Fringe Trees of Patricia Tries

Jasper Ischebeck
Goethe University Frankfurt a. M.
AofA 2023 Taipei
arXiv:2305.14900

June 28, 2023

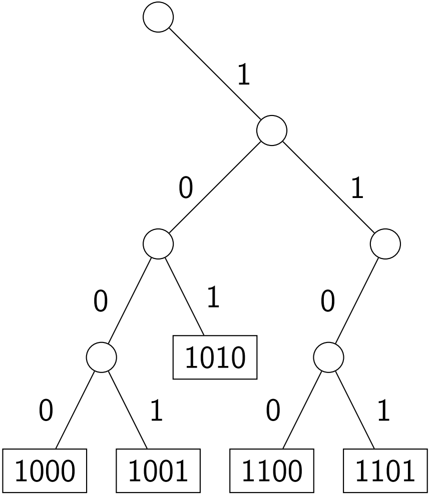
Introduction

- Patricia tries are data structures used to store and retrieve strings
- Fixed finite alphabet $\mathscr{A}$
- Patricia tries are subtrees of $\mathscr{A}^*$ (seen as a labeled, infinite tree)
- Sample i.i.d. infinite strings (=sequences) with each character i.i.d. with a distribution p on $\mathscr{A}$
- For any string $\alpha = a_1 \dots a_n$ write $p_\alpha := p(\{a_1\}) \dots p(\{a_n\})$

Construction of the Trie

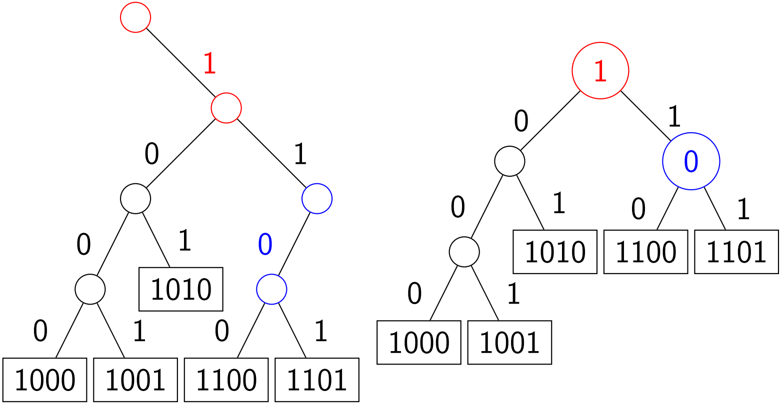
Start with a set $\mathfrak{X}$ of strings

- If $\mathfrak{X} = \emptyset$, the trie is empty.
- If $|\mathfrak{X}| = 1$, we store the string in a leaf and are finished
- Else we split $\mathfrak{X}$ on the first character of the string and have a trie as subtree for every starting character.



Patricia Trie

By compressing the nodes of a trie T with only one child into chains, we get the patricia trie $\mathrm{pat}\, T$.



- We can see pat as a function from tries to patricia tries
- Write $\mathscr{T}_n$ for the trie from n i.i.d. strings
- Write $\mathscr{P}_n := \mathrm{pat}\,\mathscr{T}_n$ for the patricia trie of n i.i.d. strings

Properties

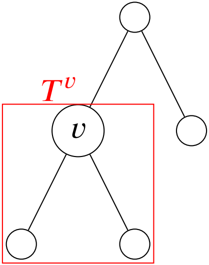
- Patricia tries were introduced in 1968 independently by Morrison (1968) and Gwehenberger (1968)
- Practical Algorithm To Retrieve Information Coded In Alphanumeric, Trie is from ReTRIEval.
- Tries and patricia tries as well as some key properties included in Knuth's Art of Computer Programming (Knuth 1973)
- Since then, many other properties have been studied, e.g. the number of visited nodes in a search (Szpankowski 1990) or the profile (Devroye 2005) etc. for many sources of random strings
- Because of the similarity, patricia tries and tries can often be handled with the same methods

Patricia Tries

- For tries, there are efforts to handle multiple properties at once, for example Fuchs, Hwang, and Zacharovas (2014) using analytic methods.
- Janson gives a general theorem for additive functionals in 2020 using probabilistic methods
- We have shown how to reduce properties of patricia tries to tries and leverage these results.

Fringe Trees

- Let T be a tree
- For $v \in T$ the *fringe tree* T^v is the subtree consisting of v and its descendants in T
- The *random fringe tree* T^* is the fringe tree T^v of a uniformly chosen $v \in T$.



Additive functionals

- Let φ be a function on trees to $\mathbb{R}$, called *toll function*
- Then Φ defined by

$$\Phi(T) := \sum_{v \in T} \varphi(T^v)$$

is its *additive functional*.

Additive functionals

- Let φ be a function on trees to $\mathbb{R}$, called *toll function*
- Then Φ defined by

$$\Phi(T) := \sum_{v \in T} \varphi(T^v)$$

  is its *additive functional.*
- For an additive functional Φ we can define its pullback on tries as $\widehat{\Phi}(T) = \Phi(\operatorname{pat} T)$.
- Let $\widehat{\varphi}$ be the toll function for $\widehat{\Phi}$.
- We call Φ *increasing* if $\widehat{\Phi}(T) \leq \widehat{\Phi}(T')$ for trees $T \subseteq T'$
- If φ is bounded, $\widehat{\varphi}$ is also bounded.

CLT with all moments

We say $X_n \overset{d}{\approx} Y_n$ with all moments if $\mathbb{E}[f(X_n) - f(Y_n)] \to 0$ for every bounded continuous function f and for $f(x) = x^a, a \in \mathbb{R}$.

Theorem (CLT with all moments)

For an increasing additive functional Φ with a bounded toll function, and thus also for the difference of two such Φ, we have approximation in distribution

$$\frac{\Phi(\mathscr{P}_n) - \mathbb{E}[\Phi(\mathscr{P}_n)]}{\sqrt{n}} \overset{d}{\approx} N(0, \sigma^2(\log n)),$$

with all moments, where σ^2 is a bounded function that is $\log(p_a)$-periodic for every $a \in \mathscr{A}$.

CLT – Notes

- Because we have convergence of all moments we get a strong law of large numbers as corollary. (answering a problem in Janson (2022))
- "Bounded toll function" can be relaxed to toll functions with variance and mean of order $O(n^{1-\varepsilon})$
- With some criteria we have $\sigma^2(t) > 0$ for all t and can thus move $\sigma^2(\log n)$ into the denominator, giving convergence to N(0,1)

CLT – Notes

- Because we have convergence of all moments we get a strong law of large numbers as corollary. (answering a problem in Janson (2022))
- "Bounded toll function" can be relaxed to toll functions with variance and mean of order $O(n^{1-\varepsilon})$
- With some criteria we have $\sigma^2(t) > 0$ for all t and can thus move $\sigma^2(\log n)$ into the denominator, giving convergence to N(0,1)
- Because σ^2 is $\log(p_a)$-periodic for every $a \in \mathscr{A}$, it is also d-periodic for d the smallest common divisor of $\{\log(p_a) : a \in \mathscr{A}\}$.
- Thus, if $d = 0$ (the *non-arithmetic case*), σ^2 is constant

CLT – Notes

- Because we have convergence of all moments we get a strong law of large numbers as corollary. (answering a problem in Janson (2022))
- "Bounded toll function" can be relaxed to toll functions with variance and mean of order $O(n^{1-\varepsilon})$
- With some criteria we have $\sigma^2(t) > 0$ for all t and can thus move $\sigma^2(\log n)$ into the denominator, giving convergence to N(0,1)
- Because σ^2 is $\log(p_a)$-periodic for every $a \in \mathscr{A}$, it is also d-periodic for d the smallest common divisor of $\{\log(p_a) : a \in \mathscr{A}\}$.
- Thus, if $d = 0$ (the *non-arithmetic case*), σ^2 is constant
- σ^2 and the asymptotic behavior of $\mathbb{E}[\Phi(\mathscr{P}_n)]$ (also periodic) can be calculated with standard methods.

The induced toll function $\widehat{\varphi}$

- What is $\widehat{\varphi}$?
- From the definition,

$$\widehat{\varphi}(T) = \widehat{\Phi}(T) - \sum_{a \in \mathscr{A}} \widehat{\Phi}(T^a).$$

The induced toll function $\widehat{\varphi}$

- What is $\widehat{\varphi}$?
- From the definition,

$$\widehat{\varphi}(T) = \widehat{\Phi}(T) - \sum_{a \in \mathscr{A}} \widehat{\Phi}(T^a).$$

- If the root of T has exactly one child $a \in \mathscr{A}$, then the root gets compressed: $\operatorname{pat} T = \operatorname{pat}(T^a)$ and thus $\widehat{\varphi}(T) = 0$
- If not, the tree splits normally and $(\operatorname{pat} T)^a = \operatorname{pat}(T^a)$ for all $a \in \mathscr{A}$, so $\widehat{\varphi}(T) = \varphi(\operatorname{pat} T)$.
- So,

$$\widehat{\varphi}(T) = \varphi(\operatorname{pat} T)\mathbf{1}\{T\text{'s root has not exactly one child.}\}$$

Asymptotic moments

- First assume φ is zero for leaves ($\{\varepsilon\}$).
    - The contribution to Φ is deterministically a multiple of the amount of strings
- By looking at the trie $\widetilde{\mathscr{T}}_\lambda$ with an independent, Pois(λ)-distributed amount of strings, the subtrees also become independent tries with Poisson distributed amounts of strings
- The moments of $\widehat{\Phi}(\widetilde{\mathscr{T}}_\lambda)$ are then sums of the form $\sum_{\alpha\in\mathscr{A}^*} f(p_\alpha\lambda)$ with a function f.
- This is called poissonization

Asymptotic moments

- The function f in $\sum_{\alpha \in \mathscr{A}^*} f(p_\alpha \lambda)$ is...
- For the expectation:

$$f_E(\lambda) = \mathbb{E}[\widehat{\varphi}(\widetilde{\mathscr{T}}_\lambda)]$$

- For the variance:

$$f_V(\lambda) = 2\,\mathrm{Cov}\Big(\widehat{\varphi}(\widetilde{\mathscr{T}}_\lambda), \widehat{\Phi}(\widetilde{\mathscr{T}}_\lambda)\Big) - \mathrm{Var}\Big(\widehat{\varphi}(\widetilde{\mathscr{T}}_\lambda)\Big).$$

- For the covariance with the amount N_λ of strings:

$$f_C(\lambda) = \mathrm{Cov}\Big(\widehat{\varphi}(\widetilde{\mathscr{T}}_\lambda), N_\lambda\Big)$$

- This method is well known. Clément, Flajolet, and Vallée (2001) lists three ways to revert this process and Janson's approach is yet another
- The asymptotics of such sums can be described with Mellin transforms, given as:

$$f_E^*(s) = \int_0^\infty f_E(\lambda)\lambda^{s-1} d\lambda.$$

- To revert the Mellin transformation and the poissonization one can use ...
  - analytic methods, such as in Fuchs, Hwang, and Zacharovas (2014) and Hwang, Fuchs, and Zacharovas (2010)
  - renewal theory and that Φ is increasing, as in Janson (2022)

Asymptotic moments

Theorem (Asymptotic moments; non-arithmetic)

For an increasing additive functional Φ on patricia tries with a bounded toll function φ and $\varphi(\{\varepsilon\}) = 0$, and thus also for the difference of two such Φ, the following holds:
If $d = 0$,

$$\mathbb{E}\big[\Phi(\mathscr{P}_n)\big] = \frac{n}{H} f_E^*(-1) + o(n)$$
$$\operatorname{Var}\big(\Phi(\mathscr{P}_n)\big) = \frac{n}{H} f_V^*(-1) - \frac{n}{H^2} f_E^*(-1)^2 + o(n),$$

where H is the Shannon entropy of p.

Asymptotic moments

Theorem (Asymptotic moments; arithmetic)

For an increasing additive functional Φ on patricia tries with a bounded toll function φ and $\varphi(\{\varepsilon\}) = 0$, and thus also for the difference of two such Φ, the following holds:
If $d > 0$ let $\chi_m := 2\pi i m/d$. Then

$$\mathbb{E}\left[\Phi(\mathscr{P}_n)\right] = \frac{n}{H} \sum_{m\in\mathbb{Z}} f_E^*(-1-\chi_m) n^{\chi_m} + o(n)$$

$$\operatorname{Var}\left(\Phi(\mathscr{P}_n)\right) = \frac{n}{H} \sum_{m\in\mathbb{Z}} f_V^*(-1-\chi_m) n^{\chi_m} - \frac{n}{H^2}\left(\sum_{m\in\mathbb{Z}} f_C^*(-1-\chi_m) n^{\chi_m}\right)^2 + o(n).$$

Size of fringe patricia tries

- The natural measure for "size" of a patricia trie is the amount of strings or equivalently of leaves
- Let $\varphi_{\geq k}(T)$ be the indicator that a tree T has at least $k \geq 2$ leaves / strings
- Then, the additive functional $\Phi_{\geq k}(\mathscr{P}_n)$ is the amount of fringe trees with at least k strings
- This additive functional is increasing

Size of fringe patricia tries

- The natural measure for "size" of a patricia trie is the amount of strings or equivalently of leaves
- Let $\varphi_{\geq k}(T)$ be the indicator that a tree T has at least $k \geq 2$ leaves / strings
- Then, the additive functional $\Phi_{\geq k}(\mathscr{P}_n)$ is the amount of fringe trees with at least k strings
- This additive functional is increasing
- Let $\Phi_k := \Phi_{\geq k} - \Phi_{\geq k-1}$ be the amount of fringe trees with exactly k strings
- We can then apply the CLT with all moments to Φ_k

Expected size of fringe patricia tries

- The induced toll function $\widehat{\varphi}_k$ is then "T has k strings that don't all start with the same character."
- So

$$f_{E,k}(\lambda) = \mathbb{E}\left[\widehat{\varphi}_k(\widetilde{\mathscr{T}}_\lambda)\right] = e^{-\lambda}\frac{\lambda^k}{k!}\Big(1 - \underbrace{\sum_{a\in\mathscr{A}} p_a^k}_{=:\,\rho(k)}\Big).$$

- And the Mellin transform is

$$\begin{aligned} f^*_{E,k}(s) &= \int_0^\infty \frac{1-\rho(k)}{k!} e^{-\lambda}\lambda^{k+s-1}d\lambda \\ &= \frac{1-\rho(k)}{k!}\Gamma(s+k) \end{aligned}$$

- The mean term is $f^*_{E,k}(-1) = \frac{1-\rho(k)}{k(k-1)}$.

Size of fringe patricia tries

Theorem (I. 2023)

For $n \to \infty$ and $k \geq 2$, we have for the amount $\Phi_k(\mathscr{P}_n)$ of fringe trees with k strings in a patricia trie from n strings,

$$\frac{\Phi_k(\mathscr{P}_n) - \mathbb{E}\left[\Phi_k(\mathscr{P}_n)\right]}{\sqrt{\mathrm{Var}\left(\Phi_k(\mathscr{P}_n)\right)}} \xrightarrow{d} \mathscr{N}(0,1) \tag{1}$$

Size of fringe patricia tries

Theorem (I. 2023)

For $n \to \infty$ and $k \geq 2$, we have for the amount $\Phi_k(\mathscr{P}_n)$ of fringe trees with k strings in a patricia trie from n strings,

$$\frac{\Phi_k(\mathscr{P}_n) - \mathbb{E}\left[\Phi_k(\mathscr{P}_n)\right]}{\sqrt{\mathrm{Var}\left(\Phi_k(\mathscr{P}_n)\right)}} \xrightarrow{d} \mathscr{N}(0,1) \tag{1}$$

The asymptotics of $\mathbb{E}\left[\Phi_k(\mathscr{P}_n)\right]$ are given by

$$\frac{1}{n}\mathbb{E}\left[\Phi_k(\mathscr{P}_n)\right] = \frac{1 - \sum_{a\in\mathscr{A}} p_a^k}{H k(k-1)} + \psi_k(\log n) + o(1), \tag{2}$$

where ψ_k is a bounded, periodic function.

Other additive functionals

- Bounded toll functions already cover many properties:
- The number of k-protected nodes (nodes whose fringe trees have no leaf with depth lesser equal k)
- The independence number, domination number etc.
- With a logarithmically growing toll function, we have the shape functional (subtree size product logarithm)

Conclusion

- We have seen a CLT for additive functionals with bounded toll function
- and an application to fringe tree amounts

Conclusion

- We have seen a CLT for additive functionals with bounded toll function
- and an application to fringe tree amounts

Open questions:

- What about digital search trees?
- How to generalize to bigger toll functions?

Conclusion

- We have seen a CLT for additive functionals with bounded toll function
- and an application to fringe tree amounts

Open questions:

- What about digital search trees?
- How to generalize to bigger toll functions?

Thanks