Asymmetric Rényi Problem and PATRICIA Tries

AofA 2016

Michael Drmota, Abram Magner, Wojciech Szpankowski

July 3, 2016

Rényi's Problem

We have a set X of objects and a set A of labels, along with a **hidden** bijective labeling $\phi : X \to A$.

Algorithmic task: Recover the labeling ϕ using as few queries as possible.

A **query** take the form of a subset B of labels. An answer to the query B is the set of objects with labels in B: $\phi^{-1}(B)$.

Rényi's problem in a nutshell: How many **random queries** are needed to recover ϕ in its entirety?

Object: x	Label: $\phi(x)$
1	d
2	e
3	a
4	c
5	b

Query: B	Response: $\phi^{-1}(B)$
{ a, c, e}	{2, 3, 4 }
{ d }	{ 1 }
{ a, b, c, d }	{ 1, 3, 4, 5 }

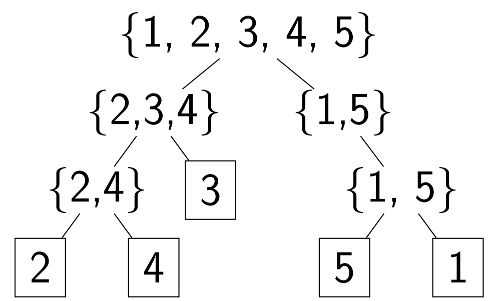
Partition Refinement Tree View of the Rényi Process

A sequence of queries corresponds to a refinement of partitions of the item set:

Example:

$\phi : 1 \to d, 2 \to e, 3 \to a, 4 \to c, 5 \to b$.

1. $B_1 = \{b, d\} \mapsto \{1, 5\}$
2. $B_2 = \{a, b, d\} \mapsto \{1, 3, 5\}$,
3. $B_3 = \{a, c, d\} \mapsto \{1, 3, 4\}$,



Level $j \geq 0$ of the tree $\iff$ partition $\mathcal{P}_j$.

Right child node $\iff$ subset of objects in parent set contained in the response to the jth query.

Singletons are only explicitly depicted in the first level in which they appear.

Probabilistic Models for Queries

Original query model: fix a bias $p \geq 1/2$. Independently, include each label in a query with probability p.

Probabilistic Models for Queries

Original query model: fix a bias $p \geq 1/2$. Independently, include each label in a query with probability p.

An inefficiency: some queries are **inconclusive**. They may not split all partition elements: e.g.,

$$\mathcal{P}_j = \{\{2,3,4\},\{1,5\}\}, \qquad \phi^{-1}(B_j) = \{1,3,5\}, \quad \text{or} \qquad \phi^{-1}(B_j) = \{2,4\}.$$

Probabilistic Models for Queries

Original query model: fix a bias $p \geq 1/2$. Independently, include each label in a query with probability p.

An inefficiency: some queries are **inconclusive**. They may not split all partition elements: e.g.,

$$\mathcal{P}_j = \{\{2,3,4\},\{1,5\}\}, \qquad \phi^{-1}(B_j) = \{1,3,5\}, \quad \text{or} \qquad \phi^{-1}(B_j) = \{2,4\}.$$

Eliminate inconclusiveness by refining queries before asking them:

- Start with $B_{j,0}$, a query generated as usual: e.g., $\phi^{-1}(B_{j,0}) = \{1,3,5\}$.

Probabilistic Models for Queries

Original query model: fix a bias $p \geq 1/2$. Independently, include each label in a query with probability p.

An inefficiency: some queries are **inconclusive**. They may not split all partition elements: e.g.,

$$\mathcal{P}_j = \{\{2,3,4\},\{1,5\}\}, \qquad \phi^{-1}(B_j) = \{1,3,5\}, \quad \text{or} \qquad \phi^{-1}(B_j) = \{2,4\}.$$

Eliminate inconclusiveness by refining queries before asking them:

- Start with $B_{j,0}$, a query generated as usual: e.g., $\phi^{-1}(B_{j,0}) = \{1,3,5\}$.
- To construct $B_{j,i+1}$ from $B_{j,i}$: for each label in each partition element unsplit by $B_{j,i}$, decide again whether or not to include it independently with probability p: e.g., $\phi^{-1}(B_{j,1}) = \{3\}, \phi^{-1}(B_{j,2}) = \{3,5\}$.

Probabilistic Models for Queries

Original query model: fix a bias $p \geq 1/2$. Independently, include each label in a query with probability p.

An inefficiency: some queries are **inconclusive**. They may not split all partition elements: e.g.,

$$\mathcal{P}_j = \{\{2,3,4\},\{1,5\}\}, \qquad \phi^{-1}(B_j) = \{1,3,5\}, \quad \text{or} \qquad \phi^{-1}(B_j) = \{2,4\}.$$

Eliminate inconclusiveness by refining queries before asking them:

- Start with $B_{j,0}$, a query generated as usual: e.g., $\phi^{-1}(B_{j,0}) = \{1,3,5\}$.
- To construct $B_{j,i+1}$ from $B_{j,i}$: for each label in each partition element unsplit by $B_{j,i}$, decide again whether or not to include it independently with probability p: e.g., $\phi^{-1}(B_{j,1}) = \{3\}, \phi^{-1}(B_{j,2}) = \{3,5\}$.
- Query B_j is the end result, where all partition elements are split: e.g., $\phi^{-1}(B_j) = \{3,5\}$.

Refining Inconclusive Queries: PATRICIA Trie Correspondence

Important correspondence: Partition refinement tree for the Rényi process with inconclusive query refinement $\stackrel{D}{=}$ a PATRICIA trie on n infinite random binary strings.

An object $\iff$ a string, where 1 means that the label is in a query, and 0 means that it is not.

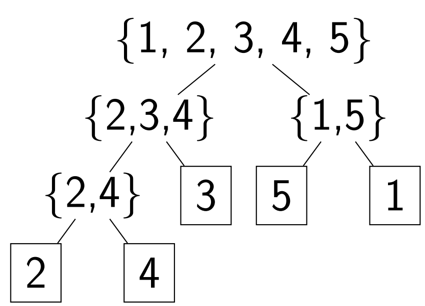
Example: $\phi : 1 \to d, 2 \to e, 3 \to a, 4 \to c, 5 \to b$.

Strings corresponding to objects:

1. d: $111\ldots$
2. e: $000\ldots$
3. a: $011\ldots$
4. c: $0001\ldots$
5. b: $110\ldots$

Queries corresponding to strings:

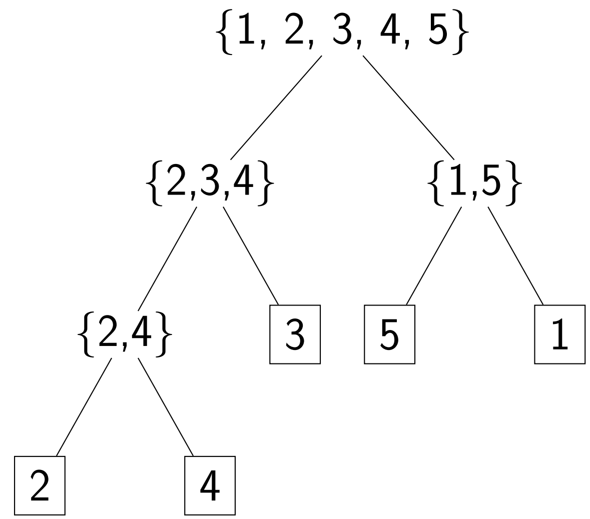
1. $B_1 = \{b, d\} \mapsto \{1, 5\}$
2. $B_2' = \{a, b, d\} \mapsto \{1, 3, 5\}$;
   $B_2 = \{a, d\} \mapsto \{1, 3\}$,
3. $B_3 = \{a, c, d\} \mapsto \{1, 3, 4\}$,



Parameters of Interest

Parameters of interest:

- Height (H_n): # of queries needed to recover ϕ entirely.
- Fillup level (F_n): # of queries needed before the first item-label pair is discovered.
- Typical depth (D_n): # of queries before a randomly chosen item's label is discovered.
- External profile at level k $(B_{n,k})$: Number of item-label pairs revealed by the kth query.



In the diagram: $H_5 = 3$, $F_5 = 2$, $\Pr[D_5 = 2] = 3/5, \Pr[D_5 = 3] = 2/5$, $B_{5,2} = 3$, $B_{5,3} = 2$.

Our Results

We have the following asymptotic expansions for the typical values of H_n and F_n:

Theorem 1 (Asymptotics for F_n and H_n)**.** *With high probability,*

$$H_n = \begin{cases} \log_{1/p} n + \frac{1}{2}\log_{p/q}\log n + o(\log\log n) & p > q = 1-p \\ \log_2 n + \sqrt{2\log_2 n} + o(\sqrt{\log n}) & p = q = 1/2 \end{cases} \tag{1}$$

and

$$F_n = \begin{cases} \log_{1/q} n - \log_{1/q}\log\log n + o(\log\log\log n) & p > q = 1-p \\ \log_2 n - \log_2\log n + o(\log\log n) & p = q = 1/2 \end{cases} \tag{2}$$

for large n*.*

Symmetric case ($p = 1/2$) was known, but **asymmetric case (**$p > 1/2$**) is new!**

Note the phase transition in the second term.

We also have **results for** D_n **via the external profile.**

Prior Work

Pittel & Rubin (1990): "How many random questions are needed to identify n distinct objects?": Two-term asymptotics for H_n in the symmetric case ($p = 1/2$) via (different) GF methods.

Prior Work

Pittel & Rubin (1990): "How many random questions are needed to identify n distinct objects?": Two-term asymptotics for H_n in the symmetric case ($p = 1/2$) via (different) GF methods.

Devroye (1992): "A note on the probabilistic analysis of PATRICIA trees": Two-term asymptotics for H_n and F_n in the symmetric case via more probabilistic methods.

Prior Work

Pittel & Rubin (1990): "How many random questions are needed to identify n distinct objects?": Two-term asymptotics for H_n in the symmetric case ($p = 1/2$) via (different) GF methods.

Devroye (1992): "A note on the probabilistic analysis of PATRICIA trees": Two-term asymptotics for H_n and F_n in the symmetric case via more probabilistic methods.

Park, Hwang, Nicodème, Szpankowski (2009): "Profile of tries": Precisely analyzed trie profiles via the Poisson transform/Mellin transform/Inverse Mellin via saddle point method pipeline.

Prior Work

Pittel & Rubin (1990): "How many random questions are needed to identify n distinct objects?": Two-term asymptotics for H_n in the symmetric case ($p = 1/2$) via (different) GF methods.

Devroye (1992): "A note on the probabilistic analysis of PATRICIA trees": Two-term asymptotics for H_n and F_n in the symmetric case via more probabilistic methods.

Park, Hwang, Nicodème, Szpankowski (2009): "Profile of tries": Precisely analyzed trie profiles via the Poisson transform/Mellin transform/Inverse Mellin via saddle point method pipeline.

Drmota & Szpankowski (2011): "The expected profile of digital search trees": Similar to above, but for digital search tree profiles.

Prior Work

Pittel & Rubin (1990): "How many random questions are needed to identify n distinct objects?": Two-term asymptotics for H_n in the symmetric case ($p = 1/2$) via (different) GF methods.

Devroye (1992): "A note on the probabilistic analysis of PATRICIA trees": Two-term asymptotics for H_n and F_n in the symmetric case via more probabilistic methods.

Park, Hwang, Nicodème, Szpankowski (2009): "Profile of tries": Precisely analyzed trie profiles via the Poisson transform/Mellin transform/Inverse Mellin via saddle point method pipeline.

Drmota & Szpankowski (2011): "The expected profile of digital search trees": Similar to above, but for digital search tree profiles.

Magner Ph.D. thesis / Magner & Szpankowski (2015): "Profiles of PATRICIA tries": Precisely analyzed distribution of the external profile in the central range.

This work: Requires **extension** of the external profile analysis to the **boundaries of the central range**.

Comparison with Tries and DSTs

Only the first terms of F_n and H_n for tries and DSTs with $p > q$ are given in the literature!

For tries:

$$H_n \sim \frac{2}{\log(1/(p^2+q^2))} \log n, \qquad F_n \sim \log_{1/q} n.$$

For DSTs:

$$H_n \sim \log_{1/p} n, \qquad F_n \sim \log_{1/q} n.$$

Proof Sketch: Derivation of Height

Goal: Define $k_* = k_*(n) = \log_{1/p} n + \psi(n)$. We want to determine $\psi(n)$ for which

$$H_n = k_* + o(\psi(n)).$$

First, we connect H_n with $B_{n,k}$: $H_n = \max\{k \ : \ B_{n,k} > 0\}$.

Then connect H_n to moments of $B_{n,k}$ via the first and second moment methods:

$$\Pr[H_n > k] \leq \sum_{j>k} \mathbb{E}[B_{n,j}] \qquad\qquad \Pr[H_n < k] \leq \frac{\mathbf{Var}[B_{n,k}]}{\mathbb{E}[B_{n,k}]^2}.$$

So we want $\psi(n)$ to satisfy

$$\mathbb{E}[B_{n,\log_{1/p} n+(1-\epsilon)\psi(n)}] \xrightarrow{n\to\infty} \infty, \qquad\qquad \mathbb{E}[B_{n,\log_{1/p} n+(1+\epsilon)\psi(n)}] \xrightarrow{n\to\infty} 0.$$

Derivation of the Height: External Profile Analysis

Ranges of behavior of $\mathbb{E}[B_{n,k}]$:

- Magner/Magner & Szpankowski (2015): Central range.

$$k \sim \alpha \log n, \qquad \alpha \in \left(\frac{1}{\log(1/q)} + \epsilon, \frac{1}{\log(1/p)} - \epsilon \right).$$

- **This work:** Boundaries of the central range.

$$H_n : k \sim \log_{1/p} n, \qquad F_n : k \sim \log_{1/q} n.$$

$$k = \log_{1/q} n \qquad\qquad k = \log_{1/p} n$$

$$k = \log_{1/q} n - \log_{1/q} \log\log n \qquad\qquad k = \log_{1/p} n + \tfrac{1}{2} \log_{p/q} \log n$$

External Profile Analysis (continued)

Basic tool chain for profile analysis:

$$\boxed{\text{Poissonization}} \longrightarrow \boxed{\text{Mellin}} \longrightarrow \boxed{\text{Inverse Mellin}} \longrightarrow \boxed{\text{De-Poissonization}}$$

Poisson transform $\tilde{G}_k(z)$ for $\mathbb{E}[B_{n,k}]$:

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + e^{-pz}(\tilde{G}_k(qz) - \tilde{G}_{k-1}(qz)) + e^{-qz}(\tilde{G}_k(pz) - \tilde{G}_{k-1}(pz)).$$

External Profile Analysis (continued)

Basic tool chain for profile analysis:

$$\boxed{\text{Poissonization}} \longrightarrow \boxed{\text{Mellin}} \longrightarrow \boxed{\text{Inverse Mellin}} \longrightarrow \boxed{\text{De-Poissonization}}$$

Poisson transform $\tilde{G}_k(z)$ for $\mathbb{E}[B_{n,k}]$:

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + e^{-pz}(\tilde{G}_k(qz) - \tilde{G}_{k-1}(qz)) + e^{-qz}(\tilde{G}_k(pz) - \tilde{G}_{k-1}(pz)).$$

Explicit formula for Mellin transform of $\tilde{G}_k(z)$:

$$G_k^*(s) := \int_0^\infty z^{s-1}\tilde{G}_k(z)\,\mathrm{d}z = (p^{-s} + q^{-s})^k A_k(s)\Gamma(s+1).$$

Fundamental strip for $\tilde{G}_k(z)$: $\Re(s) > -k-1$.

Inverting the Mellin Transform

Main new challenge of our analysis: estimate $\tilde{G}_k(n)$ by bounding inverse Mellin transform of $G_k^*(s)$:

$$\tilde{G}_k(z) = \frac{1}{2\pi i} \int_{\rho - i\infty}^{\rho + i\infty} z^{-s} G_k^*(s) \, \mathrm{d}s = \int_{\rho - i\infty}^{\rho + i\infty} J_k(z, s) \, \mathrm{d}s.$$

where $\rho > -k - 1$ and

$$J_k(n, s) = \sum_{j=0}^{k} n^{-s} (p^{-s} + q^{-s})^{k-j} \sum_{m \geq j} (p^m + q^m)(\mathbb{E}[B_{m,j}] - \mathbb{E}[B_{m,j-1}]) \frac{\Gamma(m+s)}{\Gamma(m+1)}.$$

Main Steps of Upper Bounding the Inverse Mellin Integral

$$\boxed{J_k(n,s) = \sum_{j=0}^{k} n^{-s}(p^{-s}+q^{-s})^{k-j} \sum_{m\geq j}(p^m+q^m)(\mathbb{E}[B_{m,j}] - \mathbb{E}[B_{m,j-1}])\frac{\Gamma(m+s)}{\Gamma(m+1)}.}$$

- Exponential decay of $\Gamma(s+1)A_k(s)$ along vertical lines $\Longrightarrow$

$$|\tilde{G}_k(n)| = O(J_k(n,\rho))$$

for $\rho > -k-1$.

Main Steps of Upper Bounding the Inverse Mellin Integral

$$\boxed{J_k(n,s) = \sum_{j=0}^{k} n^{-s}(p^{-s}+q^{-s})^{k-j} \sum_{m\geq j}(p^m+q^m)(\mathbb{E}[B_{m,j}] - \mathbb{E}[B_{m,j-1}])\frac{\Gamma(m+s)}{\Gamma(m+1)}.}$$

- Exponential decay of $\Gamma(s+1)A_k(s)$ along vertical lines $\implies$

$$|\tilde{G}_k(n)| = O(J_k(n,\rho))$$

  for $\rho > -k-1$.

- Estimate each jth term of $J_k(n,\rho)$ using upper bound on $\mathbb{E}[B_{m,j}]$ for j sufficiently close to m:

$$\mathbb{E}[B_{m,j}] \leq C\frac{m!}{(m-j-1)!}p^{j^2/2+j/2+o(j)}.$$

Main Steps of Upper Bounding the Inverse Mellin Integral

$$\boxed{J_k(n,s) = \sum_{j=0}^{k} n^{-s}(p^{-s}+q^{-s})^{k-j} \sum_{m\geq j}(p^m+q^m)(\mathbb{E}[B_{m,j}] - \mathbb{E}[B_{m,j-1}])\frac{\Gamma(m+s)}{\Gamma(m+1)}.}$$

- Exponential decay of $\Gamma(s+1)A_k(s)$ along vertical lines $\Longrightarrow$

$$|\tilde{G}_k(n)| = O(J_k(n,\rho))$$

  for $\rho > -k-1$.

- Estimate each jth term of $J_k(n,\rho)$ using upper bound on $\mathbb{E}[B_{m,j}]$ for j sufficiently close to m:

$$\mathbb{E}[B_{m,j}] \leq C\frac{m!}{(m-j-1)!}p^{j^2/2+j/2+o(j)}.$$

- Maximize resulting upper bound over all j:

$$\boxed{J_k(n,s) \leq p^{\nu(n,s)},}$$

  where

$$\nu(n,s) = -\frac{(s+\log_{1/p}(1+(p/q)^s)+\psi(n)+1)^2}{2}$$
$$-\log_{1/p} n \log_{1/p}(1+(p/q)^s) + \psi(n)^2/2 + o(\psi(n)^2).$$

Tightening the Upper Bound on the Poisson Transform

We now know

$$\boxed{\tilde{G}_k(n) = O(J_k(n,s)) \le p^{\nu(n,s)}}$$

for $s \in (-k-1, 0)$.

Tighten the upper bound by minimizing over s:

- $p = q = 1/2$: $\log_{1/p}(1 + (p/q)^s) = 1$, and we get

$$s_* = -\psi(n) + O(1), \qquad \nu(n, s_*) = -\log_2 n + \psi(n)^2/2 + o(\psi(n)^2).$$

Tightening the Upper Bound on the Poisson Transform

We now know

$$\boxed{\tilde{G}_k(n) = O(J_k(n,s)) \le p^{\nu(n,s)}}$$

for $s \in (-k-1, 0)$.

Tighten the upper bound by minimizing over s:

- $p = q = 1/2$: $\log_{1/p}(1 + (p/q)^s) = 1$, and we get

$$s_* = -\psi(n) + O(1), \qquad \nu(n, s_*) = -\log_2 n + \psi(n)^2/2 + o(\psi(n)^2).$$

- $p > q$: $\log_{1/p}(1 + (p/q)^s)$ is a function of s. Lambert W function asymptotics + algebra $\Longrightarrow$

$$s_* = -\log_{p/q} \log n + O(\log\log\log n).$$

Determining the Second Term of H_n

We now know

$$\boxed{\tilde{G}_k(n) \le p^{\nu(n,s_*)}.}$$

for an optimal s_* exhibiting a phase transition w.r.t. p.

Determining the Second Term of H_n

We now know

$$\boxed{\tilde{G}_k(n) \leq p^{\nu(n,s_*)}.}$$

for an optimal s_* exhibiting a phase transition w.r.t. p.

Determining a candidate $\psi(n)$:

- for $p^{\nu(n,s_*)}$ to tend to 0, ∞ w.r.t. n, need $\nu(n, s_*) \to +\infty$, $-\infty$, respectively.

Determining the Second Term of H_n

We now know

$$\boxed{\tilde{G}_k(n) \le p^{\nu(n,s_*)}.}$$

for an optimal s_* exhibiting a phase transition w.r.t. p.

Determining a candidate $\psi(n)$:

- for $p^{\nu(n,s_*)}$ to tend to 0, ∞ w.r.t. n, need $\nu(n, s_*) \to +\infty, -\infty$, respectively.
- So we need $\psi(n)$ to be such that $\nu(n, s_*) = 0$.

Determining the Second Term of H_n

We now know

$$\boxed{\tilde{G}_k(n) \le p^{\nu(n,s_*)}.}$$

for an optimal s_* exhibiting a phase transition w.r.t. p.

Determining a candidate $\psi(n)$:

- for $p^{\nu(n,s_*)}$ to tend to $0, \infty$ w.r.t. n, need $\nu(n, s_*) \to +\infty, -\infty$, respectively.
- So we need $\psi(n)$ to be such that $\nu(n, s_*) = 0$.

$$\psi(n) = \begin{cases} \sqrt{\log_2 n} + o(\sqrt{\log n}) & p = q = 1/2 \\ \frac{1}{2}\log_{p/q} \log n + O(\log\log\log n) & p > q. \end{cases}$$

Plugging in $k = \log_{1/p} n + (1 \pm \epsilon)\psi(n)$ for the upper bound on $\tilde{G}_k(n)$ gives

$$p^{\frac{\epsilon}{2}(\log_{p/q}\log n)^2 + o((\log\log n)^2)} \to 0, \qquad p^{-\frac{\epsilon}{2}(\log_{p/q}\log n)^2 + o((\log\log n)^2)} \to \infty,$$

as desired.

Possible Future Directions

- More precise asymptotics/limit laws for H_n and F_n. The limiting behavior for H_n is known for $p = q$ (Knessl & Szpankowski (1999): " Limit laws for height in generalized tries and PATRICIA tries"), but not for $p > q$.

Possible Future Directions

- More precise asymptotics/limit laws for H_n and F_n. The limiting behavior for H_n is known for $p = q$ (Knessl & Szpankowski (1999): " Limit laws for height in generalized tries and PATRICIA tries"), but not for $p > q$.
- More satisfying explanation of the phase transition for H_n in terms of the number and sizes of fringe subtrees.

Possible Future Directions

- More precise asymptotics/limit laws for H_n and F_n. The limiting behavior for H_n is known for $p = q$ (Knessl & Szpankowski (1999): " Limit laws for height in generalized tries and PATRICIA tries"), but not for $p > q$.
- More satisfying explanation of the phase transition for H_n in terms of the number and sizes of fringe subtrees.
- How does adding noise affect the process? One natural noise model: Items are dropped randomly from responses to queries, or extra items are added.

Thank you!