Average-Case Analysis of Approximate Trie Search

Moritz G. Maaß

maass@in.tum.de
Institut für Informatik
Technische Universität München

15th CPM, July 2004

Basic Definitions

- Let Σ be an alphabet of fixed size σ. Σ^* the set of all (finite) strings over Σ.
- For the string $u = u_1 \cdots u_m$ we call $|u| := m$ its length, $u_1 \cdots u_i$ a prefix, $u_j \cdots u_m$ a suffix, and $u_k \cdots u_l$ a substring.
- $d : \sigma \times \sigma \to \{0, 1\}$ be an error function and let $\hat{d} : \sigma^* \times \sigma^* \to \mathbb{N}$ be its extension to strings, such that

$$\hat{d}(u, v) = \begin{cases} \infty & \text{, if } |u| \neq |v|, \\ \sum_{i=1}^{|u|} d(u_i, v_i) & \text{, otherwise.} \end{cases}$$

Examples of Error Functions

- Hamming Distance.

$$d(x,y) = \begin{cases} 1 & \text{, if } x \neq y, \\ 0 & \text{, otherwise.} \end{cases}$$

- Number of don't-cares. Let $\mathsf{c} \in \Sigma$ be a special don't-care symbol.

$$d(x,y) = \begin{cases} 1 & \text{, if } x = \mathsf{c} \text{ or } y = \mathsf{c}, \\ 0 & \text{, otherwise.} \end{cases}$$

- Hamming Distance with don't-cares.

$$d(x,y) = \begin{cases} 1 & \text{, if } x \neq y \text{ and neither } x = \mathsf{c} \text{ nor } y = \mathsf{c}, \\ 0 & \text{, otherwise.} \end{cases}$$

Examples of Error Functions (2)

- Arithmetic Distance. Let $\Sigma = [0, \ldots, \sigma - 1]$ be ordered and let $i < (\sigma - 1)/2$ be a constant. Define $a(x, y) = \max(x, y) - \min(x, y)$.

$$d(x,y) = \begin{cases} 1 & \text{, if } \min\left(a(x,y), \sigma - a(x,y)\right) > i, \\ 0 & \text{, otherwise.} \end{cases}$$

For example, let Σ be the discretization of all angles, i.e. $\Sigma = \{[0, \frac{1}{12}\pi), \ldots, [\frac{23}{12}\pi, 2\pi)\}$, then $d(x, y)$ measures whether two angles are not too far apart.

Problem Definition

For a given alphabet, a given error function d, and a given threshold D, we define the following two-phase-problem

Input:

1. A database of strings $S = \{X^{(1)}, \ldots, X^{(n)}\} \subset \Sigma^*$ (initial phase).
2. One (or more) query strings $P \in \Sigma^*$ of length m (query phase).

Output:

1. Some data structure DS for the string database.
2. Using DS, search for each P. This may answer one of the one of the query types: *Occurrence*, *Longest Prefix*, *Count*, *All Prefixes*.

Query Types

1. **Occurrence**: Answer *YES*, if there exists a prefix $X^{(j)}[1,m]$ of $X^{(j)} \in S$ with $\hat{d}(X^{(j)}[1,m],P) \leq D$, and *NO* otherwise.
2. **Longest Prefix**: Answer j,l, if the prefix $X^{(j)}[1,l]$ of $X^{(j)} \in S$ satisfies with $\hat{d}(X^{(j)}[1,l],P[1,l]) \leq D$ and there is no i with a longer matching prefix $\hat{d}(X^{(i)}[1,l+1],P[1,l+1]) \leq D$.
3. **Count**: Answer $k = \left|\left\{j \mid \hat{d}(X^{(j)}[1,m],P) \leq D\right\}\right|$.
4. **All Prefixes**: Answer with the set of all (maximal) matches

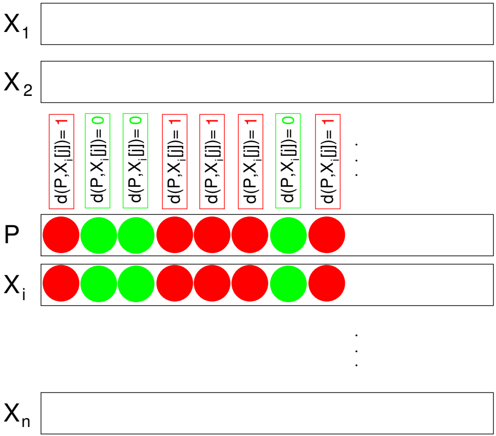
$$\left\{(j,l_j) \mid \hat{d}(X^{(j)}[1,l_j],P[1,l_j]) \leq D \text{ and } d(X^{(j)}_{l_j+1},P_{l_j+1}) > 0\right\}.$$

Overview

- Average-Case behavior of "Linear Search" (LS), which simply compares each query pattern with every database strings.
- Average-Case behavior of "Trie Search" (TS), which builds a trie from all database strings and uses the trie to speed up the pattern search.
- Asymptotically, the worst case for both algorithms for both algorithms is the same.
- There is a threshold in the number of errors, where TS has the same asymptotic running time.

LS Algorithm

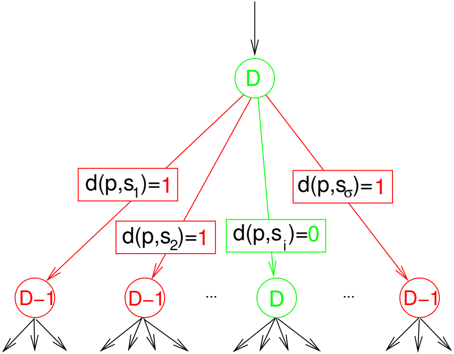
Input: Strings $X_1, \ldots, X_n$ and pattern P, bound D.
for i **from** 1 **to** n **do**
  $j := 1$
  $c := 0$
  $l := \min\{\text{length}(P), \text{length}(X_i)\}$
  while $c \leq D$ **do**
    while $j \leq l$ **and** $d(P[j], X_i[j]) = 0$ **do**
      $j := j + 1$
    $c := c + 1$
    $j := j + 1$
  if $j - 2 = l$ **then**
    report match for X_i



TS Algorithm : rfind(v, P, pos, D)

if $D \geq 0$ **then**
  if v is a leaf **then**
    report match for $X_{\text{value}(v)}$
  else if $pos >$ length(P) **then**
    for all leaves u in the subtree of v **do**
      report match for $X_{\text{value}(u)}$
  else
    for each child u of v **do**
      let c be the edge label of (u, v)
      if $d(P[pos], c) = 0$ **then**
        rfind$(u, P, pos + 1, D)$
      else
        rfind$(u, P, pos + 1, D - 1)$

Started with rfind$(root, P, 0, D)$



Probabilistic Model

- All strings are generated at random by a memoryless source.
- For the a random string $u = u_1 \cdots u_n$ and alphabet $\Sigma = \{s_1, \ldots, s_\sigma\}$

$$\Pr\{u_j = s_i\} = \frac{1}{\sigma}.$$

- We assume that all strings are infinite.
- Random variables for the number of character comparisons
  - L_n^D for the LS algorithm and
  - T_n^D for the TS algorithm.

Error Probability

The error probability of two random characters under the error function d is given by

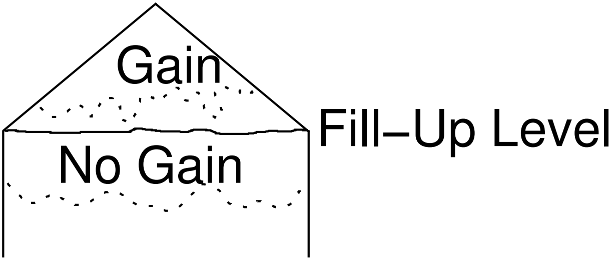
$$p = \frac{\sum_{i=1}^{\sigma} \sum_{j=1}^{\sigma} d(s_i, s_j)}{\sigma^2}.$$

We define $q := 1 - p$.

- Hamming distance: $p = 1 - \frac{1}{\sigma}$, $q = \frac{1}{\sigma}$
- Number of don't-cares: $p = \frac{2\sigma-1}{\sigma^2}$, $q = 1 - \frac{2\sigma-1}{\sigma^2}$
- Hamming distance with don't cares: $p = 1 - \frac{3\sigma-2}{\sigma^2}$, $q = \frac{3\sigma-2}{\sigma^2}$
- Arithmetic distance: $p = 1 - \frac{2i+1}{\sigma}$, $q = \frac{2i+1}{\sigma}$

Expected Results

- The expected depth of a trie is $\log_\sigma n$ (Pittel 1986, Szpankowski 1988).
- There are at most n nodes with depth below $\log_\sigma n$ in the trie.
- We expect a speed up of TS over LS when the expected number of comparisons for each string is smaller than $\log_\sigma n$.



Related Work (incomplete)

- k-d-tries: Flajolet and Puech 1986
- Average behavior of tries and suffix trees: Apostolico and Szpankowski 1992
- Regular Expressions: Baeza-Yates and Gonnet 1996
- All-Against-All matching: Baeza-Yates and Gonnet 1999
- Hybrid Indexing Method: Navarro and Baeza-Yates 2000
- Tree/Trie Traversal algorithms: Jokinen and Ukkonen 1991, Ukkonen 1993, Cobbs 1995, Schulz and Mihov 2002

Average Complexity of the LS algorithm

$$\mathbf{E}\left[L_n^D\right] = \frac{(D+1)n}{p}$$

We can even prove convergence:

$$\lim_{n\to\infty} \frac{L_n^D}{n(D+1)} = \frac{1}{p} \qquad \text{(pr.)}$$

$$\lim_{n\to\infty} \frac{L_n^D}{n(D+1)} = \frac{1}{p} \qquad \text{(a.s.)}$$

Average Complexity of the TS algorithm

$$\mathbf{E}\left[T_n^D\right] = \begin{cases} O\left((\log n)^{D+1}\right), & \text{for } D = O(1) \text{ and } q = \sigma^{-1} \\ O\left((\log_\sigma n)^D n^{\log_\sigma q + 1}\right), & \text{for } D = O(1) \text{ and } q > \sigma^{-1} \\ O(1), & \text{for } D = O(1) \text{ and } q < \sigma^{-1} \\ o(n), & \text{for } D + 1 < p \log_\sigma n \\ \Omega(n \log_\sigma n), & \text{for } D + 1 > p \log_\sigma n. \end{cases}$$

Exact Average Complexity of the TS algorithm

For Hamming Distance we have

$$\mathbf{E}\left[T_n^D\right] = \frac{\sigma(\sigma-1)^D}{(D+1)!}\left(\log_\sigma n\right)^{D+1} + O\left(\left(\log n\right)^D\right),$$

otherwise we have

$$\mathbf{E}\left[T_n^D\right] = \frac{(1-q)^D}{D!q^{D+1}}\left(\log_\sigma n\right)^D n^{\log_\sigma q+1} C(q,\sigma,n) + O\left(\left(\log n\right)^{D-1} n^{\log_\sigma q+1}\right),$$

where $C(q,\sigma,n) = \sum_{k\in\mathbb{Z}} n^{-\frac{2\pi\imath k}{\ln\sigma}}\Gamma\left(-\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}\right) = O\left(1\right)$ is a small, bounded fluctuating function.

Applications

For a $D = O(1)$ error search with an alphabet of size 4 we get

- Hamming Distance with $p = \frac{3}{4}$, $q = \frac{1}{4}$:

$$\frac{4 \cdot 3^D}{(D+1)!} (\log_4 n)^{D+1} + O\left((\ln n)^D\right)$$

- Number of don't-cares with $p = \frac{7}{16}$, $q = \frac{9}{16}$:

$$O\left((\ln n)^D n^{\log_4 \frac{9}{16}+1}\right) = O\left((\ln n)^D n^{0.59}\right)$$

- Hamming Distance with don't-cares with $p = \frac{3}{8}$, $q = \frac{5}{8}$:

$$O\left((\ln n)^D n^{\log_4 \frac{5}{8}+1}\right) = O\left((\ln n)^D n^{0.66}\right)$$

For the arithmetic distance the case $D = 0$ for various i is interesting. Let $\sigma = 24$, then $p = 1 - \frac{2i+1}{24}$, $q = \frac{2i+1}{24}$. In general

$$O\left(n^{\log_{24}\frac{2i+1}{24}+1}\right) = O\left(n^{\log_{24}(2i+1)}\right).$$

For

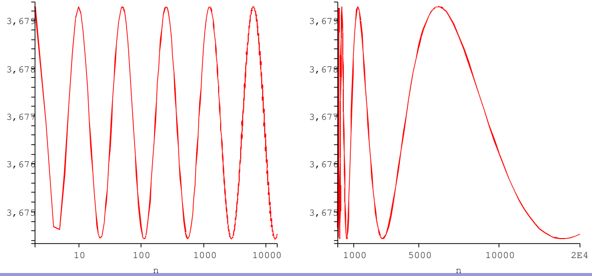
- $i = 1$ we get $O\left(n^{0.35}\right)$,
- $i = 2$ we get $O\left(n^{0.51}\right)$
- $i = 3$ we get $O\left(n^{0.61}\right)$

Hamming Distance in $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{N}\}$ with don't care symbol $\mathtt{N}$ has $p = \frac{12}{25}$, $q = \frac{13}{25}$:

$$\frac{25}{13}\frac{\left(\frac{12}{13}\right)^D}{D!}\left(\log_5 n\right)^D n^{\log_5 \frac{13}{5}} C\left(\frac{13}{25}, 5, n\right) + O\left(\left(\log n\right)^{D-1} n^{\log_5 \frac{13}{5}}\right)$$

$$\approx 1.92\frac{(0.4)^D}{D!}\left(\log_2 n\right)^D n^{0.59}\left(3.675 \pm 0.005\right) + O\left(\left(\log n\right)^{D-1} n^{0.59}\right)$$

Plots of $C\left(\frac{13}{25}, 5, n\right)$:



Average Complexity of the LS algorithm

The probability of k comparisons is

$$\Pr\left\{L_n^D = k\right\} = \sum_{i_1+\ldots+i_n=k} \prod_{j=1}^{n} \binom{i_j - 1}{D} p^{D+1} q^{i_j - D - 1}.$$

From it we can derive the probability generating function

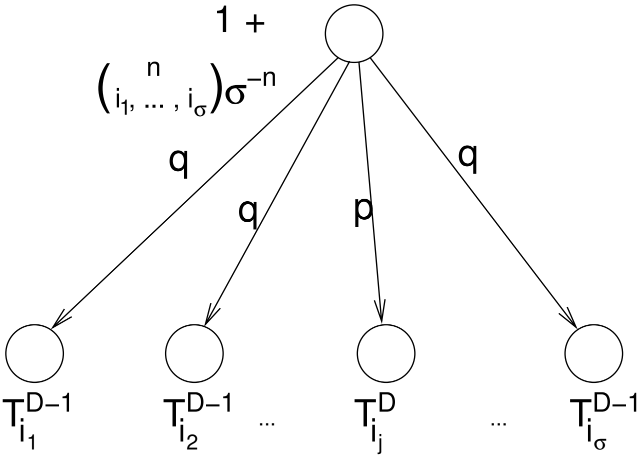
$$g_{L_n^D}(z) = \mathbf{E}\left[z^{L_n^D}\right] = \sum_{k=0}^{\infty} \Pr\left\{L_n^D = k\right\} z^k = \left(\frac{pz}{1-qz}\right)^{n(D+1)},$$

which yields the expected value $\mathbf{E}\left[L_n^D\right] = \frac{D+1}{p} n$.

Average Complexity of the TS algorithm

- Count the number of nodes visited by the search process ($\hat{=}$number of comparisons$+1$).
- The expected number of nodes is computed recursively, summing over all subtrees and distributions of strings.

Average Complexity of the TS algorithm



Average Complexity of the TS algorithm (2)

- Boundary conditions: $\mathbf{E}\left[T_n^{-1}\right] = 1$ (last mismatch) and $\mathbf{E}\left[T_0^D\right] = 0$ (no strings).
- Recursion:

$$\mathbf{E}\left[T_n^D\right] = 1 + \sum_{i_1+\cdots+i_\sigma=n} \binom{n}{i_1,\ldots,i_\sigma} \sigma^{-n} \left(\sum_{j=1}^{\sigma} p\mathbf{E}\left[T_{i_j}^{D-1}\right] + \sum_{j=1}^{\sigma} q\mathbf{E}\left[T_{i_j}^D\right] \right)$$

- For $n = 1$ we have $\mathbf{E}\left[T_1^D\right] = 1 + \frac{D+1}{p}$.

Average Complexity of the TS algorithm (3)

Compute the EGF $t^D(z)$, multiply by e^{-z}, define $\tilde{t}^D(z) = t^D(z)e^{-z}$, compare coefficients and find that for $n > 1$

$$y_n^D = \frac{(-1)^{n-1}}{1-\sigma^{1-n}q} + \frac{\sigma^{1-n}p}{1-\sigma^{1-n}q} y_n^{D-1},$$

with Boundary condition $y_n^{-1} = (-1)^{n-1}$ for $n > 0$ and $y_1^D = 1 + (D+1)/p$, $y_0^D = 0$. We get

$$y_n^D = \frac{(-1)^n \sigma^{1-n}}{1-\sigma^{1-n}} \left(\frac{\sigma^{1-n}p}{1-\sigma^{1-n}q} \right)^{D+1} - \frac{(-1)^n}{1-\sigma^{1-n}}.$$

Average Complexity of the TS algorithm (4)

We translate back to

$$\mathbf{E}\left[T_n^D\right] = n\left(1 + \frac{D+1}{p}\right) + \underbrace{\sum_{k=2}^{n}\binom{n}{k}\frac{(-1)^k}{\sigma^{k-1}-1}\left(\frac{p\sigma^{1-k}}{1-q\sigma^{1-k}}\right)^{D+1}}_{\mathcal{S}_n^{(D)}} - \underbrace{\sum_{k=2}^{n}\binom{n}{k}\frac{(-1)^k}{1-\sigma^{1-k}}}_{A_n}.$$

Average Compression Number

A similar derivation to the above shows that the sum A_n is the solution to

$$A_n = n - 1 + \sum_{i_1+\cdots+i_\sigma=n} \binom{n}{i_1,\ldots,i_\sigma} \sigma^{-n} \sum_{j=1}^{\sigma} A_{i_j},$$

which we call the average "compression number".

Lemma

The asymptotic behavior of A_n is

$$A_n = n\log_\sigma n + n\left(\frac{1}{2} - \frac{1-\gamma}{\ln\sigma} + \frac{\sum_{k\in\mathbb{Z}\setminus\{0\}} n^{-\frac{2\pi\imath k}{\ln\sigma}}\Gamma\left(-1+\frac{2\pi\imath k}{\ln\sigma}\right)}{\ln\sigma}\right) + O(1).$$

Rice's Formula

Let $f(z)$ be an analytic continuation of $f(k) = f_k$ that contains the half line $[m, \infty)$. Then

$$\sum_{k=m}^{n} (-1)^k \binom{n}{k} f_k = \frac{(-1)^n}{2\pi\imath} \int_{\mathcal{C}} f(z) \frac{n!}{z(z-1)\cdots(z-n)} dz,$$

where $\mathcal{C}$ is a positively oriented curve that encircles $[m, n]$ and does not include any of the integers $0, 1, \ldots, m-1$ or other singularities of $f(z)$.
($\Rightarrow$ Nörlund 1924)

We apply Rice's formula, let $\mathcal{C}$ grow to a large half-circle and find for $1 < \xi < 2$

$$\mathfrak{S}_n^{(D)} = \frac{1}{2\pi\imath} \int_{-\xi-\imath\infty}^{-\xi+\imath\infty} \frac{1}{\sigma^{-1-z}-1} \left(\frac{p}{\sigma^{-1-z}-q} \right)^{D+1} \mathrm{B}(n+1,z) dz + O(1) \, .$$

Since

$$\pi|z| \left| \frac{1}{\sigma^{-1-z}-1} \left(\frac{p}{\sigma^{-1-z}-q} \right)^{D+1} \mathrm{B}(n+1,z) \right| \xrightarrow[|z|\to\infty]{} 0$$

The integral needs to be simplified further by approximation of the Beta function:

$$\mathrm{B}(n+1,z)=\frac{\Gamma(n+1)\Gamma(z)}{\Gamma(n+1+z)}=\Gamma(z)n^{-z}+O\left(n^{-z-1}|z|^2\right).$$

This approximation is uniformly valid, for $\left(|z|^2\right)=o(n)$ (Tricomi and Erdélyi 1951, Fields 1970).

For $x<0$ and any strictly positive function $f(n)\in\omega(1)$ we have

$$\int_{f(n)\ln n}^{\infty}|\mathrm{B}(n,x+\imath y)|\,dy=O\left(n^{-f(n)\left(\frac{\pi}{4}-\epsilon\right)-x}\right).$$

For constant $x\notin\{0,-1,-2,\dots\}$ we have

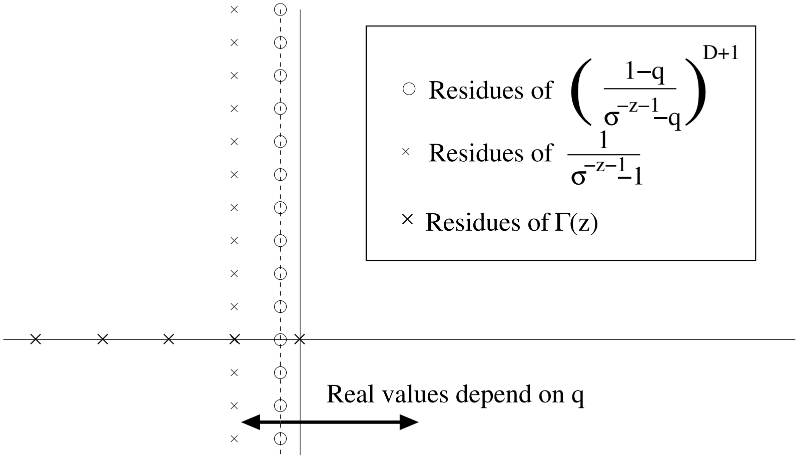
$$\int_{-\infty}^{\infty}|\mathrm{B}(n,x+\imath y)|\,dy=O\left(n^{-x}\right).$$

We are left with

$$\mathcal{I}_{\xi,n}^{(D)} := \frac{1}{2\pi\imath}\int_{-\xi-\imath\infty}^{-\xi+\imath\infty} \frac{1}{\sigma^{-1-z}-1}\left(\frac{p}{\sigma^{-z-1}-q}\right)^{D+1}\Gamma(z)n^{-z}dz,$$

which we can evaluate by the residues to the right of $\Re(z) = -\xi$.

Residues in the complex plane



The residues at $\Re(z) = -1$, A_n, and the starting terms cancel out.

$$-\left(\sum_{k\in\mathbb{Z}} \operatorname{res}\left[g(z), z = -1 + \frac{2\pi\imath k}{\ln\sigma}\right]\right) + n\left(1 + \frac{D+1}{p}\right) - A_n = O(1) .$$

Highest Order Term

We consider D, q, p, σ constant, the residues at $z = 0$ and at $\Re(z) = -\log_\sigma q - 1$ yield a multi-index sum of which we look at the term of highest order.
If $q = \sigma^{-1}$, this term is

$$-\frac{\sigma(\sigma-1)^D}{(D+1)!}\left(\log_\sigma n\right)^{D+1},$$

otherwise, this term is

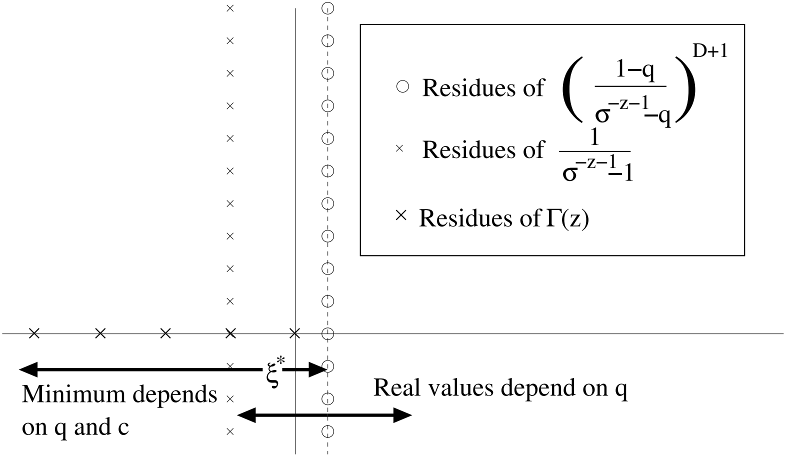
$$-\frac{(1-q)^D}{D!q^{D+1}}\left(\log_\sigma n\right)^D n^{\log_\sigma q+1}\sum_{k\in\mathbb{Z}} n^{-\frac{2\pi\imath k}{\ln\sigma}}\Gamma\left(-\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}\right).$$

Note that $\left|\sum_{k\in\mathbb{Z}} n^{-\frac{2\pi\imath k}{\ln\sigma}}\gamma_{l-i}^{\left(-\log_\sigma q-1+\frac{2\pi\imath k}{\ln\sigma}\right)}\right| = O\left(1\right).$

Assume $D+1 = c\log_\sigma n$, we can bound the integral by

$$\mathcal{I}_{\xi,n}^{(D)} \leq \frac{C}{\sigma^{\xi-1}-1} n^{\overbrace{c\log_\sigma\left(\frac{p}{\sigma^{\xi-1}-q}\right)+\xi}^{\varepsilon_{c,q,\xi}}}.$$

Residues in the complex plane (2)



Sublinear behavior for $c < p$

If the exponent $\mathcal{E}_{c,q,\xi}$ has a minimum $\xi^* < 1$, we are either left with a term $O\left(n^{\epsilon}\right)$ or we evaluate the remaining residues. This is the case if $c < p$.
For $\xi^* < 0$ we have an additional residue for the Gamma function at $z = 0$, but it is $o(n)$ for $D + 1 = c\log_{\sigma} n$.

Outlook

- Search bounded in multiple parameters.
- For (very) small D the method might be used to estimate the complexity for Edit Distance.
- Extension to indices with look-up time linear in the size of the pattern. The average size should behave similar (i.e., $O\left(n \operatorname{polylog}(n)\right)$).

Thank you!

Average-Case Analysis of Approximate Trie Search

Moritz G. Maaß

maass@in.tum.de
Institut für Informatik
Technische Universität München

15th CPM, July 2004