4) = \dfrac{q!}{q_1!q_2!q_3!q_4!}$

Number of families $C_{q,s}$ (cost of summation)

$$\begin{aligned} C_{q,s} &= \text{compositions with } s \text{ summands } \geq 0 \text{ of } q \\ &= \text{compositions with } s \text{ summands } > 0 \text{ of } q+s \end{aligned}$$

$$C_s(z) = \left(\frac{z}{1-z}\right)^s \qquad C_{q,s} = [z^{q+s}]\left(\frac{z}{1-z}\right)^s = \binom{q+s-1}{s-1}$$

$C_{q,2} = q + 1$ $\qquad$ ADN: $C_{10,4} = 286$ $\qquad$ Proteins: $C_{3,20} = 1540$

Computing the moments

The values of q_1 to q_{i-1} have been computed previously when Procedure Calcsum is entered and $d = s - i$.
$s = |\Sigma|$ and q are handled as global constants.

Procedure Calcsum (f, d, n, ϕ)**:**

$i = s - d$

$u = \sum_{k=1}^{i-1} q_k$

If $d > 1$ **Then**
 For j **To** $s - u$ **Do**
  $q_i = j$
  $f =$ **Calcsum**$(f, d-1, n, \phi)$
 End of for
Else
 $q_s = q - \sum_{k=1}^{s-1} q_k$

 $f = f + \frac{q!}{q_1! q_2! \ldots q_s!} \phi(\theta_{q_1,\ldots,q_s}, n)$

End of if
Return (f)
End of procedure

$\theta_\xi = \theta_{q_1,\ldots q_s} = n \times \omega_1^{q_1} \omega_2^{q_2} \ldots \omega_s^{q_s}$
$\phi_1 = \left(e^{-\theta_\xi}(1+\theta_\xi) + \frac{1}{2n} e^{-\theta_\xi} \theta_\xi^2 (1-\theta_\xi)\right)$ $\mu_n = m -$ **Calcsum**$(0, s, n, \phi_1)$

Profile asymptotics comparisons - Method

$\sum_{|\omega|=k}, \; Cn^{\zeta}/\sqrt{\log n}$

Trie $\mathcal{P}_n$ keys (Poisson model)

$\Delta = n^{-\lambda_1}$ $\Delta = n^{-\lambda_2}$

Trie n keys Suffix-Tree n keys $\sum_{|\omega|=k}$

Analysis

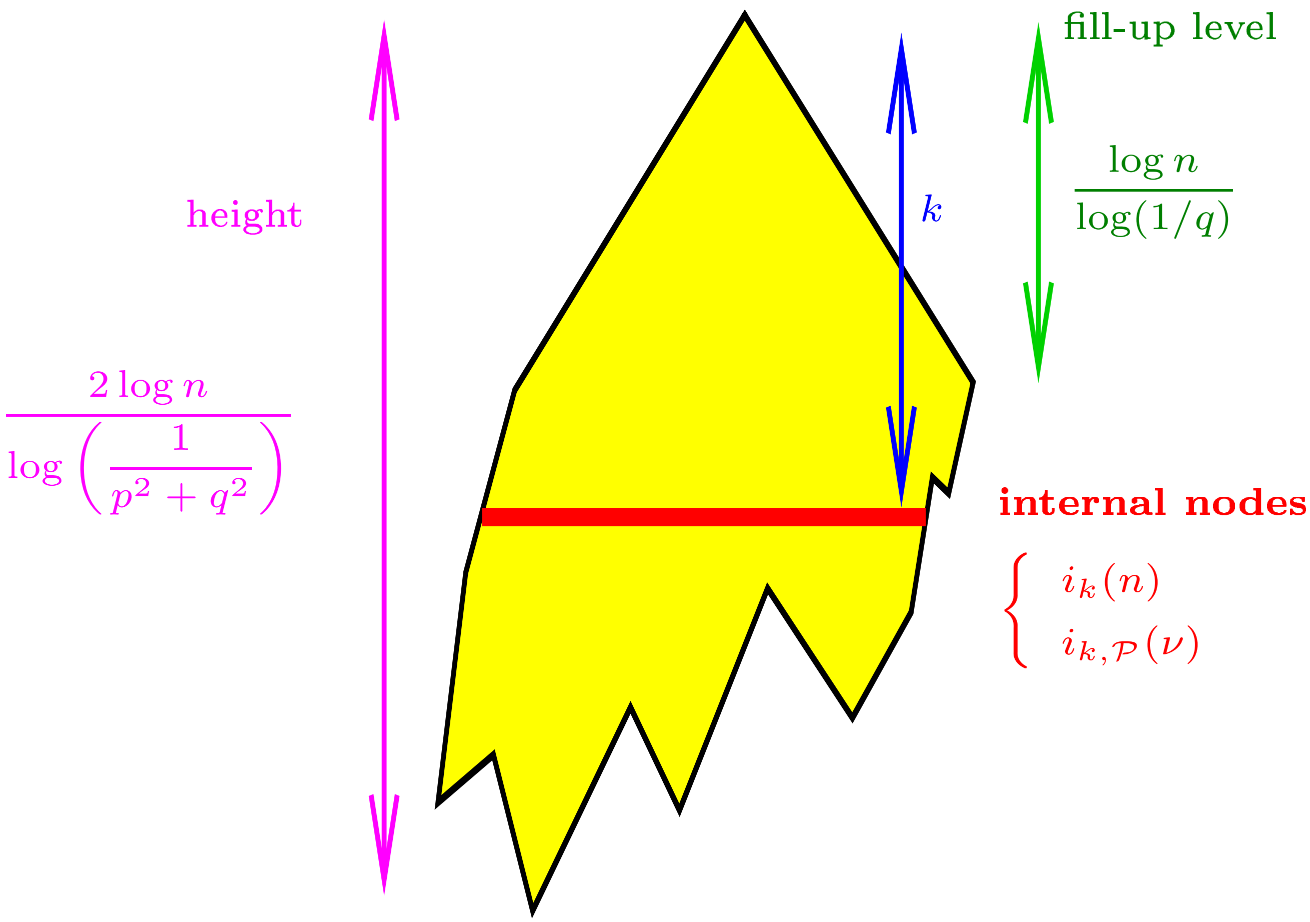
1. Trie Poisson - get rid of the sum by Mellin; inverse Mellin by saddle-point method
2. Trie - evaluate Δ between Poisson and fixed
3. evaluate Δ between Trie Poisson and Suffix-tree fixed

Part II

Trie - Asymptotic analysis

Profile of trie and suffix-tree

$$\begin{cases} n \text{ keys exactly} \\ \text{Poisson of parameter } \nu \text{ keys } - \ \mathcal{P}_\nu \end{cases}$$



Mellin transform

$$\Gamma(s) = \int_{x=0}^{\infty} e^{-x} x^{s-1} dx, \qquad \Gamma(n+1) = n!$$

Replacing e^{-x} by a function $f(x)$ gives the Mellin transform of f

$$\mathcal{M}[f(x); s] = \int_{x=0}^{\infty} f(x) x^{s-1} dx$$

$\langle \alpha, \beta \rangle$ = largest open strip of complex numbers $s = \sigma + it$ such that $\alpha < \sigma < \beta$ and $\mathcal{M}[f(x); s]$ is defined

$$f(x) \underset{x \to 0+}{=} O(x^u), \qquad f(x) \underset{x \to +\infty}{=} O(x^v), \qquad u > v$$

$$\Longrightarrow \mathcal{M}[f(x); s] \textbf{ exists} \text{ in the strip } \langle -u, -v \rangle$$

Inverse Mellin transform

$$\phi(s) = \int_{x=0}^{\infty} f(x) x^{s-1} dx \quad \Longleftrightarrow \quad f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \phi(s) x^{-s} ds \quad c \in \langle -u, -v \rangle$$

Sketch of proof (in one direction)

$$\begin{aligned}
&\frac{1}{2\pi i} \int_0^{\infty} x^{s-1} \int_{c-i\infty}^{c+i\infty} \phi(z) x^{-z} dz \\
&= \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} \phi(z) dz \int_0^1 x^{s-z-1} dx + \frac{1}{2\pi i} \int_{b-i\infty}^{b+i\infty} \phi(z) dz \int_1^{\infty} x^{s-z-1} dx \\
&= \frac{1}{2\pi i} \int_{b-i\infty}^{b+i\infty} \frac{\phi(z)}{z-s} dz - \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} \frac{\phi(z)}{z-s} dz = \phi(s)
\end{aligned}$$

where $-u < a < b < -v$

Application to the trie

$$i_{k,\mathcal{P}}(\nu) = \sum_{|\omega|=k} 1 - (1 + \pi_\omega \nu) e^{-\pi_\omega \nu}$$

$$\mathcal{M}[g(\nu); s] = \int_{\nu=0}^{\infty} g(\nu) \nu^{s-1} d\nu \qquad \Rightarrow \qquad \mathcal{M}[1 - (1+\nu)e^{-\nu}; s] = -(1+s)\Gamma(s)$$

$$|\omega| = k \text{ and } |\omega|_1 = j \qquad \Rightarrow \qquad \pi_\omega = p^j q^{k-j} \qquad (q = 1 - p)$$

$$\mathcal{M}[g(\chi\nu); s] = \chi^{-s} \mathcal{M}[g(\nu); s]$$

$$\mathcal{M}[i_{k,\mathcal{P}}(\nu); s] = -(1+s)\Gamma(s) \sum_{j=0}^{k} \binom{k}{j} p^{-js} q^{-(k-j)s} = -(1+s)\Gamma(s) \left(p^{-s} + q^{-s}\right)^k$$

Application to the trie

$$i_{k,\mathcal{P}}(\nu) = \sum_{|\omega|=k} 1 - (1+\pi_\omega \nu)e^{-\pi_\omega \nu}$$

$$\mathcal{M}[g(\nu); s] = \int_{\nu=0}^{\infty} g(\nu)\nu^{s-1} d\nu \qquad \Rightarrow \qquad \mathcal{M}[1-(1+\nu)e^{-\nu}; s] = -(1+s)\Gamma(s)$$

$$|\omega| = k \text{ and } |\omega|_1 = j \qquad \Rightarrow \qquad \pi_\omega = p^j q^{k-j} \qquad (q = 1-p)$$

$$\mathcal{M}[g(\chi\nu); s] = \chi^{-s}\mathcal{M}[g(\nu); s]$$

$$\mathcal{M}[i_{k,\mathcal{P}}(\nu); s] = -(1+s)\Gamma(s)\sum_{j=0}^{k}\binom{k}{j} p^{-js} q^{-(k-j)s} = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k$$

fundamental strip: $\Re s \in\,]-2, 0[$

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} (1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k \nu^{-s} ds$$

Properties of the integrand

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^{k}\nu^{-s}ds = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty}F(s)ds$$

Properties of the integrand

$k = O(\log(\nu))$ induces parametrization $k = \alpha \log(\nu) = \alpha\, t$

$\nu^{-s} = e^{-s\log(\nu)} = e^{-st}$

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^{k}\nu^{-s}ds = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty}F(s)ds$$

$$F(s) = e^{f(s)} = -(1+s)\Gamma(s)(p^{-s}+q^{-s})^{k}\nu^{-s} = \phi(s)\Theta(s)^{t}$$

Properties of the integrand

$k = O(\log(\nu))$ induces parametrization $k = \alpha \log(\nu) = \alpha\, t$

$\nu^{-s} = e^{-s \log(\nu)} = e^{-st}$

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} (1+s)\Gamma(s) \left(p^{-s} + q^{-s}\right)^{k} \nu^{-s} ds = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} F(s) ds$$

$$F(s) = e^{f(s)} = -(1+s)\Gamma(s)(p^{-s} + q^{-s})^{k} \nu^{-s} = \phi(s)\Theta(s)^{t}$$

$\Im(F(\sigma + ir))$ odd function of r

$\Re(F(\sigma + ir))$ even function of r

Typical case of saddle-point integral   saddle-point: $F'(\sigma) = f'(\sigma) = 0$

Remark - $\mathrm{Res}[F(s), s = 0] = -2^{k}$

Saddle-point method - Overview

$$I = \frac{1}{2\pi i} \int F(z) dz = \frac{1}{2\pi i} e^{f(z)} dz$$

saddle-point σ $\qquad F'(\sigma) = f'(\sigma) = 0$

$$e^{f(z)} = F(\sigma) \times e^{-\frac{(z-\sigma)^2}{2}|f''(\sigma)| + o((z-\sigma)^2)}$$

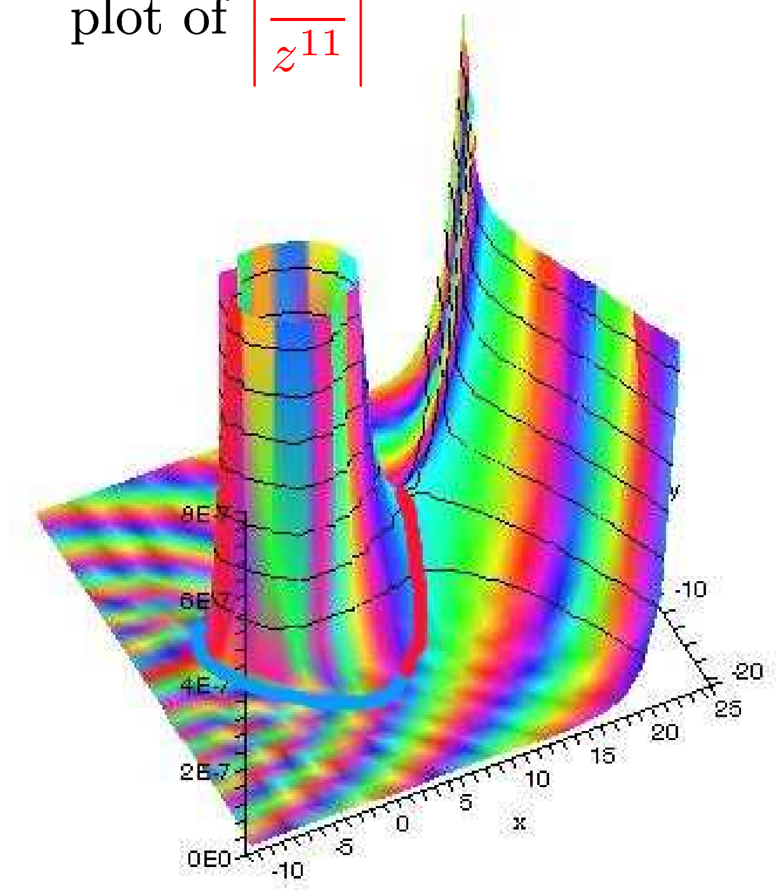
locally gaussian integral

complete with gaussian tails (Laplace method)

$$\Longrightarrow \quad I \sim \frac{F(\sigma)}{\sqrt{2\pi|f''(\sigma)|}}$$

Saddle-point method - Overview

plot of $\left|\dfrac{e^z}{z^{11}}\right|$



$$\frac{1}{10!} = \frac{1}{2\pi i}\int_{|z|=R} e^z \frac{dz}{z^{11}}$$

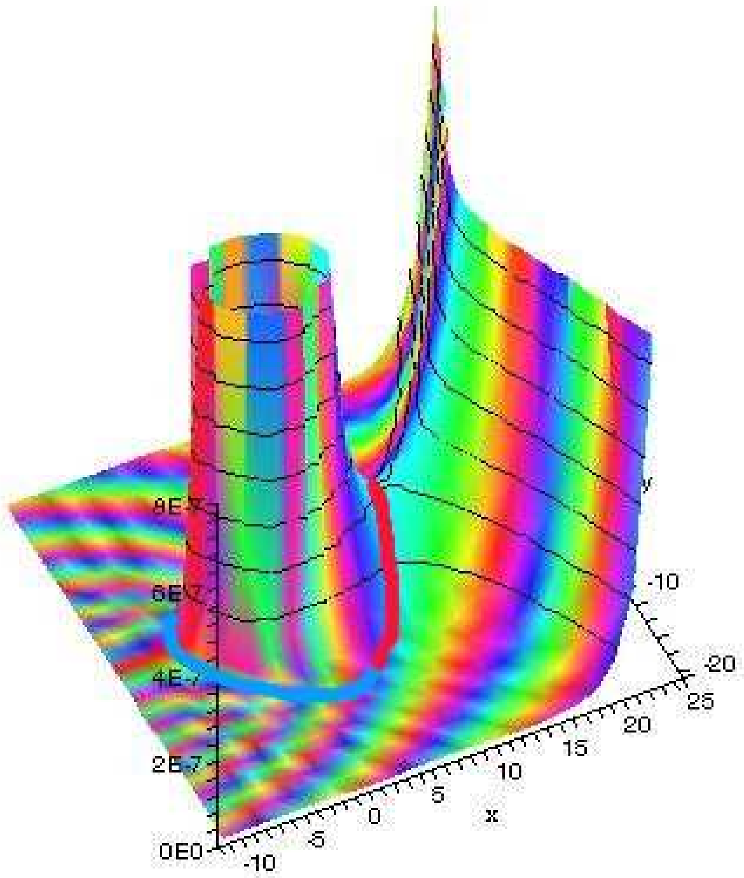
$$I = \frac{1}{2\pi i}\int F(z)dz = \frac{1}{2\pi i} e^{f(z)}dz$$

saddle-point σ $\quad F'(\sigma) = f'(\sigma) = 0$

$$\frac{1}{n!} = \frac{1}{2\pi i}\int_{|z|=R} e^z \frac{dz}{z^{n+1}}$$

$f(z) = z - (n+1)\log(z) \quad \sigma = n+1$

Saddle-point method - Overview



$$\frac{1}{10!} = \frac{1}{2\pi i}\int_{|z|=R} e^z \frac{dz}{z^{11}}$$

$$= \int_{C_1} + \int_{C_2}$$

$$I = \frac{1}{2\pi i}\int F(z)dz = \frac{1}{2\pi i}e^{f(z)}dz$$

saddle-point σ $\quad F'(\sigma) = f'(\sigma) = 0$

$$e^{f(z)} = F(\sigma) \times e^{-\frac{(z-\sigma)^2}{2}|f''(\sigma)|+o((z-\sigma)^2)}$$

locally gaussian integral

complete with gaussian tails (Laplace method)

$$\Longrightarrow \quad I \sim \frac{F(\sigma)}{\sqrt{2\pi|f''(\sigma)|}}$$

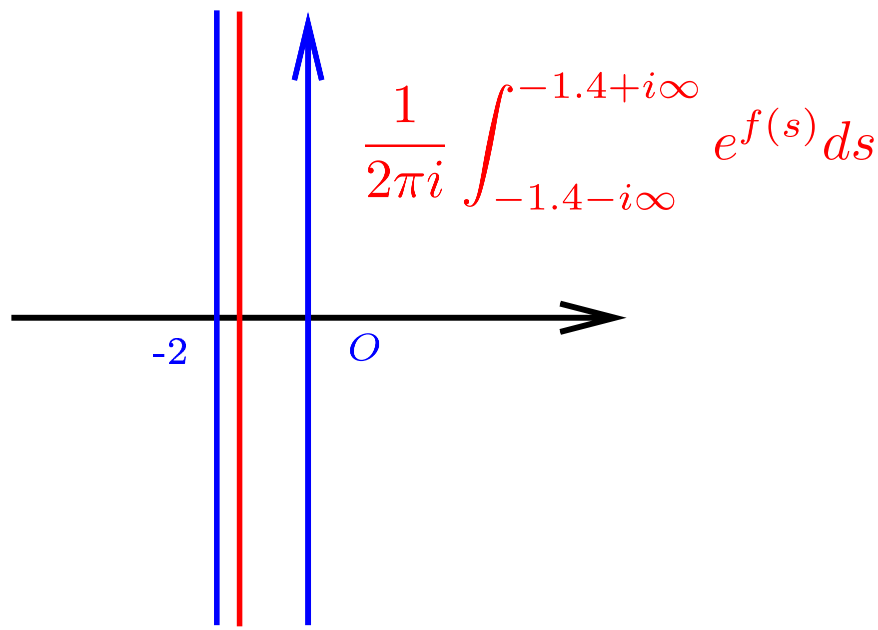
$$\frac{1}{n!} = \frac{1}{2\pi i}\int_{|z|=R} e^z \frac{dz}{z^{n+1}}$$

$$f(z) = z - (n+1)\log(z) \quad \sigma = n+1$$

$$\frac{1}{n!} \sim \frac{e^n}{n^n\sqrt{2\pi n}}$$

Numerical example

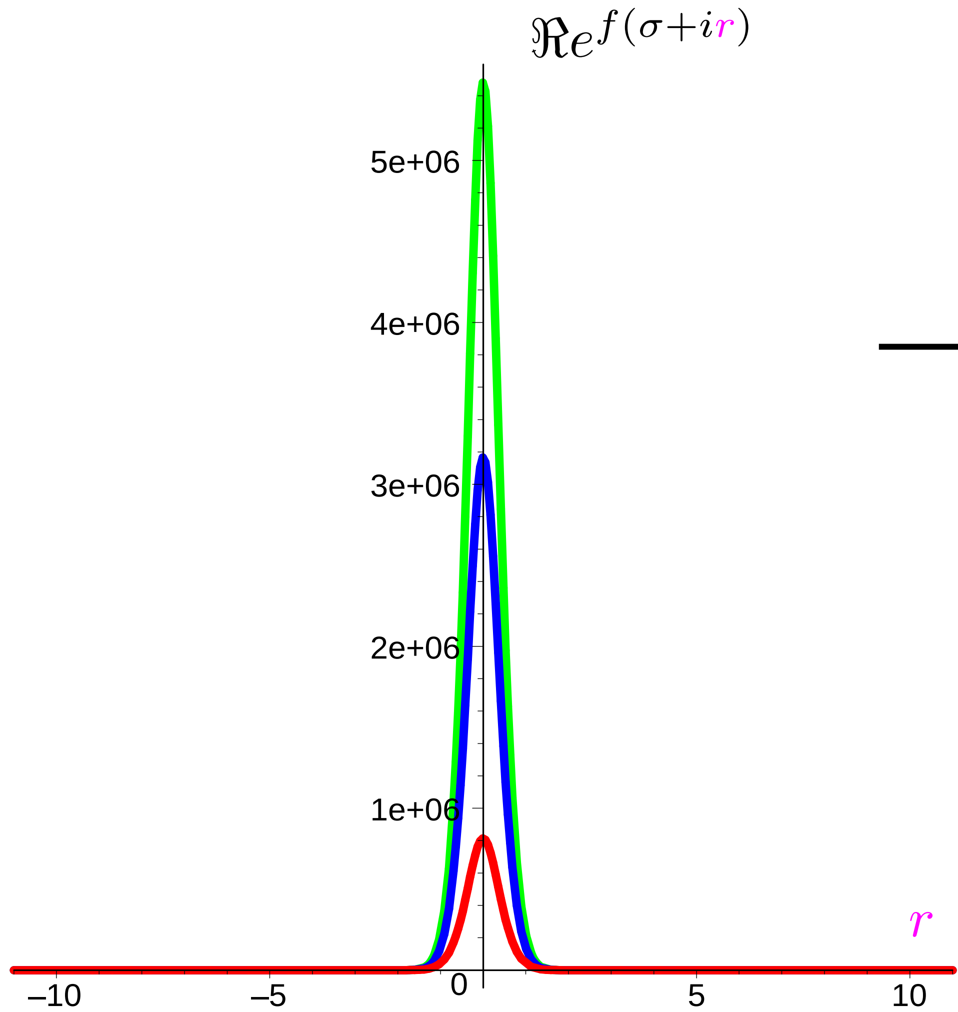
$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^{k}\nu^{-s}ds = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty}F(s)ds$$

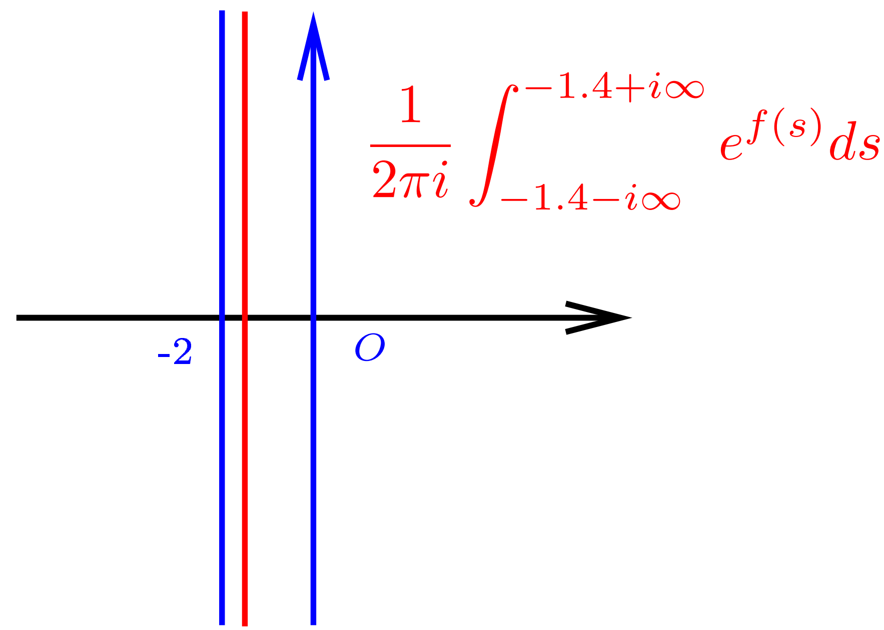


$p = 0.7 \qquad k = \alpha \times \log\nu = 1.8 \times \log\nu \qquad \sigma = -1.4$

Numerical example

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^{k}\nu^{-s}ds = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty}F(s)ds$$





$$\nu = \begin{cases} 1 \\ 4 \\ 7 \end{cases} \times 10^6$$

$$p = 0.7 \qquad k = \alpha \times \log\nu = 1.8 \times \log\nu \qquad \sigma = -1.4$$

Trie - position of the saddle-point

$$i_{k,\mathcal{P}}(\nu) = -\frac{1}{2i\pi}\int_{c-i\infty}^{c+\infty}(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k \nu^{-s}ds = \frac{1}{2i\pi}\int_{c-i\infty}^{c+\infty} e^{f(s)}ds$$

saddle-point σ verifies $f'(\sigma) = 0$

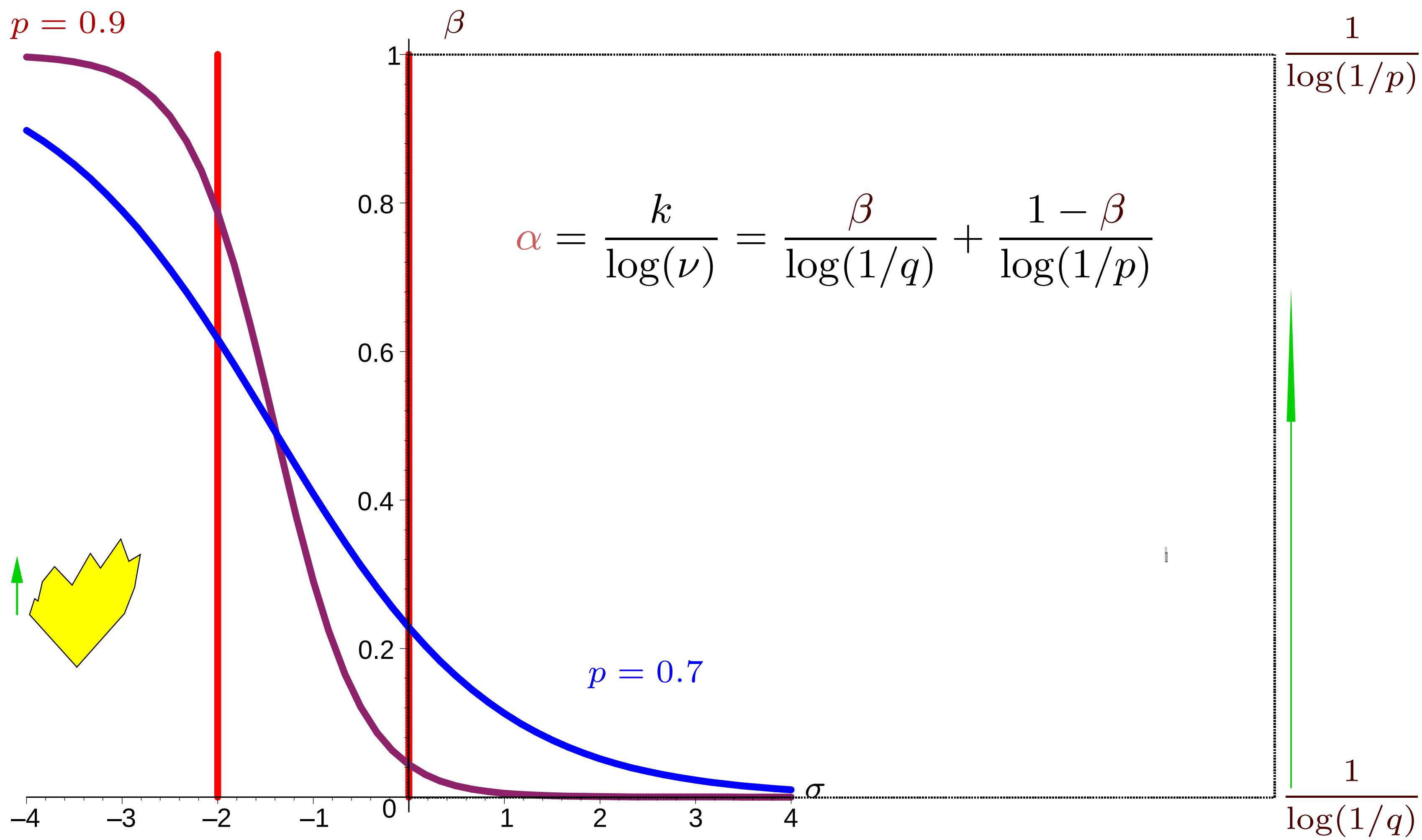
$$f'(s) = \frac{1}{1+s} + \psi(s) - k\frac{p^{-s}\log p + q^{-s}\log q}{p^{-s}+q^{-s}} - \log\nu$$

k and ν tend to infinity

$$k \times \frac{p^{-s}\log 1/p + q^{-s}\log 1/q}{p^{-s}+q^{-s}} = \log\nu$$

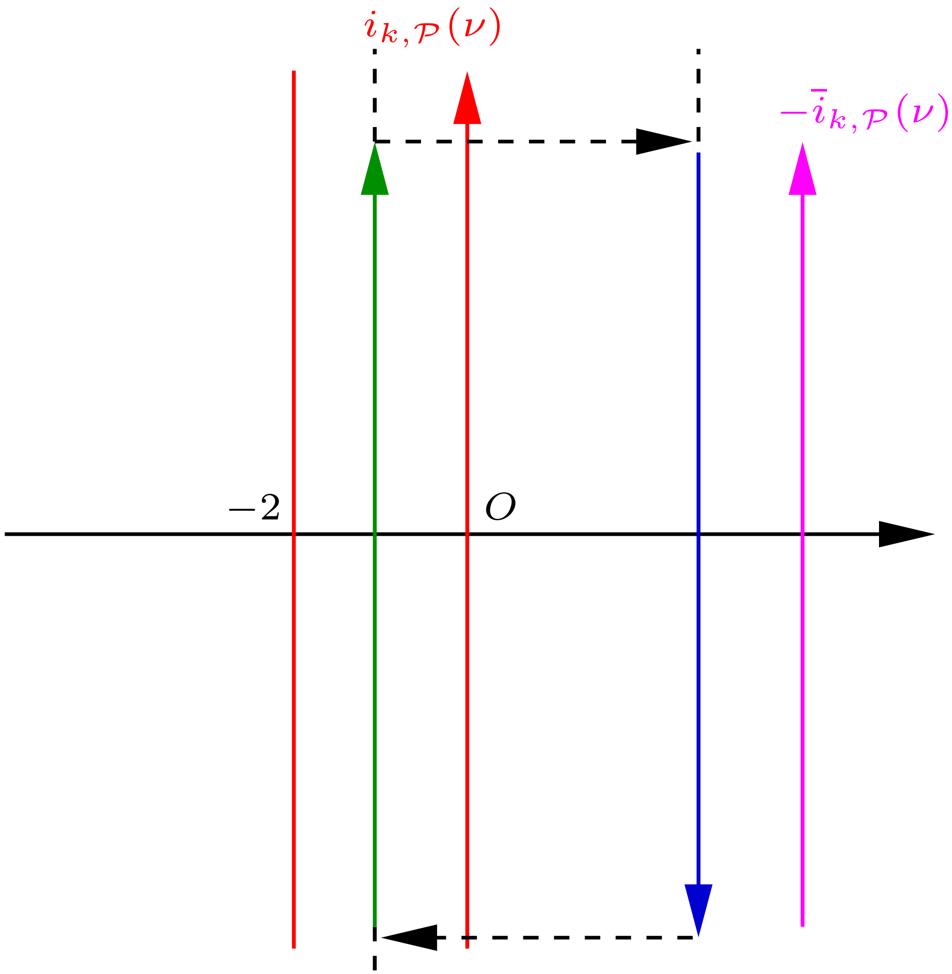
$$\alpha = \frac{k}{\log\nu} \qquad \Longrightarrow \qquad \sigma = \sigma(\alpha) = \frac{\log\left(\dfrac{1-\alpha\log 1/p}{\alpha\log 1/q - 1}\right)}{\log(p/q)} + o(1)$$

Saddle-point position



The saddle-point σ as a function of β, where β is a barycentric weight varying from 0 to 1.

Shifting the integral path

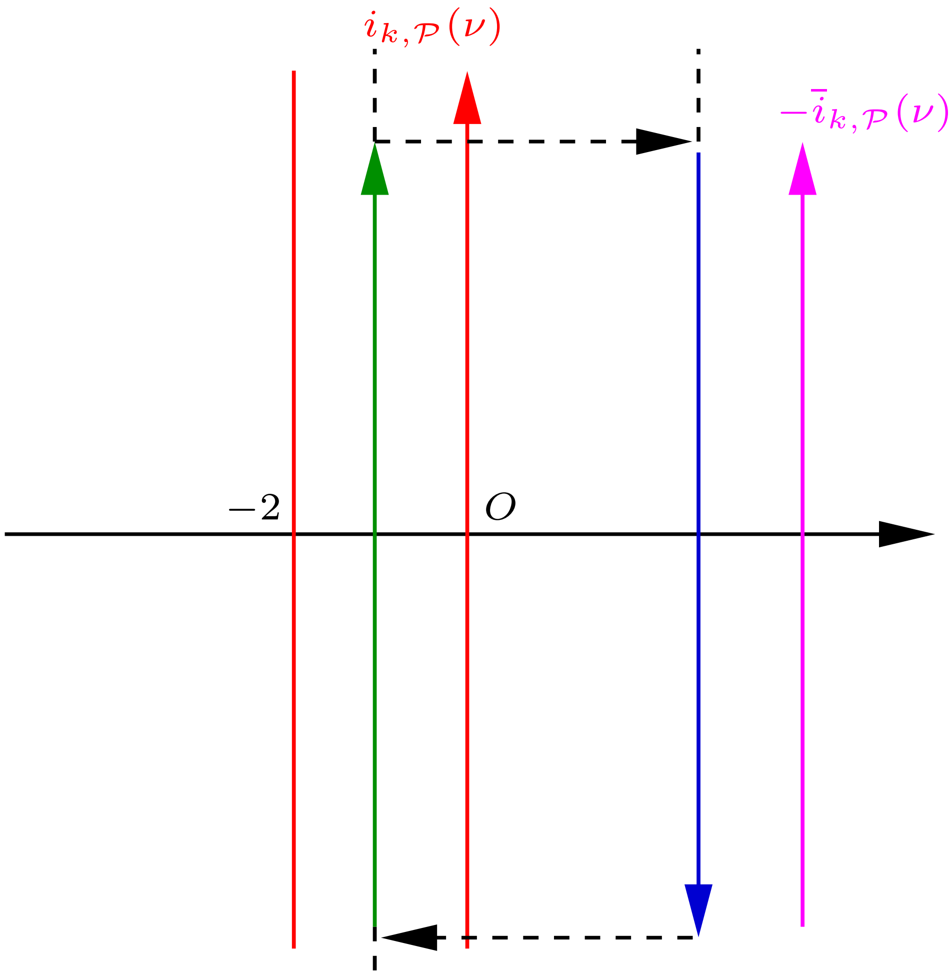


$$F(s) = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k \nu^{-s}$$

$$\mathrm{Res}(F, s=0) = \frac{-2^k}{s}$$

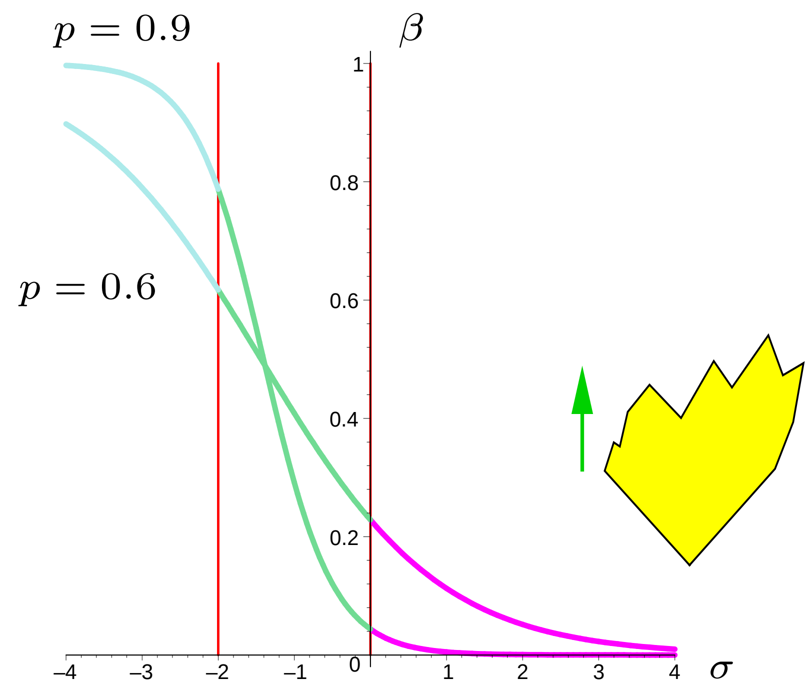
The inverse Mellin integral gives $i_{k,\mathcal{P}}(\nu)$ when $\sigma \in]-2,0[$ (number of present nodes at depth k), and $-\bar{i}_{k,\mathcal{P}}(\nu) = -2^k + i_{k,\mathcal{P}}(\nu)$ when $\sigma \in]0,+\infty]$ (number of missing nodes at depth k).

Shifting the integral path



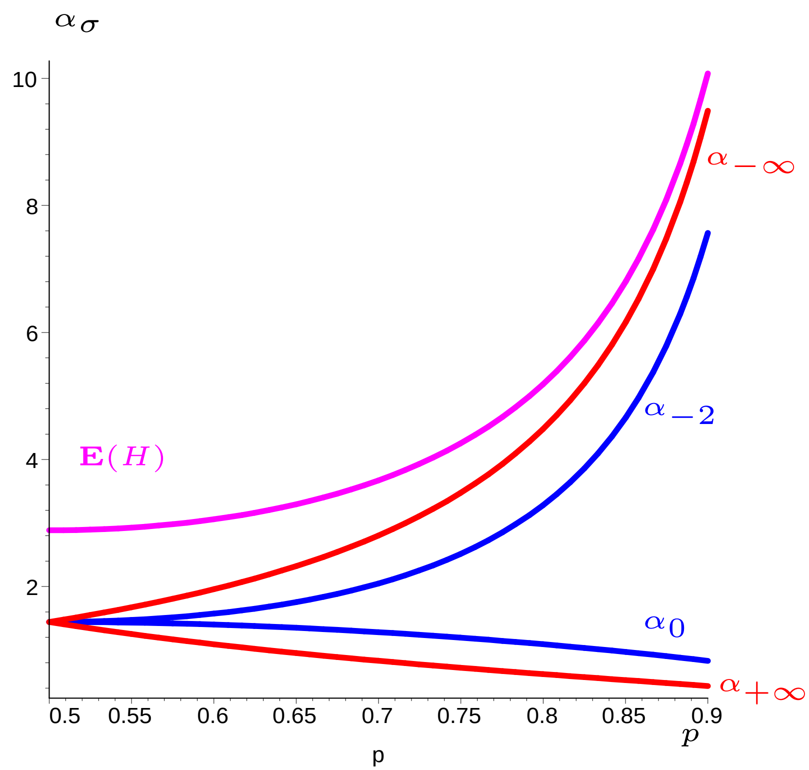
$$F(s) = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k \nu^{-s}$$

$$\mathrm{Res}(F, s=0) = \frac{-2^k}{s}$$



The inverse Mellin integral gives $i_{k,\mathcal{P}}(\nu)$ when $\sigma \in]-2, 0[$ (number of present nodes at depth k), and $-\bar{i}_{k,\mathcal{P}}(\nu) = -2^k + i_{k,\mathcal{P}}(\nu)$ when $\sigma \in]0, +\infty]$ (number of missing nodes at depth k).

Region with real saddle-point for $\alpha = k/\log \nu$



$k = \alpha \times \log \nu$

From bottom to top, the curves are

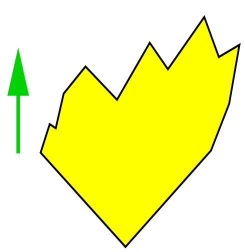
(1) $\alpha_{+\infty}(p,q) = \dfrac{1}{\log 1/q}$

(2) $\alpha_0(p,q) = \dfrac{2}{\log 1/p + \log 1/q}$

(3) $\alpha_{-2}(p,q) = \dfrac{p^2+q^2}{p^2 \log 1/p + q^2 \log 1/q}$

(4) $\alpha_{-\infty}(p,q) = \dfrac{1}{\log 1/p}$

(5) $\mathbf{E}(H) = \dfrac{2}{\log(1/(p^2+q^2))}$



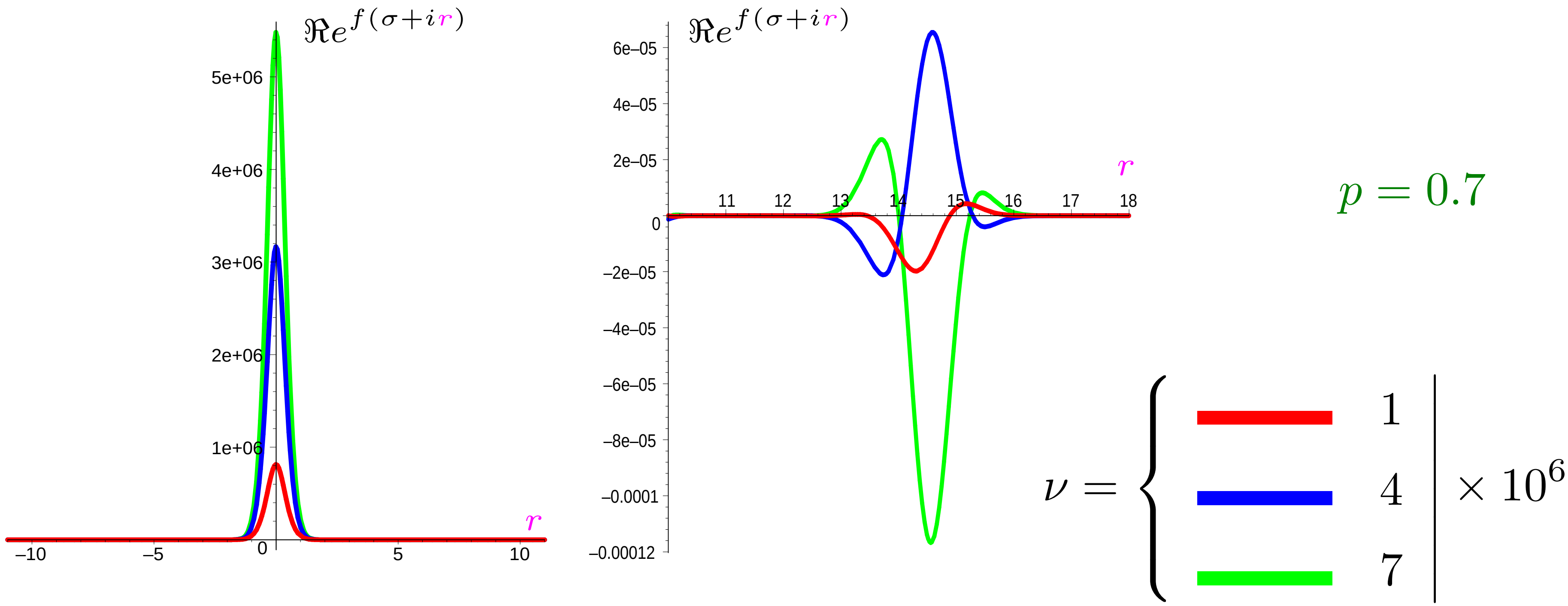
$$\left(\frac{p}{q}\right)^{\sigma} = \frac{\log(1/p)}{\log(1/q)} \times \frac{\dfrac{1}{\log(1/p)} - \alpha}{\alpha - \dfrac{1}{\log(1/q)}}$$

Perturbations

$$I(\nu) = \frac{1}{2i\pi}\int_{\sigma-i\infty}^{\sigma+i\infty} -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^{k}\nu^{-s}ds = \frac{1}{2i\pi}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{f(s)}ds$$

$s = \sigma + ir$ $\qquad$ $|\Gamma(\sigma+ir)| = O(e^{-|r|})$ as $|r| \to \infty$

$|p^{-\sigma-ir}+q^{-\sigma-ir}|$ periodic, maximum when $p^{-\sigma-ir}$ and $q^{-\sigma-ir}$ in phase



$$k = \alpha \times \log\nu = 1.8 \times \log\nu \qquad \sigma = -1.4$$

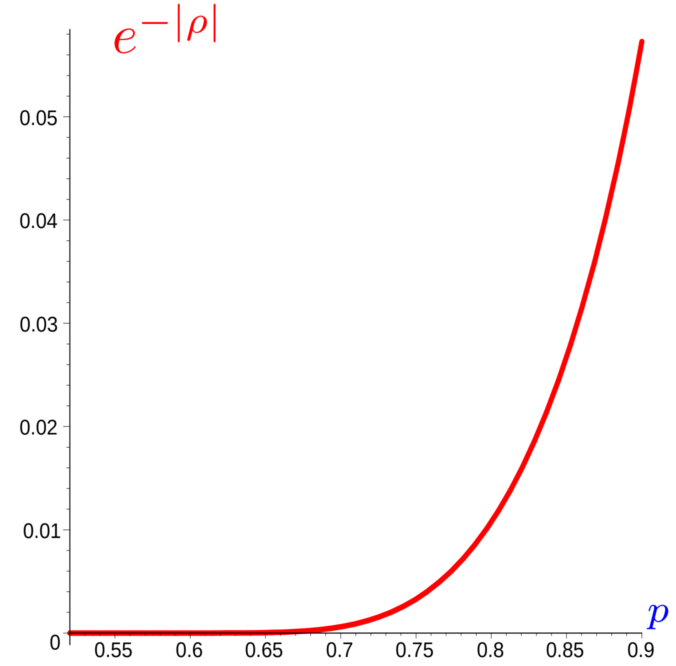
Bounding the periodic perturbation terms

$|p^{-\sigma-ir}+q^{-\sigma-ir}|$ periodic, maximum when $p^{-\sigma-ir}$ and $q^{-\sigma-ir}$ in phase

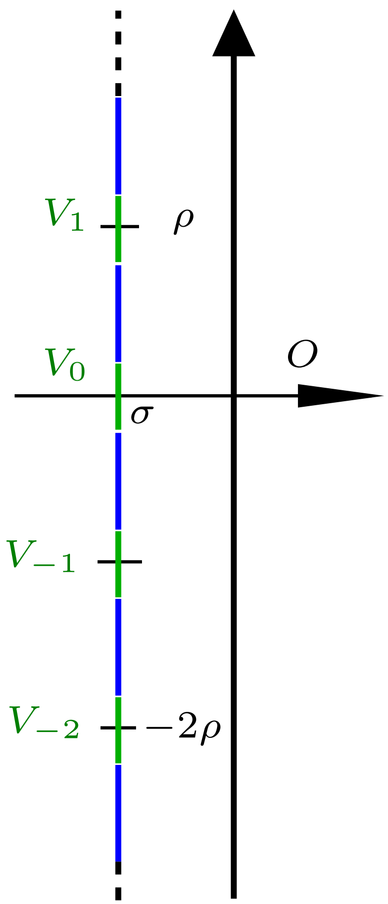
$$\Longrightarrow \quad \begin{cases} \exists\theta, 0<\theta<2\pi, \quad j_p, j_q \in \mathbb{N}, \quad j_p < j_q \\ |r|\log 1/p = \theta + 2j_p\pi \quad \text{and} \quad |r|\log 1/q = \theta + 2j_q\pi \end{cases}$$

$$\Longrightarrow \ |r| = j\rho = j \times 2\pi \times \frac{1}{\log(p/q)}$$

$$|\Gamma(\sigma+i\rho)| \sim |\Gamma(\sigma)|e^{-|\rho|}$$



Trie - end result



$$B_j = \frac{1}{2i\pi}\int_{r=j\rho-\delta}^{j\rho+\delta} e^{f(\sigma+ir)}dr \qquad |B_j| = B_0 \times c_j(\nu) e^{-|j|\rho}$$

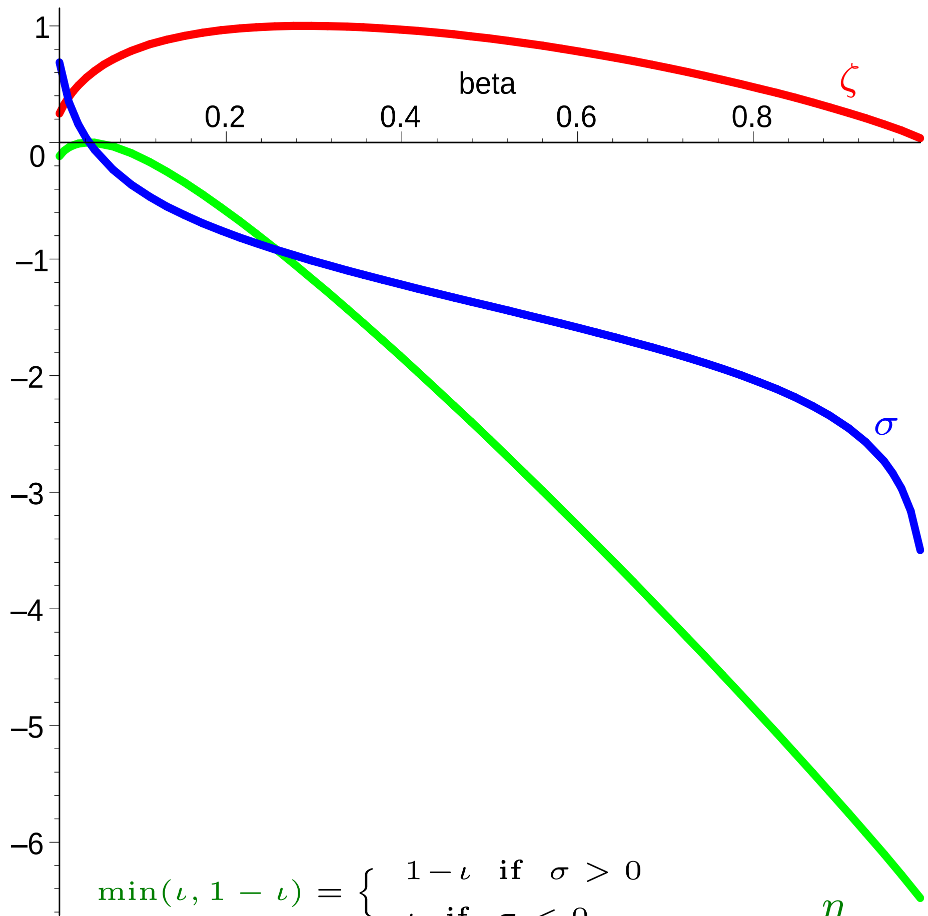
$$B_0 \sim \frac{e^{f(\sigma)}}{\sqrt{2\pi f''(\sigma)}} \quad \text{(saddle-point evaluation)}$$

$$i_{k,\mathcal{P}}(\nu) = \frac{-(1+\sigma)\Gamma(\sigma)\nu^{\alpha \log(p^{-\sigma}+q^{-\sigma}) - \sigma}}{\sqrt{2\pi\alpha \log(\nu) \times U(\sigma,p,q)}}$$

$$\times \left(1 + c(\nu)e^{-\rho}\right) \times \left(1 + O\left(\frac{1}{\sqrt{\log(\nu)}}\right)\right)$$

$$|c_j(\nu)| = O(1) \qquad |c(\nu)| = O(1)$$

Dominant power ζ in $i_{k,\mathcal{P}}(\nu)$



$$p = 0.9$$

$$\alpha = \frac{k}{\log(\nu)} = \frac{\beta}{\log(1/q)} + \frac{1-\beta}{\log(1/p)}$$

$$i_{k,\mathcal{P}}(\nu) \sim \frac{\nu^{\zeta}}{\sqrt{2\pi \log(\nu) U(\sigma)}}$$

$$\zeta = \alpha \log(p^{-\sigma} + q^{-\sigma}) - \sigma$$

$$\iota = \frac{i_{k,\mathcal{P}}(\nu)}{2^k} \approx \nu^{\eta}$$

Depoissonization "à la Ramanujan"

Ramanujan simplified entry

$$\begin{cases} h(x) \text{ of at most polynomial growth,} & \left|\dfrac{h^{(m)}(x)}{m!}\right| \le \left(\dfrac{1}{x}\right)^m \quad (x \text{ large}) \\[2ex] h_\infty(x) = e^{-x} \displaystyle\sum_{k=2}^{\infty} \frac{x^k h(k)}{k!} & \end{cases}$$

$$\Longrightarrow \quad h_\infty(x) = h(x) + x h''(x) + O\left(x^{-2}\right) \qquad (x \to \infty)$$

Reasonning by contradiction implies

$$i_k(n) = i_{k,\mathcal{P}}(n) \left(1 + O\left(n^{-(1-\epsilon)}\right)\right)$$

Part III

Suffix-tree - Asymptotic analysis

Profile asymptotics comparisons - Method

Trie $\mathcal{P}_n$ keys (Poisson model)

$\updownarrow$

Trie n keys

Profile asymptotics comparisons - Method

Trie $\mathcal{P}_n$ keys (Poisson model)

Trie n keys

Suffix-Tree n keys

Profile asymptotics comparisons - Method

Trie $\mathcal{P}_n$ keys (Poisson model) $\sum_{|\omega|=k}$

Suffix-Tree n keys $\sum_{|\omega|=k}$

Repeated words in random strings

W_n random string of size n $\qquad \Pr(1) = p = 1 - q = 1 - \Pr(0)$

$o_\omega^{(n)}$ number of occurrences of word ω in W_n

$$\Pr(o_\omega^{(n)} = 0) + \Pr(o_\omega^{(n)} = 1) + \Pr(o_\omega^{(n)} \geq 2) = 1$$

$Y_\omega^{(n)} = \mathbf{1}_{\{o_\omega^{(n)} \geq 2\}}$ indicator that a word ω is repeated W_n

$$\mathbf{E}(Y_\omega^{(n)}) = 1 - \Pr(o_\omega^{(n)} = 0) - \Pr(o_\omega^{(n)} = 1)$$

$Y^{(n)}$ counts the number of repeated words in W_n

$$\mathbf{E}(p_k^{(S)}(n)) = \mathbf{E}(Y^{(n)}) = 2^k - \sum_{|\omega|=k} \Pr(o_\omega^{(n)} = 0) - \sum_{|\omega|=k} \Pr(o_\omega^{(n)} = 1)$$

$$Y(z) = \sum_{n \geq 0} Y^{(n)} z^n = \frac{2^k}{1-z} - \sum_{|\omega|=k} O_\omega^{(0)}(z) - \sum_{|\omega|=k} O_\omega^{(1)}(z)$$

Languages and autocorrelation

$\mathcal{L} \subseteq \{0,1\}^\star$ $\qquad$ $\mathcal{L}(z) = \sum_{\omega \in \mathcal{L}} \pi_\omega z^{|\omega|} = \sum_{n \geq 0} l_n z^n$

$\pi_\omega = \Pr(\omega)$ $\qquad$ $l_n = \Pr(\omega \in \mathcal{L}) \quad \text{if } |\omega| = n$

autocorrelation set of word ω

$$\mathcal{A}_\omega = \Big\{ h; \quad \omega.h = u.\omega \qquad \text{and} \qquad |h| < |\omega| \Big\}$$

$$\mathcal{A}_{ababa} = \{\epsilon, ba, baba\}$$

$ababa$	ϵ
$ababa$	ba
$ababa$	$baba$

Languages decomposition

First $\mathcal{F} = \{\ w = u.\omega \quad \text{et} \quad \not\exists r, s,\ w = r.\omega.s\ \}$

$aaaaaababa \subset \mathcal{F}, \quad bbbbbabababa \not\subset \mathcal{F}$

Ultimate $\mathcal{U} = \{w, \quad \not\exists r, s,\ \omega.w = r.\omega.s\}$ $\qquad \mathcal{O}^{(1)} = \mathcal{F}\mathcal{U}$

$ababa \qquad\qquad ababa$

$aabbbabbbbbb \subset \mathcal{U} \qquad babbbbbbbbb \not\subset \mathcal{U}$

No occurrences $\mathcal{O}^{(0)} = \Sigma^\star - \Sigma^\star.\omega.\Sigma^\star = \{w,\ \not\exists r, s,\ w = r.\omega.s\}$

$$\begin{cases} \mathcal{O}^{(0)}\, x = \mathcal{O}^{(0)} + \mathcal{F} - \epsilon \\ \mathcal{O}^{(0)}\, \omega = \mathcal{F}\mathcal{A}_\omega \end{cases} \Longrightarrow \begin{cases} F(z) = \dfrac{\pi_\omega z^{|\omega|}}{K_\omega(z)} \\ U(z) = \dfrac{1}{K_\omega(z)} \\ \mathcal{O}^{(0)}(z) = \dfrac{\mathcal{A}_\omega(z)}{K_\omega(z)} \end{cases} \qquad \frac{1}{K_\omega(z)} = \frac{1}{\pi_\omega z^{|\omega|} + (1-z)\mathcal{A}_\omega(z)}$$

Expectation of number of repeated words

$$O_\omega^{(1)} = \mathcal{F}_\omega \mathcal{U}_\omega \Longrightarrow O_\omega^{(1)}(z) = F(z)U(z)$$

Generating function for the repeated words

$$P_k^{(S)}(z) = \sum_{n\geq 0} p_k^{(S)}(n) z^n = \frac{2^k}{1-z} - \sum_{|\omega|=k} \left(\frac{\mathcal{A}_\omega(z)}{K_\omega(z)} + \frac{\pi_\omega z^{|\omega|}}{K_\omega(z)^2} \right)$$

$$\frac{1}{K_\omega(z)} = \frac{1}{\pi_\omega z^{|\omega|} + (1-z)\mathcal{A}_\omega(z)}$$

Suffix-tree - Asymptotic expansion for $i_k^{(S)}(n)$

$$i_k^{(S)}(n) = [z^n]P_k^{(S)}(z) \approx 2^k - [z^n]\frac{\pi_\omega z^{|\omega|}}{\left(\pi_\omega z^{|\omega|} + (1-z)\mathcal{A}_\omega(z)\right)^2}$$

dominant singularity $\rho_\omega = 1 + o(1) \;\Rightarrow\;$ bootstrapping, Cauchy integration

$$i_k^{(S)}(n) = \sum_{|\omega|=k} 1 - \left(1 + \frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) e^{-\frac{n\pi_\omega}{\mathcal{A}_\omega(1)}}$$

$$+\, i_{k,\mathcal{P}}^{(T)}(n)\left(O(n^{-\upsilon}) + O\left(n^{-\alpha \log(1/p)}\right)\right)$$

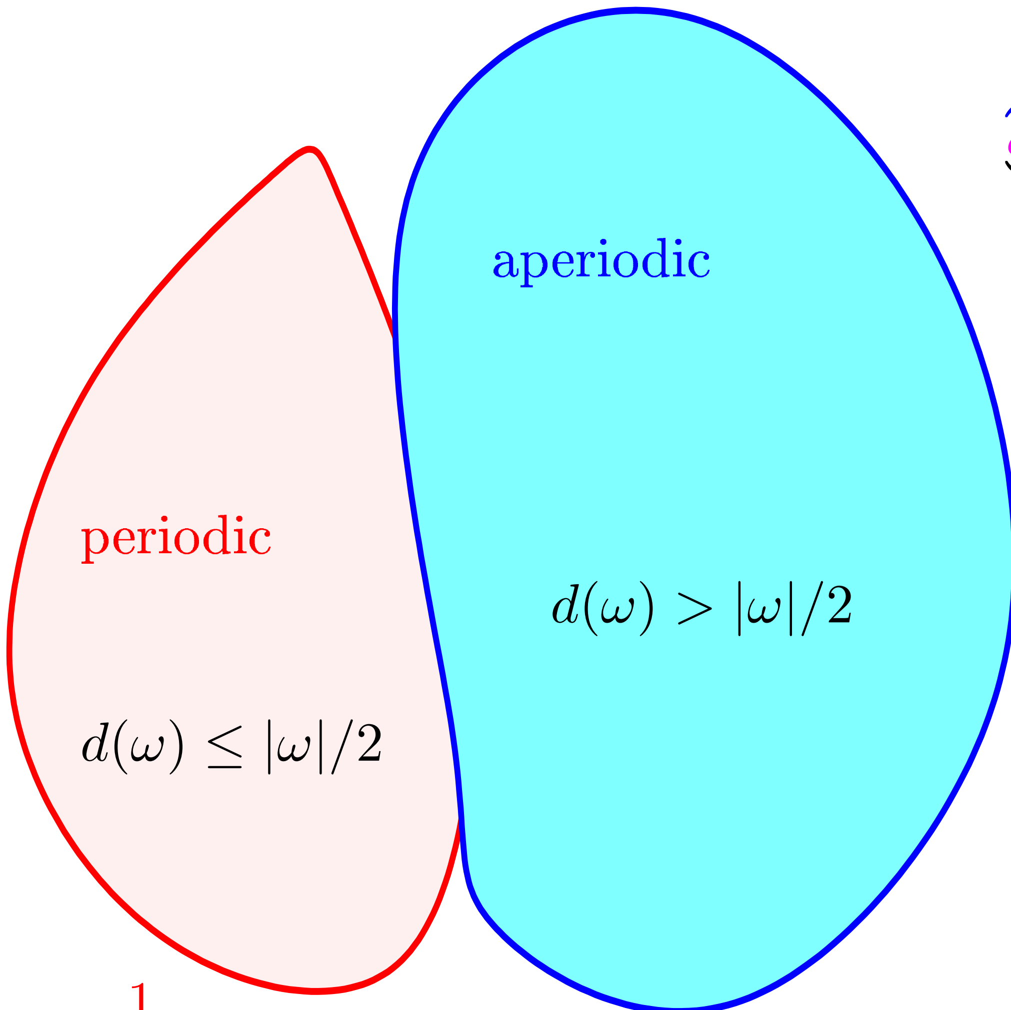
$$\alpha > \alpha_\sigma^{(S)} = (1+\sigma)/\log\left(\frac{p^{-\sigma}+q^{-\sigma}}{p^2+q^2}\right) \quad \Longrightarrow \quad \upsilon > 0$$

Periodic and aperiodic words

basic period $d = d(\omega)$

$\underbrace{\overbrace{aab}^{d=3}aabaa}_{|\omega|=8}$

$\underbrace{\overbrace{aaaaaaab}^{d=8}a}_{\omega=9}$



$1 \leq \mathcal{A}_\omega(1) \leq \dfrac{1}{1-p}$

$1 \leq \mathcal{A}_\omega(1) \leq 1 + \dfrac{p^{k/2}}{1-p}$

Trie Poisson versus suffix-tree

suffix-tree $\quad i_k^{(S)}(n) \approx a_k^{(S)}(n) = \sum_{|\omega|=k} 1 - \left(1 + \frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) e^{-\frac{n\pi_\omega}{\mathcal{A}_\omega(1)}}$

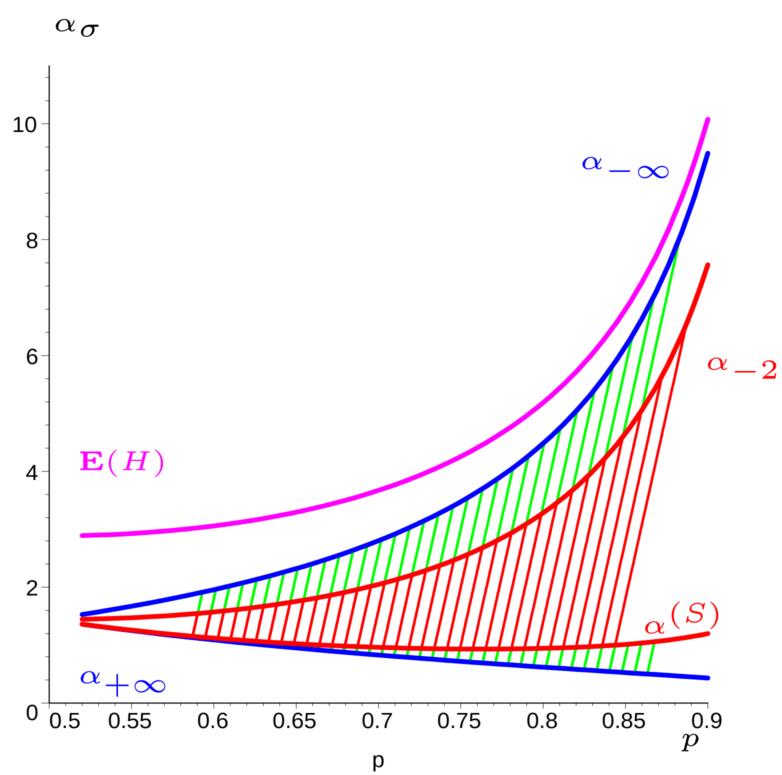
trie $\quad i_{k,\mathcal{P}}^{(T)}(n) = \sum_{|\omega|=k} 1 - (1 + n\pi_\omega)e^{-n\pi_\omega}$

use Mellin transform of $g(\chi x)$ with $g(x) = 1 - (1+x)e^{-x}$

and bounds for χ of $\dfrac{1}{\mathcal{A}_\omega(1)}$ on periodic and aperiodic words

$$\left| i_k^{(S)}(n) - i_k^{(T)}(n) \right| = i_k^{(T)}(n) \times O(n^{-\lambda}) \qquad \lambda > 0$$

Summarizing



$k = \alpha \times \log \nu$

From bottom to top, the curves are

(1) $\alpha_{+\infty}(p,q) = \dfrac{1}{\log 1/q}$

(2) $\alpha^{(S)}$ verifies

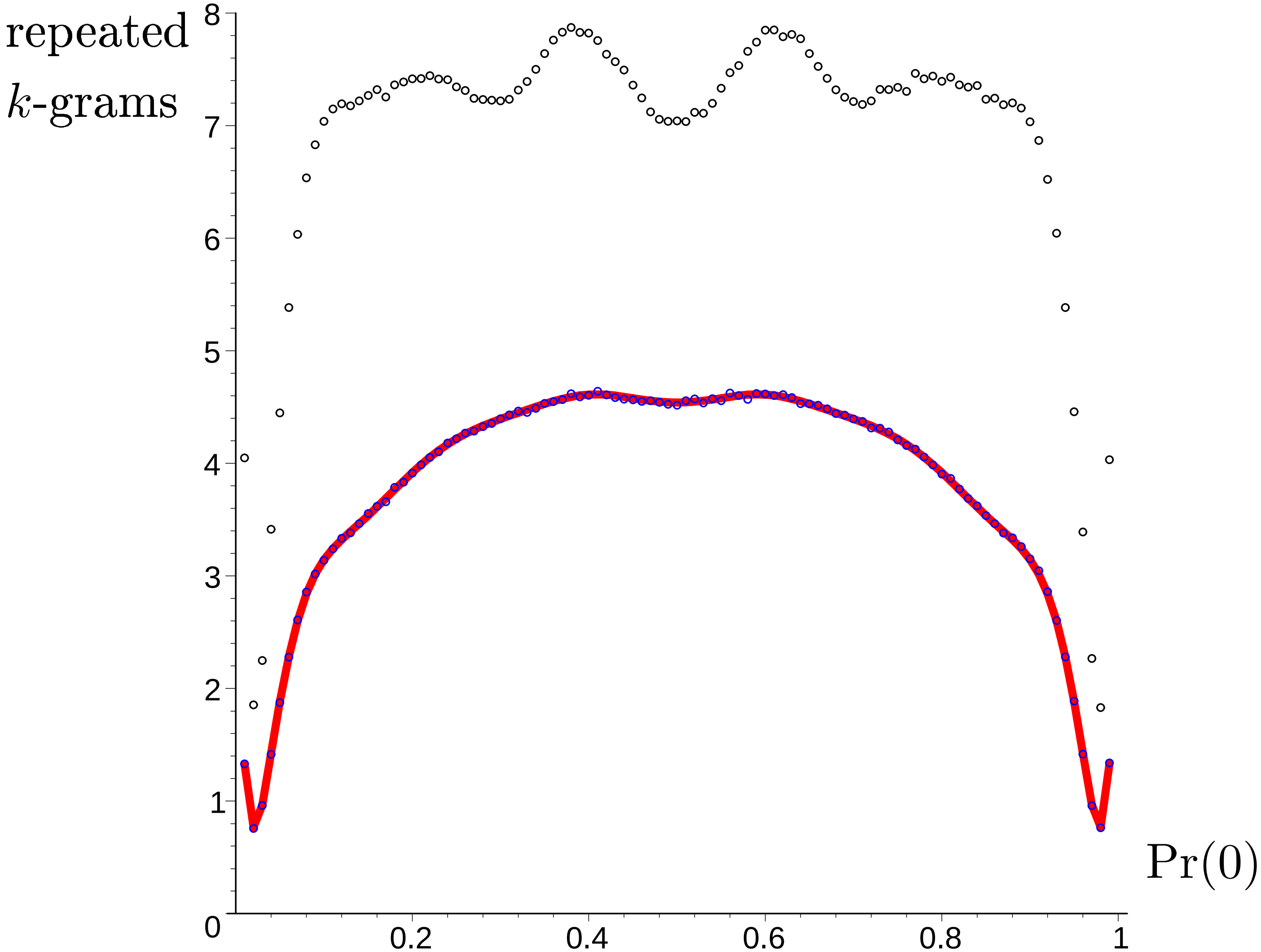
$$\alpha - (1+\sigma(\alpha))/\log\left(\frac{p^{-\sigma(\alpha)} + q^{-\sigma(\alpha)}}{p^2+q^2}\right) > 0$$

(3) $\alpha_{-2}(p,q) = \dfrac{p^2+q^2}{p^2 \log 1/p + q^2 \log 1/q}$

(4) $\alpha_{-\infty}(p,q) = \dfrac{1}{\log 1/p}$

(5) $\mathbf{E}(H) = \dfrac{2}{\log(1/(p^2+q^2))}$

Bad news - Standard deviation



$n = 300$ $\Sigma = \{0, 1\}$ $k = 10$

theoretical - trie (solid line)
simulations for trie (blue circles)
simulations for suffix-tree (black circles)

References

– *J. Fayolle*, 2004, trie and suffix-tree

– *Flajolet*, 2005^+, Mellin transform, saddle-point method

– *Jacquet, Szpankowski*, 1994, trie and suffix-tree

– *P.N.*, 2005, AofA05

– *Nielsen*, 1905, Gamma function, Mellin transform

– *Park, Szpankowski*, 2005, profile of tries, SODA05

– *Rahmann, Rivals*, 2003, missing words in texts