MÉMOIRE DE DEA

JULIEN CLÉMENT

Table des matières

Introduction 2
1. Analyse de complexité 2
2. Représentation des nombres 3
2.1. Système de numération en base 2 3
2.2. développement en fraction continue 5
3. Trie 6
3.1. Définitions 7
3.2. Exemples d'applications 7
3.3. Trie et représentation des nombres 9
3.4. Paramètres 10
4. Paramètres des tries de nombre 11
4.1. Notations. 11
4.2. Modèle probabiliste 12
4.3. Longueur de cheminement externe 13
4.4. Nombre de noeuds internes 13
4.5. Hauteur 14
5. Transformée de Mellin 15
5.1. Définitions 15
5.2. Propriétés fonctionnelles 15
5.3. Propriétés asymptotiques 16
5.4. Sommes harmoniques 17
6. Trie binaire : comportement asymptotique 18
6.1. Longueur de cheminement externe 18
6.2. Nombre de noeuds internes 20
6.3. Hauteur 21
7. Continuants, Opérateur de Ruelle Mayer 23
7.1. Continuants 23
7.2. Opérateur de Ruelle Mayer 24
8. Trie en fraction continue : comportement asymptotique 26
8.1. Longueur de cheminement externe et nombre de noeuds internes 26
8.2. Hauteur 32
8.3. Grandeurs fondamentales de l'analyse 33
9. Conclusion 34

Date: 1996.

Références 35

Introduction

Le «tri» par comparaisons de n données a une complexité, en nombre de comparaisons, en $O(n \log n)$. On attribue usuellement à chaque comparaison un coût unité, ce qui ne se justifie que dans le cas où les clés sont de nature simple, distribuées de manière relativement uniforme et indépendantes les unes des autres.
Lorsque ces hypothèses ne sont plus vérifiées, et lorsqu'on veut réaliser des comparaisons qui ne soient pas entachées d'erreur, il n' est plus légitime d'attribuer à chaque comparaison un coût unitaire.
Le problème se pose déjà de manière naturelle lors de la comparaison exacte de 2 rationnels, ou, ce qui est équivalent dans l'évaluation exacte du signe d'un déterminant 2×2. Cette évaluation est fondamentale en géométrie algorithmique plane, et doit être effectuée de manière exacte et en simple précision. L'idée est alors de développer les deux rationnels dans un système de numération convenable. Le système binaire est le premier qui vient à l'esprit, mais il s'avère que la numération en fraction continue est au moins aussi efficace.
Cela suggère l'étude d'arbres digitaux,appelés *tries*, fondés non plus sur la numération binaire, mais sur la numération en fraction continue (ce qui constitue en quelque sorte une généralisation de la comparaison de 2 rationnels).
L'analyse en moyenne des trois principaux paramètres des ces tries, à savoir nombre de noeuds, longueur de cheminement externe et hauteur, laisse à penser que le trie en fraction continue est une structure de donnée qui se révèle au moins aussi efficace que son homologue binaire.
On va commencer par établir un cadre précis et définir ce qu'est l'analyse en moyenne d'un algorithme. Les algorithmes de développement en base 2 et en fraction continue seront ensuite présentés et permettront d'introduire la notion très importante d'intervalle fondamental. Il faut donner une définition des tries, qui sera plutôt orientée vers la représentation des nombres, alors que, habituellement, ils servent surtout dans les applications liés au stockage de mots (dictionnaires dans les traitements de texte par exemple). Après avoir défini les paramètres principaux, on effectuera l'analyse en moyenne des tries de nombres. La forme des expressions trouvées suggére alors alors clairement l'utilisation de la transformée de Mellin pour en extraire le comportement asymptotique. Le cas en fraction continue, quant à lui, en plus de la transformée de Mellin, nécessite l'introduction de l'opérateur de Ruelle Mayer pour mener finalement l'étude à bien.

1. Analyse de complexité

L'analyse d'un algorithme consiste en la caractérisation des ressources de calcul nécessaires (le plus souvent temps et espace) à l'exécution de l'algorithme. La taille d'une donnée, définie sur un ensemble de données I, est en général reliée assez "naturellement" à la place mémoire utilisée. Ainsi la taille d'un tableau sera souvent le nombre d'éléments du tableau. Pour un entier m, on aura recours au nombre de bits nécessaires à sa représentation. Une fois la taille

définie, on regroupe les données par taille, et on considère l'ensemble des données de taille n

$$I_n = \{x \in I \mid |x| = n\},$$

où $|x|$ désigne la taille de x. A un algorithme $\mathcal{A}$ fonctionnant sur cet ensemble de données, on associe alors un paramètre μ défini sur I (en général à valeurs entières) soit à l'exécution même de cet algorithme sur la donnée x (place mémoire, nombre d'opérations fondamentales comme des comparaisons, des affectations...), soit à la configuration de sortie produite (comme le degré du PGCD dans le calcul du PGCD de 2 polynômes de $\mathbb{Z}[X]$ par l'algorithme d'Euclide). Une analyse de complexité cherche à caractériser l'évolution d'un paramètre en fonction de la taille n de la donnée. Si μ_n est la restriction de μ à I_n, on peut définir la complexité dans le meilleur des cas et dans le pire des cas, M_n et P_n par

$$M_n = \inf\{\mu_n(x) \mid x \in I_n\}, \quad P_n = \sup\{\mu_n(x) \mid x \in I_n\}.$$

Ces deux notions sont intéressantes puisqu'elles donnent un encadrement des paramètres μ_n et permettent de "prédire" si un algorithme terminera et à partir de quand on peut espérer le voir terminer. L'analyse de complexité en moyenne permet de préciser encore les choses. Cela nécessite de définir un modèle probabiliste sur I_n précisant la répartition des données. La variable μ_n devient alors une variable aléatoire. La complexité en moyenne est alors l'espérance de la variable μ_n :

$$\overline{\mu_n} = \mathrm{E}[\mu_n] = \sum_k k \Pr[\mu_n(X) = k],$$

Le plus souvent, les modèles probabilistes reposent sur le modèle uniforme sur I_n. Un exemple classique de modèle probabiliste est celui des tris à base de comparaisons qui considère n nombres réels tirés de façon indépendante selon une loi de probabilité continue sur $[0, 1]$. Cela est équivalent à choisir une permutation de manière uniforme dans $\{1, \ldots, n\}$ (en raisonnant directement sur l'ordre). Ces diverses notions sont bien distinctes. Un algorithme peut être très peu efficace sur certaines configurations dans le pire des cas et pratiquement linéaire en moyenne, ce qui en fait un algorithme efficace la plupart du temps (donc utilisable en pratique, quitte à laisser de coté les configurations gênantes).

2. Représentation des nombres

Il y a diverses façons de représenter les nombres réels de l'intervalle $\mathcal{D} = [0, 1[$ (c'est à dire à partie entière nulle). On se limitera ici à deux représentations : la représentation usuelle en base 2 et la représentation fondée sur le développement en fraction continue.

2.1. Système de numération en base 2.

L'opérateur de décalage T associé au développement en base 2 est défini pour un réel x de $]0, 1[$ par

$$T(x) = \{2x\} = 2x - \lfloor 2x \rfloor,$$

où $\{u\}$ et $\lfloor u \rfloor$ désignent respectivement la *partie fractionnaire* et la *partie entière* de u. On pose alors $m(x) := \lfloor 2x \rfloor$, c'est le premier chiffre du développement de x en base 2. A l'aide de cet opérateur T, on construit le développement du réel $x \in \mathcal{D} = [0, 1[$ en base 2 par l'algorithme suivant :

Algorithme Développement-Base2(x)

$k := 1$;
Tant que $x \neq 0$ faire
$m_k := m(x)$;
$x := T(x)$;
$k := k + 1$;

Remarque – Le nombre 0 est donc représenté par une liste vide.

Appliqué à un rationnel, ce calcul du développement en base 2 présente des inconvénients bien connus. En particulier, le développement peut être infini (c'est le cas dès que x n'est pas un nombre rationnel dyadique). De plus, il convient de remarquer que tout nombre rationnel dyadique a un développement propre fini et un développement impropre infini (finissant par une suite de 1) ; $(0, 1, 1, 1, \ldots)$ et (1) représentent le même nombre $\frac{1}{2}$. L'algorithme présenté ici nous donne le développement propre. Afin de simplifier le traitement et de rentrer dans le cadre des tries tel qu'il sera défini dans une prochaine partie, on préfère considérer que tout nombre réel $x \in [0, 1[$ posséde un développement infini. L'algorithme du calcul du développement en base 2 doit être modifié pour fournir autant de termes du développement que désiré. Ceci est réalisé en étendant l'opérateur de décalage en 0 en posant $T(0) = 0$ et en supprimant la condition d'arrêt. Si on considère le nombre

$$x = \sum_{j=1}^{n} m_j 2^{-j},$$

il est naturel de compléter l'écriture du nombre par des 0 pour avoir un développement infini, c'est à dire de poser

$$x = \sum_{j=1}^{\infty} {m'}_j 2^{-j},$$

avec $m'_j = m_j$ si $1 \leq j \leq n$ et $m'_j = 0$ sinon. On peut produire dès lors autant de chiffres que voulu. En résumé, l'algorithme modifié calcule une suite d'itérés

$$x_0, \quad x_1 = T(x_0), \quad x_2 = T(x_1), \quad \ldots \quad , \quad x_n = T(x_{n-1}), \ \ldots \ ,$$

et l'on obtient l'expression en base 2 de $x = x_0$

$$x_0 = \sum_{j=1}^{\infty} m_j 2^{-j}, \qquad m_j = m\left(T^{j-1}(x)\right).$$

Inversement on peut aussi, connaissant le développement binaire d'un nombre jusqu'au "chiffre" k et l'itéré x_{k+1}, reconstruire la suite des itérés jusqu'à x_0. En effet, on remarque que les branches inverses de T sont les 2 applications

$$h_m(x) = \frac{x + m}{2} \qquad (m \in \{0, 1\}).$$

La relation

$$x_0 = h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}(x_k),$$

définit une application affine h associée à un k-uplet $(m_1, \ldots, m_k)$. Le nombre $x = x_0$ est donc représenté par la suite infinie des termes de son développement $(m_1, m_2, m_3, \ldots)$ et l'on a le diagramme suivant :

$$x_0 \underset{h_{m_1}}{\overset{T}{\longrightarrow\atop\longleftarrow}} x_1 \underset{h_{m_2}}{\overset{T}{\longrightarrow\atop\longleftarrow}} x_2 \underset{h_{m_3}}{\overset{T}{\longrightarrow\atop\longleftarrow}} \cdots .$$

Il est naturel de considérer l'ensemble des nombres dont le développement binaire commence par $m_1, \ldots, m_k$. C'est un intervalle, appelé *intervalle fondamental*, égal par définition à

$$\mathcal{D}_{m_1,m_2,\ldots,m_k} = h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}(\mathcal{D}),$$

où $\mathcal{D} = [0, 1[$. Le paramètre k est appelé *rang* de l'intervalle fondamental. On note que $[0, 1[$ est la réunion disjointes des intervalles fondamentaux de niveau k, pour chaque $k \geq 0$. La relation

$$h(x) = h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}(x) = \frac{x}{2^k} + \sum_{j=1}^{j=k} m_j 2^{-j}.$$

montre qu'un intervalle fondamental de niveau k est de longueur 2^{-k}.

2.2. **développement en fraction continue.** Le développement en fraction continue est défini dans l'optique du paragraphe précédent en introduisant un autre opérateur de décalage $U :]0, 1[\to]0, 1[$ défini par :

$$U(x) = \left\{ \frac{1}{x} \right\}.$$

On pose alors $m(x) = \lfloor \frac{1}{x} \rfloor$. C'est le premier quotient partiel, qui fournit aussi le premier chiffre du développement de x en fraction continue. A l'aide de cet opérateur U, on construit le développement du réel x en fraction continue par l'algorithme suivant :

Algorithme Développement-FC(x)

$k := 1$;

Tant que $x \neq 0$ faire

$m_k := m(x)$;

$x := U(x)$;

$k := k + 1$;

Remarquons que le développement d'un rationnel est toujours fini, ce qui est d'ailleurs caractéristique par rapport au développement en n'importe quelle base. Puisqu'on veut que tout nombre de $[0, 1[$ ait un développement infini, l'algorithme du calcul du développement en fraction continue doit être modifié pour fournir autant de termes du développement que désiré. A cette fin, on prolonge T en 0 en posant $T(0) = 0$, ce qui nous permet de produire une suite infinie d'itérés. Ensuite, par souci d'homogénéité dans la suite, le chiffre m_j sorti lorsque $x_{j-1} = 0$ est le symbole ∞ (c'est à dire une constante plus grande que toutes les

autres) c'est à dire que $m(0) := \infty$. Un important cas particulier est 0 dont la représentation est $(\infty, \infty, \ldots)$. On peut à nouveau construire la suite des itérés

$$x_0, \quad x_1 = U(x_0), \quad x_2 = U(x_1), \quad \ldots \quad , \quad x_n = U(x_{n-1}), \ \ldots \ ,$$

Les termes $(m_1, m_2, \ldots)$ du développement en fraction continue de $x = x_0$ sont reliés à la suite des itérés grâce à la formule

$$m_j = m(U^{j-1}(x)) = \lfloor 1/U^{j-1}(x) \rfloor.$$

On a la relation $x_j = h_{m_j}(x_{j+1})$, où les branches inverses de U sont

$$h_m(x) = \frac{1}{m + x} \qquad (m \geq 1).$$

Introduire le symbole ∞ permet de rester cohérent. On retrouve les applications réciproques h_m permettant de reconstruire la suite des itérés en "remontant". Il suffit en effet de remarquer, au moins formellement, que $h_\infty(x) = 0$ pour $x \in [0, 1[$. Ainsi $\frac{1}{2}$ s'écrit $(2, \infty, \infty, \ldots)$ et on peut dessiner le *diagramme des itérés*

$$x_0 = \frac{1}{2} \quad \underset{h_2}{\overset{U}{\longrightarrow\atop\longleftarrow}} \quad x_1 = 0 \quad \underset{h_\infty}{\overset{U}{\longrightarrow\atop\longleftarrow}} \quad x_2 = 0 \quad \underset{h_\infty}{\overset{U}{\longrightarrow\atop\longleftarrow}} \quad \cdots .$$

La relation

$$x_0 = h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}(x_k)$$

définit une homographie h de hauteur k, associée à un k-uplet $(m_1, m_2, \ldots, m_k)$ d'entiers $m_i \geq 1$,

$$h(x) = \cfrac{1}{m_1 + \cfrac{1}{m_2 + \cfrac{1}{\ddots \cfrac{}{m_k + x}}}}.$$

L'ensemble des nombres réels dont le développement en fraction continue commence par $m_1, \ldots, m_k$ constitue également un intervalle fondamental, image de $\mathcal{D} = [0, 1[$ par

$$h = h_{m_1} \circ h_{m_2} \circ \cdots \circ h_{m_k}.$$

La réunion des intervalles fondamentaux de rang k est, selon la parité de k, $[0, 1[$ ou $]1, 0]$ (puisque l'homographie h correspondante est croissante ou décroissante selon la parité de k).

3. Trie

Le terme "trie" provient de la contraction de 2 mots anglais : "tree" et "retrieval". C'est un arbre planaire (dont les fils sont ordonnés), qui est utilisé dans la recherche lexicographique.

3.1. **Définitions.**

– Soit $A \subseteq \mathbb{N}$ un ensemble d'éléments appelés *chiffres* et $A^{\mathbb{N}}$ l'ensemble des suites infinies de chiffres de A

$$A^{\mathbb{N}} = \{\alpha = (a_1, a_2, \ldots) |\ \forall j \geq 1,\ a_j \in A\}.$$

– On définit les fonctions *tête* **car** et *queue* **cdr** (ainsi nommées en référence au langage LISP)

$$\begin{array}{rrcl} \mathbf{car}: & A^\infty & \to & A \\ & \alpha = (a_1, a_2, \ldots) & \mapsto & \mathbf{car}(\alpha) = a_1 \end{array}$$

.

$$\begin{array}{rrcl} \mathbf{cdr}: & A^\infty & \to & A^\infty \\ & \alpha = (a_1, a_2, \ldots) & \mapsto & \mathbf{cdr}(\alpha) = (a_2, a_3, \ldots) \end{array}$$

Remarque – L'application **cdr** définit en fait un une opération de décalage.

– Soit X une multipartie (avec répétitions possibles) finie d'éléments de $A^{\mathbb{N}}$. On construit le trie associé à X, noté $\mathcal{T}(X)$ grâce aux règles récursives suivantes :
  – si $|X| = 0$, alors $\mathcal{T}(X)$ est *vide*.
  – si $X = \{\alpha\}$, alors $\mathcal{T}(X)$ est une *feuille* étiquetée α.
  – si $|X| \geq 2$, en ordonnant les éléments de

$$\mathbf{car}(X) = \{\mathbf{car}(\alpha) \mid \alpha \in X\},$$

  on peut écrire $\mathbf{car}(X) = \{t_1, \ldots, t_r\}$ avec $t_1 < \ldots < t_r$. Le trie $\mathcal{T}(X)$ est un *noeud interne*, noté $\bullet$, ayant pour descendance le r-uplet ordonné $(\mathcal{T}(X_1), \ldots, \mathcal{T}(X_r))$, avec

$$X_i = \{\mathbf{cdr}(\alpha) \mid \mathbf{car}(\alpha) = t_i, \alpha \in X\}.$$

Au vu de cette définition, plusieurs remarques peuvent être faites. Tout d'abord, si les éléments de X ne sont pas tous distincts, le trie est infini. Ensuite, la fonction essentielle d'un trie est de séparer les éléments de X. Dès qu'un élément de X est isolé, il constitue une feuille du trie. Ainsi le nombre de feuilles du trie est égal au nombre d'éléments du multiensemble X (si le trie est fini). Il est également intéressant de noter que le degré du trie (c'est à dire le nombre maximal de branchements d'un noeud du trie) est au plus égal à $\min(card(A), n)$, où A est l'ensemble des chiffres et n le cardinal de l'ensemble X. Si $A = \{0, 1\}$ (cas binaire), le trie résultant sera binaire.

Exemple –*cas binaire.* Considérons 5 chaines de bits A, B, C, D, E dont les préfixes de longueur 7 sont $0000100, 0001110, 0010100, 1001001, 1001101$. Le trie correspondant est dessiné sur la figure 1.

3.2. **Exemples d'applications.** Les tries sont souvent utilisés dans des applications qui travaillent sur un *alphabet* fini de lettres A. On se ramène alors à la définition en codant les lettres par des entiers. Les tries sont à la base de nombreuses applications. C'est une bonne structure de donnée pour représenter *un ensemble de mots sur un alphabet*, car il est facile de supprimer, rechercher, ou encore ajouter un mot (figure 2). Ces aspects ne sont cependant pas abordés dans le cadre de ce mémoire. Dans la pratique, les mots sont évidemment finis. Mais le trie peut tout de même être construit grâce aux règles récursives de la définition si aucun mot n'est le préfixe d'un autre (par exemple, l'adjectif "uni" est un préfixe du nom

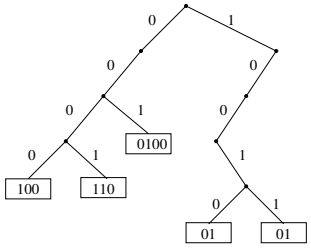

FIG. 1. exemple de trie binaire pour 5 chaines de bits $0000100, 0001110, 0010100, 1001001, 1001101$.

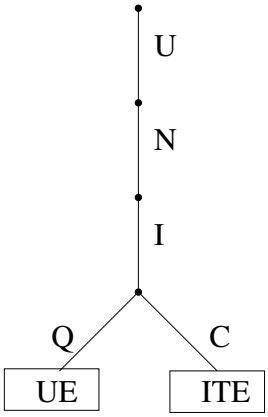

FIG. 2. exemple de trie représentant {UNIQUE, UNICITE}.

commun "unicité"). Dans le cas contraire, il faut rajouter une information dans le trie (en introduisant par exemple un nouveau caractère dit de "terminaison"). Le fait de ne considérer que des mots infinis nous garantit qu'aucun mot n'est le préfixe d'un autre (à moins de lui être égal), ce qui permet de simplifier l'analyse en moyenne sans prete de généralité, puisque sous tout modèle probabiliste réaliste, la probabilité que 2 mots infinis soient égaux est nulle. Pour une implémentation réelle du trie, il faut également rajouter des informations sur les liens, de façon, lors du parcours d'une branche, à pouvoir reconstituer le mot.

3.2.1. *Recherche.* Le principe est le suivant : pour tester si un mot x est présent parmi un ensemble de mots $\mathcal{M}$, on commence par construire le trie associé à l'ensemble $\mathcal{M}$. Ensuite, on parcourt le mot x lettre par lettre en empruntant la branche correspondante du trie jusqu'à un blocage éventuel (c'est à dire jusqu'à ce que la lettre sur le flot d'entrée ne corresponde à aucun fils du noeud courant). S'il y a blocage sur un noeud interne, le mot n'est pas présent. Si on arrive à une feuille, il reste à comparer la fin du mot et la chaîne de symboles stockée sur cette feuille.
La recherche de motifs (ou encore "pattern matching") est un outil de base de l'informatique. Le but à atteindre est simple : savoir si un motif a une occurence dans un texte. La structure de trie est pleinement adaptée à ce type de problème. On peut grandement améliorer la technique de recherche du paragraphe précédent pour s'adapter à une situation donnée. Dans le cas où le texte est fixé (typiquement un *dictionnaire*) et où de nombreux motifs doivent

être recherchés, on peut construire *une fois pour toutes* un trie codant le dictionnaire, ce qui permet de réduire de façon sensible les temps de recherche. Considérons un texte de longueur N : $m_1 m_2 \dots m_N$ sur un alphabet A. On construit le trie (appelé "suffix trie") associé à l'ensemble des suffixes du texte

$$\{m_1 m_2 \dots m_N, m_2 m_3 \dots m_N, \dots, m_{N-1} m_N, m_N\}.$$

La recherche d'un motif (variable) dans le texte (fixe) est à peu près la même que pour la recherche de base, on descend dans le trie en empruntant la branche correspondant au motif. Si on bloque sur un noeud interne, le motif n'est pas présent. Si on finit sur un noeud interne du trie, le motif est présent. Enfin si on termine sur une feuille sans avoir "épuisé" le motif, il faut vérifier la concordance entre la fin du motif et le début de la chaîne, étiquette de la feuille. Cette méthode est très efficace puisqu'un parcours du motif suffit pour savoir s'il est présent ou non. Un autre exemple de recherche de motifs concerne la configuration inverse de celle qu'on vient de voir, à savoir le cas où plusieurs motifs (fixés) doivent être recherchés dans un texte (variable). On commence par construire un trie sur l'ensemble des motifs. Pour chaque position i dans le texte, on parcourt le trie à partir de la racine. Si on aboutit à une feuille, un des motifs est présent sinon on recommence à la position $i+1$ dans le texte. Cette méthode peut être affinée en introduisant une fonction de retour plus adaptée (afin de ne pas revenir en cas d'echec à la position $i+1$).

3.2.2. *Codage,décodage.* Considérons un codage binaire sans préfixe $\mathcal{C}$ de N symboles. L'ensemble des N chaînes binaires forme un ensemble de mots dont aucun n'est le préfixe d'un autre, et permet donc de construire un trie binaire, à partir duquel on peut obtenir un code $\mathcal{C}'$ au moins aussi "court" que $\mathcal{C}$. En effet, on fait correspondre à chaque feuille du trie un des symboles. Il suffit alors de coder chaque symbole par la branche menant à la feuille le représentant. Le code $\mathcal{C}'$ est plus court en ce sens que la longueur d'une branche, et donc la chaîne binaire résultante, est inférieure ou égale à la longueur du mot binaire initial. On pourrait d'ailleurs continuer dans ce sens et en faire un Patricia trie, arbre localement complet). La technique de *décodage* est très simple et n'utilise en fait que la structure d'arbre binaire sous-jacente. On part de la chaîne de bits à décoder et du trie à N feuilles étiquetées par les N symboles. On descend dans le trie en fonction des bits rencontrés (à gauche si 0, à droite si 1). Lorsqu'une feuille est atteinte, le symbole correspondant est produit. On revient ensuite à la racine et on recommence avec le reste de la chaîne de bits. **exemple.** Si le codage des symboles A, C, D, E, G, O sont respectivement $0000, 0001, 01, 10, 110, 111$. Le trie est représenté sur la figure 3 et on code le mot "DECODAGE" par 01100001111010000011010. Si les symboles les plus souvent utilisés correspondent à des feuilles de faible profondeur, le codage sera d'autant plus efficace. La technique de codage de Fano-Shannon (dont les performances se rapprochent du codage de Huffman) utilise une approche descendante connaissant les fréquences des symboles d'un texte qui rappelle le processus de construction d'un trie.

3.3. **Trie et représentation des nombres.** La construction d'un trie à partir d'un ensemble X de chaînes de symboles permet grâce à une lecture des feuilles de gauche à droite de trier *lexicographiquement* cet ensemble. En effet, par définition , on commence par partager l'ensemble X en r sous-ensembles ordonnés selon r lettres. Chaque sous-ensemble est ensuite traité récursivement. Considérons n nombres de l'intervalle $[0, 1[$. Un trie peut être

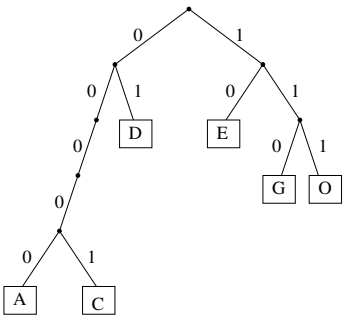

FIG. 3. Trie associé aux chaines $0000, 0001, 01, 10, 110, 111$ codant respectivement A, C, D, E, G, O.

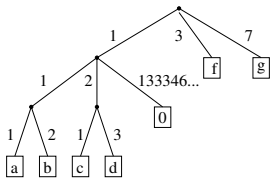

FIG. 4. Trie en fraction continue associé à $\{a, b, c, d, e, f, g\}$.

construit d'après leurs représentations (en base 2 ou en fraction continue). C'est ce genre de structure qui va être analysée par la suite, en choisissant de tirer les nombres de manière indépendante selon une loi uniforme sur $[0, 1[$. En base 2, construire le trie et lire les feuilles de gauche à droite équivaut exactement à l'algorithme de tri "radix sort", puisque tri lexicographique des représentations en base 2 (ou n'importe quelle base) et tri des nombres eux-mêmes coïncident. Dans le cas des fractions continues, il est aisé de faire un parcours adéquat des feuilles du trie pour obtenir la liste de nombres triés. L'opérateur de décalage étant décroissant, cela oblige à parcourir les noeuds d'un niveau du trie dans un sens ou dans l'autre selon la parité du niveau si on veut un parcours ordonné d'un niveau. **Exemples de tries.** En prenant un exemple tiré de [6], avec les nombres

$$\begin{aligned}
a &= \phi - 1 = (\mathbf{1}, 1, 1, 1, 1, 1, \ldots) \\
b &= \gamma = (\mathbf{1}, 1, 2, 1, 2, 1, \ldots) \\
c &= \exp(1) - 2 = (\mathbf{1}, 2, 1, 1, 4, 1, \ldots) \\
d &= \log 2 = (\mathbf{1}, 2, 3, 1, 6, 3, \ldots) \\
e &= \exp(\pi\sqrt{263}) = (\mathbf{1}, 1333462407511, 1, 8, 1, \ldots) \\
f &= 2^{1/3} - 1 = (\mathbf{3}, 1, 5, 1, 1, 4, \ldots) \\
g &= \phi - 1 = (\mathbf{7}, 15, 1, 292, 1, 1, \ldots),
\end{aligned}$$

on obtient le trie en fraction continue de la figure 4.

3.4. **Paramètres.** Il existe plusieurs paramètres naturels sur les tries qui permettent d'analyser le comportement d'algorithmes basés sur ce type de structure.

– La *profondeur* (ou niveau) P d'un noeud du trie se définit récursivement par

  - $P(r) = 0$ si le noeud r est la racine
  - $P(n) = P(père(n)) + 1$ sinon, $père(n)$ étant le noeud père du noeud n.

  La profondeur d'un noeud est aussi le nombre de liens le séparant de la racine.
- La *hauteur* H du trie est la profondeur maximale atteinte parmi l'ensemble des noeuds du trie. C'est aussi la longueur maximale d'une branche du trie, et donc le nombre de comparaisons maximal pour départager 2 éléments.
- La *taille* T du trie est le nombre de noeuds internes du trie. Ce paramètre est lié à la taille mémoire nécessaire au stockage du trie, le nombre total de noeuds étant égal à la somme de la taille et du nombre n de feuilles.
- La *longueur de cheminement externe* L est la somme des profondeurs des feuilles. Ce paramètre mesure le nombre de comparaisons nécessaires à la construction du trie. En divisant par n, le nombre de feuilles, on obtient la profondeur moyenne d'une feuille.

4. Paramètres des tries de nombre

On construit généralement un trie à partir de n mots. Il y a d'autres manières de générer un trie. On pourrait construire le "suffix trie" d'un texte de taille n par exemple. L'analyse en moyenne n'est bien sûr plus la même. Notre approche est orientée vers la représentation des nombres. On va chercher à analyser dans cette partie les paramètres d'un trie construit d'après les représentations de n nombres. Pour effectuer l'analyse en moyenne, cela nécessite de définir un modèle probabiliste.

4.1. Notations.

On peut prendre des notations communes pour les systèmes de représentation en base 2 et en fraction continue, ce qui permet de donner un cadre commun pour l'analyse en moyenne. On définit donc

$$H_\ell = \begin{cases} \{id_{[0,1[}\} & \text{si } \ell = 0 \\ \{h = h_{m_1} \circ \ldots \circ h_{m_\ell} \mid m_i \geq 1\} & \text{si } \ell > 0 \end{cases}$$

et enfin

$$H = \bigcup_{\ell \in \mathbb{N}} H_\ell,$$

où les H_ℓ sont des ensembles disjoints. Les applications h_m sont définies par

$$h_m(x) = \begin{cases} \frac{m+x}{2} & \text{pour la représentation en base 2,} \\ \frac{1}{m+x} & \text{pour la représentation en fraction continue.} \end{cases}$$

On appelle *hauteur* d'une application $h \in H$, notée $|h|$, l'indice ℓ tel que $h \in H_\ell$ (comme les ensembles sont disjoints, ℓ est déterminé de façon unique). Dans la suite, on confond souvent $h = h_{m_1} \circ \ldots \circ h_{m_\ell} \in H_\ell$ avec le noeud du trie associé à la branche $(m_1, \ldots, m_\ell)$. La hauteur de h est alors la profondeur de ce noeud dans le trie. On note également l'intervalle fondamental de rang ℓ associé à h

$$\mathcal{D}_h = h(\mathcal{D}).$$

La longueur de l'intervalle fondamental $\mathcal{D}_h$ est noté u_h.

4.2. **Modèle probabiliste.** Le modèle, que l'on va utiliser ici, et qui permet de simplifier grandement les calculs, se base sur une approche légèrement différente de celle qui paraît a priori "naturelle". Au lieu de considérer n (n étant fixé) variables indépendantes sur $]0,1[$ $X_1, \ldots, X_n$, on considère une variable de Poisson N de paramètre n, puis N variables $X_1, \ldots, X_N$ indépendantes uniformes sur $]0,1[$. Si N est de Poisson de paramètre n, alors les 2 modèles, appelés respectivement modèle de Bernouilli et de Poisson, sont proches en moyenne.

4.2.1. *Variable aléatoire de Poisson.* Une variable aléatoire N de Poisson de paramètre λ vérifie :

$$Pr[N = k] = e^{-\lambda} \frac{\lambda^k}{k!}$$

On peut aisément calculer la moyenne et la variance :

$$\mathrm{E}[N] = \mathrm{V}ar[N] = \lambda.$$

4.2.2. *Calcul préliminaire.* Il s'agit d'évaluer la probabilité d'un évenement qui interviendra souvent par la suite, et qui met en avant une propriété essentielle du modèle de Poisson : Si l'on se place dans le modèle de Poisson de taux λ, le nombre déléments tombant dans un intervalle de longueur x est distribué selon une loi de Poisson de paramètre λx, et ce de manière indépendante de ce qui se passe dans tout autre autre sous intervalle disjoint. Soient r intervalles disjoints $U_1, \ldots, U_r$ de $[0,1]$ de longueurs respectives $x_1, \ldots, x_r$. Considérons l'événement suivant $A^{\lambda}_{n_1,\ldots,n_r}(x_1, \ldots, x_r)$: "Parmi $X_1, \ldots, X_N$ variables indépendantes uniformes sur $[0,1]$ (N étant une variable de Poisson de paramètre λ), pour tout $i \in \{1, \ldots, r\}$, il existe exactement n_i variables aléatoires dans l'intervalle U_i". Notons $y = 1 - \sum_{i=1}^{r} x_i$, c'est à dire la longueur du complémentaire de la réunion des U_i, et $n = \sum_{i=1}^{r} n_i$. On calcule

$$\begin{aligned}
Pr[A^{\lambda}_{n_1,\ldots,n_r}(x_1, \ldots, x_r)] &= \sum_{k=n_1+\ldots+n_r}^{\infty} \Pr[N = k] \frac{k!}{n_1! \ldots n_r!(k-n)!} \, x_1^{n_1} \ldots x_r^{n_r} y^{k-n} \\
&= \sum_{k=n}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \frac{k!}{n_1! \ldots n_r!(k-n)!} \, x_1^{n_1} \ldots x_r^{n_r} y^{k-n} \\
&= e^{-\lambda} \frac{x_1^{n_1}}{n_1!} \ldots \frac{x_r^{n_r}}{n_r!} \sum_{k=n}^{\infty} \frac{\lambda^k}{k!} \frac{k!}{(k-n)!} y^{k-n} \\
&= e^{-\lambda} \frac{x_1^{n_1}}{n_1!} \ldots \frac{x_r^{n_r}}{n_r!} \sum_{k=0}^{\infty} \frac{\lambda^{k+n}}{k!} y^k \\
&= e^{-\lambda} \frac{x_1^{n_1}}{n_1!} \ldots \frac{x_r^{n_r}}{n_r!} \lambda^n e^{\lambda y} \\
&= e^{-\lambda(x_1+\ldots+x_r)} \frac{(\lambda x_1)^{n_1}}{n_1!} \ldots \frac{(\lambda x_r)^{n_r}}{n_r!} \\
&= \prod_{i=0}^{r} \Pr[A^{\lambda}_{n_i}(x_i)]
\end{aligned}$$

4.3. **Longueur de cheminement externe.** On se place tout d'abord à $N = k$ fixé (modèle de Bernouilli). Evaluons la probabilité de l'évenement : $[D_{i,k} > \ell]$ = "la feuille contenant X_i est à une profondeur $D_{i,k} > \ell$". Cela revient à calculer la probabilité qu'il existe un intervalle fondamental $\mathcal{D}_h$ avec $|h| > \ell$ tel que X_i et une variable X_j parmi les $k-1$ autres variables restantes (c'est à dire $j \neq i$) soient dans $\mathcal{I}_h$. On calcule facilement la probabilité de cet évenement

$$\Pr[D_{i,k} > \ell] = \sum_{|h|=\ell} u_h \sum_{m=1}^{k-1} C_{k-1}^{\ell} {u_h}^{\ell} (1-u_h)^{k-1-\ell} = \sum_{h \in H_\ell} u_h (1-(1-u_h)^{k-1}),$$

d'où l'espérance de $D_i^{(k)}$ (qui est indépendante de i).

$$\mathrm{E}[D_{i,k}] = \sum_{\ell \geq 0} \Pr[D_{i,k}] = \sum_{h \in H} u_h (1-(1-u_h)^{k-1}).$$

La longueur de cheminement externe est $L_k = \sum_{i=1}^{k} D_{i,k}$. On obtient alors l'espérance

$$\mathrm{E}[L_k] = \sum_{i=1}^{k} \mathrm{E}[D_{i,k}] = k \sum_{h \in H} u_h (1-(1-u_h)^{k-1}).$$

Dans le modèle de Poisson, N étant une variable de Poisson de paramètre n, on obtient

$$\begin{aligned}
\mathrm{E}[L_N] &= \sum_{k=0}^{\infty} \Pr[N=k]\ \mathrm{E}[L_k] \\
&= \sum_{k=1}^{\infty} e^{-n} \frac{n^k}{k!} k \sum_{h \in H} u_h (1-(1-u_h)^{k-1}) \\
&= \sum_{h \in H} e^{-n} n u_h \sum_{k=1}^{infty} \frac{n^{k-1}}{(k-1)!} (1-(1-u_h)^{k-1}) \\
&= \sum_{h \in H} u_h n e^{-n} \sum_{k=0}^{\infty} (\frac{n^k}{k!} - \frac{[n(1-u_h)]}{k!}) \\
&= \sum_{h \in H} u_h n e^{-n} (e^n - e^{n(1-u_h)}) \\
&= \sum_{h \in H} n u_h (1-e^{-n u_h})
\end{aligned}$$

4.4. **Nombre de noeuds internes.** On se place à nouveau à $N = k$ fixé. On considère l'événement $B_{h,\ell}$: "Le noeud associé à h de niveau ℓ est interne". On peut exprimer cet événement sous la forme : "il existe 2 variables X_i au moins se trouvant dans l'intervalle fondamental $\mathcal{D}_h$ de rang ℓ". Le nombre moyen de noeuds internes de niveau ℓ est égal à

$$\mathrm{E}[T_{k,\ell}] = \sum_{h \in H_\ell} \Pr[A_{h,\ell}] = \sum_{h \in H_\ell} (1-(1-u_h)^k - k u_h (1-u_h)^{k-1}).$$

Le nombre moyen de noeuds internes du trie $\mathrm{E}[T_k]$ se calcule en sommant sur tous les niveaux ℓ

$$\mathrm{E}[T_k] = \sum_{h\in H}(1-(1-u_h)^k - ku_h(1-u_h)^{k-1}).$$

Enfin, en passant au modèle de Poisson, on trouve l'espérance du nombre de noeuds internes exprimée en fonction des intervalles fondamentaux

$$\begin{aligned}
\mathrm{E}[T_N] &= \sum_{k=0}^{\infty}\Pr[N=k]\sum_{h\in H}(1-(1-u_h)^k - ku_h(1-u_h)^{k-1}) \\
&= \sum_{h\in H} e^{-n}\sum_{k=1}^{\infty}\frac{n^k}{k!}(1-(1-u_h)^k - ku_h(1-u_h)^{k-1}) \\
&= \sum_{h\in H} e^{-n}(e^n - e^{1-nu_h} - nu_h e^{1-nu_h}) \\
&= \sum_{h\in H}(1-e^{-nu_h}(1+nu_h))
\end{aligned}$$

4.5. **Hauteur.** On se place cette fois-ci directement dans le modèle de Poisson en considérant N variables $X_1,\ldots,X_N$, N étant une variable aléatoire de Poisson de paramètre λ. On considère l'événement $[H_\lambda \leq \ell]$: "Le trie construit à partir de $X_1,\ldots,X_N$ est de hauteur inférieure ou égale à ℓ". La probabilité de cet événement est celle que tous les intervalles fondamentaux de rang ℓ contiennent au plus une variable X_i. Dans le modèle de Poisson, le nombre de X_i tombant sur un intervalle est indépendant de ce qui se passe à l'extérieur de l'intervalle. On peut donc écrire :

$$\Pr[H_\lambda \leq \ell] = \prod_{h\in H_\ell}(e^{-nu_h} + nu_h e^{-nu_h})$$

On peut regrouper les derniers résultats dans la proposition suivante

Proposition 1. *Dans le modèle de Poisson,les espérances de la longueur de cheminement externe, du nombre de noeuds internes et de la hauteur du trie construit à partir de N variables indépendantes $X_1,\ldots,X_N$ (N étant une variable de Poisson de paramètre n), s'écrivent respectivement*

$$\begin{aligned}
\mathrm{E}[L_N] &= \sum_{h\in H} nu_h(1-e^{-nu_h}), \\
\mathrm{E}[T_N] &= \sum_{h\in H}(1-e^{-nu_h}(1+nu_h)), \\
\mathrm{E}[H_N] &= \sum_{\ell\geq 0}(1-\prod_{h\in H_\ell} e^{-nu_h}(1+nu_h))
\end{aligned}$$

On peut tout de suite remarquer que $\mathrm{E}[L_n]$ et $\mathrm{E}[T_n]$ sont de la forme

$$\sum_{h\in H}\lambda_k f(\mu_k n).$$

Une telle fonction est appelée somme harmonique (et les λ_k et μ_k sont appelées amplitudes et fréquences). L'outil de choix pour accéder au comportement asymptotique d'une telle fonction est la transformée de Mellin.

5. Transformée de Mellin

5.1. **Définitions.** Dans la suite, la notation $< \alpha, \beta >$ désignera la bande ouverte du plan complexe $\{s \in \mathbb{C} \mid \Re(s) \in]\alpha, \beta[\ \}$.

Définition 1 (Transformée de Mellin)**.** *Soit $f :]0, \infty[\to \mathbb{R}$ une fonction localement sommable sur $]0, \infty[$, la transformée de Mellin est définie par*

$$\mathcal{M}[f(x); s] = f^\star(s) = \int_0^{+\infty} f(x) x^{s-1} dx$$

La plus grande bande $< \alpha, \beta >$ sur laquelle l'intégrale converge est appelée *bande fondamentale* de $f^\star$.

Exemples : Les seules fonctions dont il sera fait usage dans la suite sont les suivantes (les bandes fondamentales sont précisées pour chaque fonction)

$$e^{-x} \to \Gamma(s) \qquad s \in < 0, +\infty >$$
$$e^{-x} - 1 \to \Gamma(s) \qquad s \in < -1, 0 >$$

Remarque – En coupant l'intégrale en 2, $\int_0^\infty = \int_0^1 + \int_1^\infty$, on voit que les conditions

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

garantissent, si $u > v$, que $f^\star(s)$ existe dans la bande $< -u, -v >$. Ainsi apparaît le rapport entre le développement asymptotique de f en 0 (resp. en $+\infty$) et la frontière gauche (resp. droite) de la bande fondamentale de $f^\star$.

5.2. **Propriétés fonctionnelles.** Soit $f :]0, \infty[\to \mathbb{R}$ une fonction localement sommable sur $]0, \infty[$ dont la transformée de Mellin admet $< \alpha, \beta >$ comme bande fondamentale. On peut établir directement les propriétés suivantes :

(i)	$\mathcal{M}[f(\mu x); s] = \mu^{-s} f^\star(s)$	$s \in < \alpha, \beta >,\ \mu > 0$
(ii)	$\mathcal{M}[\lambda f(x); s] = \lambda f^\star(s)$	$s \in < \alpha, \beta >$
(iii)	$\mathcal{M}[\sum_{k \in K} \lambda_k f(\mu_k x); s] = (\sum_{k \in K} \lambda_k \mu_k^{-s}) f^\star(s)$	$s \in < \alpha, \beta >,\ \mu_k > 0,$ K fini
(iv)	$\mathcal{M}[f(\frac{1}{x}); s] = f^\star(-s)$	$s \in < -\beta, -\alpha >$

(i) et (ii) s'obtiennent facilement grâce à la linéarité de l'intégration et un changement de variable. (iii) découle directement de (i) et (ii). On verra plus loin qu'on peut étendre cette formule sous certaines conditions au cas où K est infini, faisant intervenir une série de Dirichlet. Enfin, un changement de variable permet d'obtenir (iv). Cette formule permet de ne chercher que le comportement asymptotique en 0^+, puisque le développement asymptotique de $g(x)$ en $+\infty$ se ramène à celui de $f(x) = g(\frac{1}{x})$ en 0^+.

5.3. **Propriétés asymptotiques.** Il y a une correspondance précise entre le développement asymptotique d'une fonction en 0 (resp. $+\infty$), et les pôles de la transformée de Mellin dans un demi-plan gauche (resp. droit) par rapport à la bande fondamentale. Chaque terme du développement asymptotique de f de la forme $x^c(\log x)^k$ correspond à un pôle d'ordre $k+1$ de sa transformée $f^\star$ en $s=-c$. Mais cette correspondance sera utilisée dans le sens inverse. Pour déterminer le développement asymptotique de $F(x)$, on calcule la transformée de Mellin $F^\star(s)$ et chaque pôle de $F^\star$ donnera un terme du développement asymptotique de F. Soit une fonction Φ méromorphe en $s=s_0$. La fonction Φ se développe en une série de Laurent au voisinage de s_0

$$\Phi(s)=\sum_{k=-\infty}^{+\infty} c_k(s-s_0)^k.$$

La fonction $\Phi(s)$ a un *pôle* d'ordre r si $r>0$ et $c_{-r}\neq 0$, et est *holomorphe* en s_0 si $c_k=0$ pour tout $k<0$. On définit la *partie singulière* de Φ en $s=s_0$ par

$$\sum_{k=-\infty}^{-1} c_k(s-s_0)^k.$$

Définition 2. *Soit Φ méromorphe sur Ω et $\mathcal{S}\subseteq\Omega$ l'ensemble des pôles de Φ dans Ω. La* partie singulière *de Φ sur Ω est la somme formelle des parties singulières de Φ en chacun des points de $\mathcal{S}$.*

notation – Si E est la partie singulière de Φ sur Ω, on utilise la notation

$$\Phi(s)\asymp E \qquad (s\in\Omega).$$

Exemple :

$$\frac{1}{s^2(s+1)}\asymp[\frac{1}{s+1}]+[\frac{1}{s^2}-\frac{1}{s}] \qquad (s\in<-2,2>)$$

La notion de partie singulière d'une fonction méromorphe sur un ensemble $\Omega\subseteq\mathbb{C}$ étant définie, on peut énoncer le théorème suivant :

Théorème 1. *Soit f une fonction continue sur $]0,+\infty[$ dont la transformée de Mellin $f^\star$ admet une bande fondamentale non vide $<\alpha,\beta>$. On suppose que*

(i) *$f^\star(s)$ admet un prolongement méromorphe sur $<\gamma,\beta>$ avec $\gamma<\alpha$, et est analytique sur $\Re(s)=\gamma$.*

(ii) *Il existe un réel $\eta\in]\alpha,\beta[$, un entier $r>1$ et une suite réelle $(T_j)_{j\in\mathbb{N}}$ strictement croissante divergeant vers $+\infty$ tels que*

$$f^\star(s)=O(|s|^{-r}),$$

sur la réunion des segments $\{s\in\mathbb{C}\mid \Re(s)\in[\gamma,\eta]\ ,\Im(s)=T_j\}$, quand $j\to+\infty$.

Si $f^\star(s)$ admet comme partie singulière

$$f^\star(s)\asymp\sum_{(\zeta,k)\in A)} d_{\zeta,k}\frac{1}{(s-\zeta)^k}\ (s\in<\gamma,\eta>),$$

alors le développement asymptotique de $f(x)$ en 0^+ est

$$f(x) = \sum_{(\zeta,k)\in A)} d_{\zeta,k} \frac{(-1)^{k-1}}{(k-1)!} x^{-\zeta} (\log x)^k + O(x^{-\gamma}).$$

Dans la suite, on s'est limité au développement en 0^+ car c'est le seul dont on fera usage.
Fluctuations périodiques. Un pôle de $f^\star$ en un point $\zeta = \sigma + it$ non réel introduit dans le développement asymptotique un terme de la forme

$$x^{-\zeta} = x^{-\sigma} e^{-it \log x}$$

qui contient une composante périodique en $\log x$ de période $\frac{2\pi}{t}$. Ce sont ces phénomènes de fluctuations qui rendent la transformée de Mellin si utile puisqu'ils sont, autrement, difficilement accessibles par d'autres méthodes.

5.4. **Sommes harmoniques.**

Définition 3. *Une somme de la forme*

$$G(x) = \sum_{k\in K} \lambda_k g(\mu_k x) \qquad (K \text{ fini ou infini})$$

est appelée somme harmonique. Les ensembles $\{\lambda_k\}$ et $\{\mu_k\}$ forment respectivement l'ensemble des amplitudes et des fréquences. La fonction $g(x)$ est appelée fonction de base de la somme harmonique.

Dans la suite, on se restreint à l'étude de sommes harmoniques dont les fréquences μ_k tendent vers $+\infty$ quand $k \to +\infty$. Si μ_k tend vers 0, on peut toujours se ramener à ce cas en considérant $G(\frac{1}{x})$ et en changeant de manière adéquate la fonction de base g. La série de Dirichlet associée à la somme harmonique, si les μ_k sont entiers, est

$$\Lambda(s) = \sum_{k\in K} \lambda_k {\mu_k}^{-s}.$$

La transformée de Mellin d'une somme harmonique finie s'écrit

$$G^\star(s) = \Lambda(s) g^\star(s).$$

On peut étendre cette formule du produit aux sommes harmoniques infinies sous les conditions du théorème des sommes harmoniques. Ce théorème permet de garantir la convergence d'une somme harmonique et d'en obtenir son développement asymptotique.

Théorème 2 (Sommes harmoniques)**.** *Soit $G(x)$ une somme harmonique,*

$$G(x) = \sum_{k\in K} \lambda_k g(\mu_k x),$$

telle que :

- *g est continue sur $]0, +\infty[$ et $g^\star$ a pour bande fondamentale $< \alpha, \beta >$;*
- *$\Lambda(s) = \sum_{k\in K} \lambda_k {\mu_k}^{-s}$ la série de Dirichlet associée à $G(x)$ a un demi-plan de convergence simple $\Re(s) > \sigma$.*

Supposons de plus que

(i) *$\sigma < \beta$, c'est à dire que le demi-plan de convergence simple de $\Lambda(s)$ a une intersection non vide avec la bande fondamentale $< \alpha, \beta >$. Notons $\alpha' = \max(\alpha, \sigma)$.*

(ii) *Il existe un nombre réel $\gamma < \alpha$ pour lequel $g^\star$ et Λ admettent un prolongement méromophe sur $< \gamma, \beta >$ et soient analytiques sur $\Re(s) = \gamma$.*

(iii) *Il existe un réel $\eta \in]\alpha', \beta[$ et une suite réelle $(T_j)_{j\in\mathbb{N}}$ strictement croissante divergeant vers $+\infty$ tels que*

$$g^\star(s) = O(|s|^{-k}),\ \textit{pour tout entier } k > 0,$$
$$\Lambda(s) = O(|s|^{r}),\ \textit{pour un entier } r > 0,$$

sur la réunion des segments $\{s \in \mathbb{C} \mid \Re(s) \in [\gamma, \eta]\ , \Im(s) = T_j\}$ quand $j \to +\infty$.

Alors la somme harmonique $G(x)$ converge sur $]0, +\infty[$ et sa transformée $G^\star(s)$ est bien définie sur $< \alpha', \beta >$ et s'écrit $G^\star(s) = \Lambda(s)g^\star(s)$. Si $G^\star(s)$ admet de plus comme partie singulière

$$G^\star(s) \asymp \sum_{(\zeta,k)\in A} d_{\zeta,k} \frac{1}{(s-\zeta)^k} \qquad (s \in < \gamma, \eta >),$$

alors le développement asymptotique de $G(x)$ en 0^+ est

$$G(x) = \sum_{(\zeta,k)\in A} d_{\zeta,k} \frac{(-1)^{k-1}}{(k-1)!} x^{-\zeta} (\log x)^k + O(x^{-\gamma}).$$

6. Trie binaire : comportement asymptotique

On a déjà exprimé l'espérance des paramètres en fonction des intervalles fondamentaux, ou plutôt en fonction des longueurs des intervalles fondamentaux appelés u_h. Dans le cas binaire, on a déja établi qu'un intervalle fondamental de rang k est de longueur $u_h = 2^{-k}$.

6.1. Longueur de cheminement externe.

Soit F la somme harmonique définie par

$$F(x) = \sum_{h\in H} f(u_h x), \text{avec } f(x) = x(1 - e^{-x}).$$

L'espérance $\mathrm{E}[L_N]$ est égale à $F(n)$. On veut déterminer le développement asymptotique de $F(x)$ quand $x \to \infty$. Comme $u_h \to 0$ quand $|h| \to +\infty$ et que, on va plutôt chercher (pour se mettre dans le cadre du théorème) le développement asymptotique $G(x) = F(1/x)$ en 0^+. On a

$$G(x) = \sum_{h\in H} g(\mu_h x), \text{avec } g(x) = f(1/x) \text{ et } \mu_h = {u_h}^{-1}.$$

On calcule la transformée de Mellin de g

$$g^\star(s) = f^\star(-s) = -\Gamma(-s+1) \qquad (s \in < 1, 2 >).$$

La série de Dirichlet associée à $G(x)$, bien définie car les ${u_h}^{-1} = 2^k$ sont entiers, est

$$\Lambda(s) = \sum_h ({u_h}^{-1})^{-s} = \sum_{k=0}^{\infty} 2^k (2^k)^{-s} = \frac{1}{1 - 2^{1-s}}$$

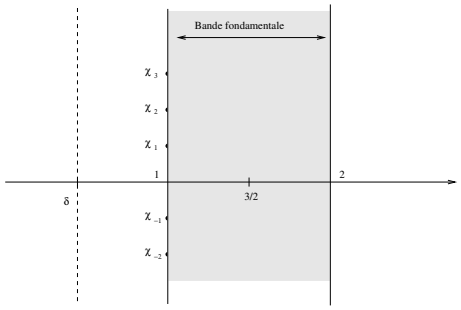

FIG. 5. Situation des pôles et de la bande fondamentale dans le plan complexe.

La série de Dirichlet $\Lambda(s)$ a un demi-plan de convergence simple $\Re(s) > 1$. Les pôles de Λ sont situés sur une droite verticale $\Re(s) = 1$ aux points

$$\chi_k = 1 + \frac{2\mathrm{i}k\pi}{\log 2},$$

pour $k \in \mathbb{Z}$. La figure 5 précise les conditions d'application du théorème (les pôles qui nous intéressent sont représentés par des points). Appliquons le théorème des sommes harmoniques

- $g^\star$ a pour bande fondamentale $< 1, 2 >$;
- Λ a pour demi-plan de convergence simple $\Re(s) > 1$ d'intersection non vide avec $< 1, 2 >$;
- Λ et $g^\star$ admettent un prolongement méromorphe sur $\mathbb{C}$ et sont analytiques sur $\Re(s) = \delta$ pour tout réel $\delta < 1$;
- Soit $\delta < 1$. On considère la suite (T_j) définie par

$$T_j = \frac{(2k+1)\pi}{\log 2}.$$

Alors sur la réunion des segments, défini pour éviter les pôles de Λ, on a

$$\{s \in \mathbb{C} \mid \Re(s) \in [\delta, \frac{3}{2}] \; et \; \Im(s) = T_j\},$$

qui est défini pour éviter les pôles de Λ, on a

$$|\Lambda(\sigma + \mathrm{i}T_j)| = |\frac{1}{1 - 2^{1-\sigma-\mathrm{i}T_j}}| \leq \frac{1}{1 + 2^{1-3/2}} = O(|s|^0).$$

De plus, la version complexe de la formule de Stirling

$$|\Gamma(\sigma + \mathrm{i}t)| \sim \sqrt{2\pi}|t|^{\sigma-1/2}e^{-\pi|t|/2} \qquad (t \to +\infty),$$

permet d'obtenir pour tout $r > 1$,quand $j \to +\infty$, que

$$|g^\star(\sigma + \mathrm{i}T_j)| \;\; = \;\; |\Gamma(1 - \sigma - \mathrm{i}T_j)| = O(|s|^{-r}).$$

Donc le théorème s'applique : $G(x)$ converge sur $]0,+\infty[$. La partie singulière de $G^\star(s) = \Lambda(s)g^\star(s)$ est

$$\begin{aligned} G^\star(s) &\asymp \frac{1}{\log 2}\frac{1}{(s-1)^2} + (\frac{1}{2} + \frac{\gamma}{\log 2})\frac{1}{s-1} \\ &\quad + \sum_{k\in\mathbb{Z}^\star} \frac{1}{\log 2}\frac{\Gamma(1-\chi_k)}{s-\chi_k} \qquad (s \in <\delta, \frac{3}{2}>), \end{aligned}$$

où γ désigne la constante d'Euler. Il faut pour la calculer se rappeler que

$$\begin{aligned} \Gamma(s) &= \frac{1}{s} - \gamma + O(s) \qquad (s \to 0) \\ (1-2^{-s})^{-1} &= \frac{1}{\log 2}\frac{1}{s} + \frac{1}{2} + O(s) \qquad (s \to 0) \end{aligned}$$

On obtient alors le développement de G(x) quand $x \to 0$

$$\begin{aligned} G(x) &= \frac{-1}{\log 2}x^{-1}\log(x) + (\frac{1}{2} + \frac{\gamma}{\log 2})x^{-1} \\ &\quad + \frac{x^{-1}}{\log 2}P(\log_2 x) + O(x^{-\delta}). \end{aligned}$$

Le polynôme $P(\log_2 x)$ regroupe les termes imaginaires liés aux χ_k.

$$P(\log_2 x) = \sum_{k\in\mathbb{Z}^\star} \Gamma(\frac{2\mathrm{i}k\pi}{\log 2})e^{-2\mathrm{i}k\pi\log_2(x)}.$$

Du fait de la forte décroissance de $\Gamma(\frac{2\mathrm{i}k\pi}{\log 2})$ quand $k \to +\infty$, $|P(\log_2 x)|$ est majoré par 10^{-6}. Par un changement de variables $\frac{1}{x} \mapsto x$, on trouve le développement de $F(x)$ en $+\infty$, pour tout $c > 1$,

$$\begin{aligned} \mathrm{E}[L_N] &= F(n) \\ &= \frac{1}{\log 2}n\log(n) + (\frac{1}{2} + \frac{\gamma}{\log 2})n \\ &\quad - \frac{n}{\log 2}P(\log_2 n) + O(n^{-c}). \end{aligned}$$

6.2. **Nombre de noeuds internes.** Le développement asymptotique du nombre de noeuds internes se calcule de la même façon que pour la longueur de cheminement externe. On considère la somme harmonique

$$F(x) = \sum_{h\in H} f(u_h x), \text{avec } f(x) = (1 - e^{-x}(1+x)).$$

L'espérance $\mathrm{E}[T_N]$ est égale à $F(n)$. On veut déterminer le développement asymptotique de $F(x)$ quand $x \to \infty$. De même que précédemment, on va chercher le développement asymptotique $G(x) = F(1/x)$ en 0^+. On a

$$G(x) = \sum_{h\in H} g(\mu_h x),$$

avec $g(x) = f(1/x)$ et $\mu_h = {u_h}^{-1}$. On calcule la transformée de Mellin de g

$$g^\star(s) = f^\star(-s) = -(-s+1)\Gamma(-s) \qquad (s \in <0, 2>).$$

La série de Dirichlet associée à $G(x)$ est toujours

$$\Lambda(s) = \frac{1}{1-2^{1-s}} \qquad (s \in <1, +\infty>).$$

Le théorème des sommes harmoniques s'applique encore. Soit $\delta < 0$. Cette fois-ci, $G^\star$ n'a que des pôles simples sur $<\delta, \frac{3}{2}>$ situés en 0 et aux points

$$\chi_k = 1 + \frac{2\mathrm{i}k\pi}{\log 2}, \text{ pour } k \in \mathbb{Z}.$$

La partie singulière de $G^\star(s) = \Lambda(s)g^\star(s)$ s'écrit

$$\begin{aligned} G^\star(s) \asymp\ & \frac{1}{\log 2}\frac{1}{s-1} + \frac{1}{s} \\ & + \sum_{k\in\mathbb{Z}^\star} \frac{1}{\log 2}\frac{(\chi_k - 1)\Gamma(-\chi_k)}{s-\chi_k} \qquad (s \in <\delta, \frac{3}{2}>). \end{aligned}$$

On a besoin des 2 développements suivants

$$\begin{aligned} \Gamma(s) &= -\frac{1}{s+1} + O(1) \qquad (s \to -1) \\ (1-2^{-s})^{-1} &= \frac{1}{\log 2}\frac{1}{s} + \frac{1}{2} + O(s) \qquad (s \to 0). \end{aligned}$$

On obtient immédiatement le développement de $G(x)$ quand $x \to 0$

$$G(x) = \frac{1}{\log 2}x^{-1} + 1 + \frac{x^{-1}}{\log 2}Q(\log_2 x) + O(x^{-\delta}).$$

Le polynôme $Q(\log_2 x)$ est un polynôme regroupant les termes imaginaires liés aux χ_k. $Q(u)$ est périodique de période 1, et d'amplitude négligeable devant 1. Par le changement de variables $\frac{1}{x} \mapsto x$, on trouve, pour tout $c > 1$, le développement de $F(x)$ en $+\infty$

$$F(x) = \frac{1}{\log 2}x + 1 - \frac{x}{\log 2}Q(\log_2 x) + O(x^{-c}).$$

6.3. **Hauteur.** La hauteur ne se traite pas aussi simplement puisqu'elle ne s'exprime pas directement comme une somme harmonique.

$$\mathrm{E}[H_N] = \sum_{\ell\geq 0}(1 - \prod_{h\in H_\ell} e^{-nu_h}(1+nu_h)).$$

On va ici se limiter à la majoration de $\mathrm{E}[H_N]$ grâce à l'inégalité

$$(1+x)e^{-x} \geq e^{-\frac{x^2}{2}}.$$

On en déduit

$$\mathrm{E}[H_N] \leq \sum_{\ell\geq 0}(1 - \prod_{h\in H_\ell} e^{-(nu_h)^2/2}) = \sum_{\ell\geq 0}(1 - e^{\frac{-n^2}{2}S_\ell^{(2)}}),$$

avec

$$S_\ell^{(2)} = \sum_{h \in H_\ell} u_h^2.$$

Dans le cas binaire, $S_\ell^{(2)} = 2^\ell 2^{-2\ell} = 2^{-\ell}$. On va donc considérer la somme harmonique $F(x)$

$$F(x) \;=\; \sum_{\ell \geq 0} f(\mu_h x)$$

avec $f(x) = 1 - e^{-x}$ et $\mu_h = 2^{-l}$. On a alors $\mathrm{E}[H_N] \leq F(n^2/2)$. Il reste à chercher le développement asymptotique de $G(x) = F(\frac{1}{x})$ en 0^+. On calcule

$$\begin{aligned} g^\star(s) &= -\Gamma(-s) \qquad (s \in <0,1>) \\ \Lambda(s) &= \frac{1}{1-2^{-s}} \qquad s \in <1,+\infty> \end{aligned}$$

Les deux fonctions admettent bien un prolongement méromorphe sur $\mathbb{C}$. Soit $\delta < 0$. La partie singulière de $G^\star(s)$ sur $<\delta, \frac{1}{2}>$ est

$$\begin{aligned} G^\star(s) \;\asymp\; & \frac{1}{\log 2}\frac{1}{s^2} + (\frac{1}{2} + \frac{\gamma}{\log 2})\frac{1}{s} \\ & + \sum_{k \in \mathbb{Z}^\star} \frac{1}{\log 2}\frac{\Gamma(-\chi_k)}{s - \chi_k} \qquad (s \in <\delta, \frac{1}{2}>), \end{aligned}$$

où γ désigne la constante d'Euler et $\chi_k = 2\mathrm{i}k\pi/\log 2$. On obtient immédiatement le développement de $G(x)$ quand $x \to 0$

$$\begin{aligned} G(x) \;=\; & \frac{-1}{\log 2}\log(x) + (\frac{1}{2} + \frac{\gamma}{\log 2}) \\ & + \frac{1}{\log 2} R(\log_2 x) + O(x^{-\delta}). \end{aligned}$$

Encore une fois $R(\log_2 x)$ est périodique d'amplitude négligeable devant 1. Le développement asymptotique de $F(x)$ est donc, pour tout $c > 0$,

$$F(x) \;=\; \frac{1}{\log 2}\log(x) + (\frac{1}{2} + \frac{\gamma}{\log 2}) - \frac{1}{\log 2} R(\log x) + O(x^c),$$

et

$$F(n^2/2) \;=\; \frac{2}{\log 2}\log(x) + (\frac{-1}{2} + \frac{\gamma}{\log 2}) - \frac{1}{\log 2} R(\log_2(n^2/2) + O(n^{-2c}).$$

On a donc :

$$E[H_N] \leq \frac{2}{\log 2}\log(x) + (\frac{-1}{2} + \frac{\gamma}{\log 2}) - \frac{1}{\log 2} R(\log 2^2/2) + O(n^{-2c}).$$

Remarque :

(1) On peut montrer en utilisant une méthode plus fine que la hauteur a effectivement pour espérance en $\frac{2}{\log 2}\log(x) + O(1)$. La majoration utilisée est une "bonne" majoration.

(2) L'utilisation de séries génératrices peut conduire aux mêmes résultats. Cette méthode se base sur des récurrences traduisant la décomposition de l'arbre en sous-arbres issus de la racine, et peut s'appliquer ici, car lorsque l'on tire un nombre selon une loi uniforme sur $]0,1[$, chaque bit de ce nombre est indépendant des autres et est égal 0 ou 1 avec une probabilité $\frac{1}{2}$. Pour calculer le nombre de noeuds internes d'un trie sur k chaînes binaires, il faut remarquer que le nombre de noeuds internes d'un arbre est égale à la somme des nombres de noeuds internes de chaque sous arbres auquel on ajoute 1 pour la racine. En moyenne on a

$$T_k = 1 + \sum_{m=0}^{k} \frac{\binom{m}{k}}{2^k}(T_m + T_{k-m}).$$

En appliquant des méthodes classiques sur la série génératrice résultante, on peut résoudre cette récurrence. On retrouve alors bien l'espérance du nombre de noeuds internes dans le modèle de Bernouilli.

(3) On peut reprendre l'étude pour le cas biaisé des tries binaire, où chaque bit d'une chaine est toujours indépendant des autres, mais où le bit tiré est 0 ou 1 avec une probabilité p et $(1-p)$. L'analyse ne demande que très peu de modifications et conduit à des expressions similaires.

7. Continuants, Opérateur de Ruelle Mayer

La longueur de l'intervalle fondamental de rang k associé à h, composée de k applications h_m, est notée u_h. Dans le cas binaire, on a déjà obtenu la relation $u_h = 2^{-k}$. Lorsque h est la composée d'homographies, on peut calculer u_h en introduisant l'opérateur de Ruelle Mayer et les polynômes continuants.

7.1. **Continuants.** A un réel $x \in \mathcal{I} =]0,1[$, on associe la suite $x_0 = x, x_1, x_2, \ldots, x_k, \ldots$ des itérés de x. Si le k-ème itéré existe, l'exécution de l'algorithme des fractions continues sur l'entrée x_0 se traduit par un développement en fraction continue

$$x_0 = \cfrac{1}{m_1 + \cfrac{1}{m_2 + \cfrac{1}{\ddots \cfrac{}{m_k + x_k}}}},$$

où les entiers m_j sont supérieurs ou égaux à 1. La relation $x_k = h(x_0)$ définit une homographie h de hauteur k, associée à un k-uplet $(m_1, m_2, \ldots, m_k)$ d'entiers $m_i \geq 1$,

$$h(x) = \cfrac{1}{m_1 + \cfrac{1}{m_2 + \cfrac{1}{\ddots \cfrac{}{m_k + x}}}}.$$

Une telle homographie s'exprime alors à l'aide des continuants

$$h(z) = \frac{P_k + zP_{k-1}}{Q_k + zQ_{k-1}},$$

où

$$\begin{array}{rclrcl} Q_k & = & Q_k(m_1, \ldots, m_k), & Q_{k-1} & = & Q_{k-1}(m_1, \ldots, m_{k-1}), \\ P_k & = & Q_{k-1}(m_2, \cdots, m_k), & P_{k-1} & = & Q_{k-2}(m_2, \ldots, m_{k-1}). \end{array}$$

Les polynômes continuants sont définis par récurrence

$$Q_k(m_1, m_2, \ldots, m_k) = m_k Q_{k-1}(m_1, \ldots, m_{k-1}) + Q_{k-2}(m_1, \ldots, m_{k-2}), \tag{3}$$

avec $Q_0 = 1$, $Q_1(m_1) = m_1$. Le polynôme continuant $Q_k(m_1, m_2, \ldots, m_k)$ est aussi la somme de tous les monômes obtenus en barrant deux variables consécutives $m_i m_{i+1}$ dans le produit $m_1 m_2 \cdots m_k$. Les continuants vérifient une propriété de *symétrie*

$$Q_k(m_1, \ldots, m_k) = Q_k(m_k, \ldots, m_1),$$

et l'*identité du déterminant*

$$Q_k P_{k-1} - Q_{k-1} P_k = (-1)^k.$$

L'intervalle fondamental $h(\mathcal{D})$ s'exprime en fonction des continuants :

$$h(\mathcal{D}) = \begin{cases} [\frac{P_k}{Q_k}, \frac{P_{k-1}+P_k}{Q_{k-1}+Q_k}[& \text{si } k \text{ est pair,} \\ ]\frac{P_{k-1}+P_k}{Q_{k-1}+Q_k}, \frac{P_k}{Q_k}] & \text{si } k \text{ est impair,} \end{cases}$$

et est de longueur égale à

$$|h(\mathcal{D})| = \frac{1}{Q_k(Q_k + Q_{k-1})}.$$

7.2. **Opérateur de Ruelle Mayer.** Dans toute la suite, $\mathcal{J}$ et $\mathcal{V}$ désignent respectivement le segment et le disque ouvert de centre 1 et de rayon 5/4. Soit l'ensemble $A_\infty(\mathcal{V})$ formé par les fonctions f holomorphes dans $\mathcal{V}$ et continues sur $\bar{\mathcal{V}}$. Muni de la norme sup $|.|$ définie par

$$|f| = \mathrm{S}up\ \{|f(z)|, z \in \mathcal{V}\},$$

cet ensemble est un *espace de Banach.*

Définition 4. *On définit les opérateurs $\mathcal{G}_s$ (dits de Ruelle Mayer) sur $A_\infty(\mathcal{V})$, pour un paramètre s vérifiant $\Re(s) > 1$, par la relation*

$$\mathcal{G}_s[f](z) = \sum_{m \geq 1} \frac{1}{(m+z)^s}\ f(\frac{1}{m+z}).$$

La proposition justifiant l'introduction de cet opérateur est la suivante :

Proposition 2. *L'itéré d'ordre k de cet opérateur engendre alors les continuants d'ordre k dans le sens suivant*

$$\mathcal{G}_s^k[f](z) = \sum_{\substack{m_1 \ldots m_k \\ m_i \geq 1}} \frac{1}{(Q_{k-1}z + Q_k)^s}\ f(\frac{P_{k-1}z + P_k}{Q_{k-1}z + Q_k}),$$

et, en particulier

$$\mathcal{G}_s^k[f](0) = \sum_{\substack{m_1 \dots m_k \\ m_i \geq 1}} \frac{1}{Q_k^s}\ f(\frac{P_k}{Q_k}) \ = \sum_{\substack{m_1 \dots m_k \\ l_i \geq 1}} \frac{1}{Q_k^s}\ f(\frac{Q_{k-1}}{Q_k}),$$

la seconde égalité étant due aux propriétés de symétrie.

Preuve

Elle se fait par récurrence en utilisant les propriétés des continuants.

– Si k=1,

$$\begin{aligned} \mathcal{G}_s[f](z) &= \sum_{m\geq 1} \frac{1}{(m+z)^s}\ f(\frac{1}{m+z}) \\ &= \sum_{m\geq 1} \frac{1}{(Q_0 z + Q_1)^s}\ f(\frac{P_0 z + P_1}{Q_0 z + Q1}). \end{aligned}$$

car les premiers continuants vérifient $P_0 = 0, Q_0 = P_1 = 1, Q_1(m) = m$.

– Supposons maintenant que

$$\mathcal{G}_s^k[f](z) = \sum_{\substack{m_1 \dots m_k \\ m_i \geq 1}} \frac{1}{(Q_{k-1} z + Q_k)^s}\ f(\frac{P_{k-1} z + P_k}{Q_{k-1} z + Q_k}),$$

alors

$$\begin{aligned} \mathcal{G}_s^{k+1}[f](z) &= \sum_{m\geq 1} \frac{1}{(m+z)^s} \sum_{\substack{m_1 \dots m_k \\ m_i \geq 1}} \frac{1}{(\frac{Q_{k-1}}{m+z} + Q_k)^s}\ f(\frac{\frac{P_{k-1}}{m+z} + P_k}{\frac{Q_{k-1}}{m+z} + Q_k}) \\ &= \sum_{\substack{m_1 \dots m_{k+1} \\ m_i \geq 1}} \frac{1}{(Q_{k-1} + m_{k+1} Q_k + Q_k z)^s}\ f(\frac{P_{k-1} + m_{k+1} P_k + P_k z}{Q_{k-1} + m_{k+1} Q_k + Q_k z}) \\ &= \sum_{\substack{m_1 \dots m_{k+1} \\ m_i \geq 1}} \frac{1}{(Q_k z + Q_{k+1})^s}\ f(\frac{P_k z + P_{k+1}}{Q_k z + Q_{k+1}}). \end{aligned}$$

La dernière égalité est obtenue en réduisant au même dénominateur et en utilisant la relation de récurrence $Q_{k+1} = m Q_k + Q_{k-1}$.

L'opérateur $\mathcal{G}_s$ a de multiples propriétés, mais dans le cadre de l'étude des tries basés sur le développement en fraction continue , ce sont surtout les propriétés spectrales dans un voisinage de $s = 2$ et $s = 4$ qui nous intéressent.

7.2.1. *Les propriétés spectrales de l'opérateur de Ruelle Mayer.* La principale caractéristique de l'opérateur de Ruelle Mayer est qu'il admet un spectre discret. De plus, lorsque s est réel ($s > 1$), l'opérateur $\mathcal{G}_s$ satisfait à une propriété de Perron-Frobenius :

Théorème 3 (Mayer)**.** *Pour un réel $s > 1$, l'opérateur $\mathcal{G}_s : A_\infty(\mathcal{V}) \to A_\infty(\mathcal{V})$ a une valeur propre dominante $\lambda(s)$ qui est simple et strictement plus grande que toutes les autres valeurs propres en valeur absolue. Le vecteur propre correspondant f_s est strictement positif sur $\mathcal{J}$. L'opérateur adjoint $\mathcal{G}_s^\star : A_\infty^\star(\mathcal{V}) \to A_\infty^\star(\mathcal{V})$ a un vecteur propre dominant $f_s^\star$ correspondant*

à $\lambda(s)$ qui vérifie $f_s^\star[f] > 0$ si $f > 0$ sur $\mathcal{J}$. Si $\mathcal{P}_s$ désigne la projection sur le sous-espace dominant définie par $\mathcal{P}_s = f_s^\star \otimes f_s$, alors $\mathcal{G}_s$ admet la représentation

$$\mathcal{G}_s = \lambda(s)\mathcal{P}_s + \mathcal{N}_s,$$

où $\mathcal{P}_s \circ \mathcal{N}_s = \mathcal{N}_s \circ \mathcal{P}_s = 0$. Le rayon spectral de $\mathcal{N}_s$ est strictement plus petit que $\lambda(s)$. Si f est une fonction de $A_\infty(\mathcal{V})$ strictement positive sur $\mathcal{J}$, on a

$$\mathcal{G}_s^k[f](z) = \lambda(s)^k f_s^\star[f]\, f_s(z) \ + \mathcal{N}_s^k[f](z),$$

pour tout $k \geq 1$ et tout z dans $\mathcal{V}$.

Au voisinage de l'axe réel, on peut appliquer la théorie des perturbations et montrer que l'opérateur $\mathcal{G}_s$ conserve ses propriétés spectrales dominantes au voisinage d'un paramètre s réel, ce qui est décrit dans le théorème suivant

Théorème 4 (Hensley, Faivre)**.** *Soit $\sigma > 1$ un nombre réel fixé. Il existe un voisinage complexe Ω de σ pour lequel les propriétés spectrales dominantes de $\mathcal{G}_\sigma$ se prolongent à $\mathcal{G}_s$: les quatre quantités $\lambda(s)$, f_s, $f_s^\star$ (et donc $\mathcal{P}_s$), $\mathcal{N}_s$ y sont bien définies, y représentent les objets spectraux dominants de $\mathcal{G}_s$ et sont analytiques en s ; de plus, le rayon spectral de $\mathcal{N}_s$ est strictement inférieur à $|\lambda(s)|$. Sur le même voisinage Ω, et pour toute fonction f de $A_\infty(\mathcal{V})$ strictement positive sur $\mathcal{J}$, on a*

$$\mathcal{G}_s^k[f](z) = \lambda(s)^k \ f_s(z) \ f_s^\star[f] \ + \mathcal{N}_s^k[f](z)$$

pour tout $k \geq 1$, et pour tout point z de $\mathcal{V}$.

Pour $s = 2$, l'opérateur $\mathcal{G}_s$ a des propriétés spectrales dominantes bien connues ; la valeur propre dominante $\lambda(s)$ est égale à 1, la valeur propre subdominante, également simple, et déterminée d'abord par Wirsing, vaut $\lambda^{(2)}(s) \approx -0.303663$; le vecteur propre dominant f_2, correspondant à la densité–limite de l'algorithme des fractions continues, et le vecteur propre dominant $f_2^\star$ de l'opérateur adjoint sont tous deux explicites et égaux respectivement à

$$f_2(z) = \frac{1}{1+z} \quad \text{et} \quad f_2^\star[f] = \frac{1}{\log 2} \int_0^1 f(x)dx.$$

Pour $s = 4$, la valeur propre dominante vaut approximativement

$$\lambda(4) \approx 0.1995;$$

cette valeur apparaît naturellement dans l'analyse en moyenne d'algorithmes s'apparentant à des extensions de l'algorithme d'Euclide en dimension 2 (comme la réduction des réseaux).

8. Trie en fraction continue : comportement asymptotique

8.1. **Longueur de cheminement externe et nombre de noeuds internes.** La série de Dirichlet associée aux sommes harmoniques en jeu est

$$\Lambda(s) = \sum_{h \in H} u_h^s,$$

où u_h est maintenant la longueur un intervalle fondamental associé à h une homographie de hauteur k. Il convient donc d'en faire l'étude.

8.1.1. *Etude de la série de Dirichlet* $\Lambda(s)$. Cette fonction s'exprime grâce aux itérés de l'opérateur de Ruelle Mayer. En effet, on calcule :

$$\begin{aligned}\Lambda(s) &= \sum_{h\in H} u_h^s \\ &= \sum_{k\geq 0}\sum_{|h|=k} \frac{1}{{Q_k}^s(Q_{k-1}+Q_k)^s} \\ &= \sum_{k\geq 0}\sum_{|h|=k} \frac{1}{{Q_k}^{2s}(1+\frac{Q_{k-1}}{Q_k})^s} \\ &= \sum_{k\geq 0} \mathcal{G}_{2s}^k[\frac{1}{(1+x)^s}](0).\end{aligned}$$

Formellement, on a

$$\Lambda(s) = (I-\mathcal{G}_2 s)^{-1}[\frac{1}{(1+x)^s}](0).$$

D'après les propriétés spectrales de $\mathcal{G}_{2s}$, $\Lambda(s)$ posséde un pôle simple en $s=1$. En effet, au voisinage de $s=1$, $\mathcal{G}_{2s}$ s'écrit

$$\mathcal{G}_s^k[f](z) = \lambda(s)^k\ f_s(z)\ f_s^\star[f]\ + \mathcal{N}_s^k[f](z),$$

où

$$f_2(z) = \frac{1}{1+z}, f_2^\star(f) = \frac{1}{\log 2}\int_0^1 f(x)dx.$$

Dans notre cas, $f(z) = \frac{1}{(1+z)^s}$, on en déduit que, dans un voisinage de $s=1$,

$$\Lambda(s) = \frac{1}{1-\lambda(2s)} f_{2s}(z)\ f_{2s}^\star[f](0)\ + (I-\mathcal{N}_{2s})^{-1}[f](0).$$

Le rayon spectral de $\mathcal{N}_{2s}$ étant strictement inférieur à 1, $(I-\mathcal{N}_2 s)^{-1}$ est analytique en $s=1$.

Le résidu de Λ en $s=1$ est donc

$$\lim_{s\to 1}(s-1)\frac{1}{\log 2}\frac{1}{1-\lambda(2s)}\int_0^1 \frac{dt}{1+t} = -\frac{1}{2\lambda'(2)} = \frac{6\log 2}{\pi^2}.$$

Pour appliquer le théorème des sommes harmoniques, il faut encore trouver un réel $\delta<1$ tel que Λ soit analytique sur $\Re(s)=\delta$ et aussi prouver que Λ lorsque l'on s'éloigne de l'axe réel à l'interieur du'une bande verticale ne croisse pas trop vite. C'est l'objet de la proposition suivante :

Proposition 3. *$\Lambda(s)$ n'a pas de pôles sur $<\frac{1}{2},1>$ et, dans le domaine $\mathcal{H}$, est en $O(|s|^r)$ pour un $r\geq 0$, avec*

$$\mathcal{H} = \{s\in\mathbb{C}\mid \frac{1}{2}<\alpha<\Re(s)<\beta<2 \text{ et } |Im(s)\geq 4\}.$$

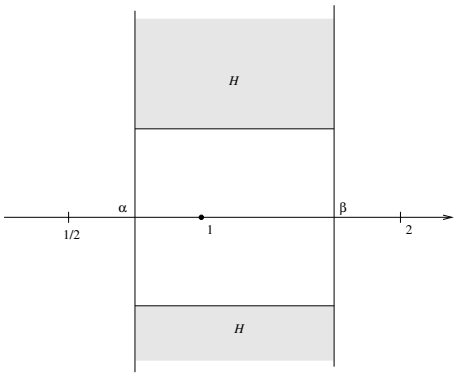

FIG. 6. situation pour la proposition 3.

Pour clarifier les choses on peut se référer à la figure 6. **Preuve :** La fonction Λ s'écrit en fonction des continuants

$$\Lambda(s) = \sum_h {u_h}^s = 1 + \sum_{k=1}^{\infty} \sum_{|h|=k} \frac{1}{{Q_k}^s (Q_k + Q_{k-1})^s}.$$

Or d'après une propriétés des continuants :

$$Q_{k-1}(m_1, \ldots, m_{k-1}) = P_k(m_{k-1}, \ldots, m_1),$$

donc

$$\Lambda(s) \;=\; 1 + \sum_{k=1}^{\infty} \sum_{\substack{m_1,\ldots,m_k \\ m_i \geq 1}} \frac{1}{{Q_k}^s (Q_k + P_k)^s}.$$

A chaque fraction irréductible c/d de l'intervalle $]0,1[$, on peut associer 2 développements en fraction continue, l'un propre où le dernier quotient vérifie $m_k \geq 1$, l'autre impropre où le dernier quotient vérifie $m_{k+1} = 1$. Ainsi, tout couple (c,d) d'entiers vérifiant $\gcd(c,d) = 1$, $d \geq 2$ et $0 < c < d$ peut s'écrire de 2 manières différentes comme un couple (P_k, Q_k). Par ailleurs, le nombre 1 s'écrit de manière unique $(1,1)$ et on obtient donc

$$\begin{aligned} \Lambda(s) &= 1 + \frac{1}{2^s} + 2 \sum_{\substack{d \geq 2 \\ 0<c<d \\ \gcd(c,d)=1}} \frac{1}{d^s (c+d)^s} \\ &= 1 + \frac{1}{2^s} + 2 \sum_{\substack{d \geq 2 \\ 0<c<d \\ \gcd(c,d)=1}} \frac{1}{d^s (c+d)^s} \end{aligned}$$

On considère les 2 ensembles

$$\begin{aligned} \Omega &= \{(d,c) \in \mathbb{Z}^2 \mid d \geq 2, 0 < c < d\} \\ \Omega' &= \{(d,c) \in \mathbb{Z}^2 \mid d \geq 2, 0 < c < d, \gcd(c,d) = 1\}. \end{aligned}$$

Tout élément (d,c) de Ω s'écrit de manière unique $(d=ud', c=uc')$ avec $(c',d') \in \Omega'$ et $u>1$ (c'est la propriété du pgcd).

$$\begin{aligned}
\sum_{(d,c)\in\Omega} \frac{1}{d^s(c+d)^s} &= \sum_{u\geq 1} \sum_{(c,d)\in\Omega'} \frac{1}{(ud)^s(uc+ud)^s} \\
&= \sum_{u\geq 1} \frac{1}{u^{2s}} \sum_{(c,d)\in\Omega'} \frac{1}{d^s(c+d)^s} \\
&= \zeta(2s) \sum_{(c,d)\in\Omega'} \frac{1}{d^s(c+d)^s} \\
&= \zeta(2s) \sum_{\substack{d\geq 2 \\ 0<c<d \\ \gcd(c,d)=1}} \frac{1}{d^s(c+d)^s}
\end{aligned}$$

La condition $\gcd(d,c)=1$ peut donc être supprimée en divisant par $\zeta(2s)$. On a alors :

$$\begin{aligned}
\Lambda(s) &= 1+\frac{1}{2^s}+\frac{2}{\zeta(2s)} \sum_{d\geq 2} \frac{1}{d^s} \sum_{0<c<d} \frac{1}{(c+d)^s} \\
&= 1+\frac{1}{2^s}+\frac{2}{\zeta(2s)} \sum_{d\geq 2} \frac{1}{d^s} \sum_{d<c<2d} \frac{1}{c^s}
\end{aligned}$$

Dans la double somme, on a besoin de bornes non strictes sur c (pour utiliser une majoration par des intégrales). On a

$$\begin{aligned}
\sum_{d\geq 2} \frac{1}{d^s} \sum_{d<c<2d} \frac{1}{c^s} &= \sum_{d\geq 1} \frac{1}{d^s} \sum_{d<c<2d} \frac{1}{c^s} \\
&= \sum_{d\geq 1} \frac{1}{d^s} \sum_{d\leq c\leq 2d} \frac{1}{c^s} - \sum_{d\geq 1} \frac{1}{d^s}\left(\frac{1}{d^s}+\frac{1}{(2d)^s}\right) \\
&= \sum_{d\geq 1} \frac{1}{d^s} \sum_{d\leq c\leq 2d} \frac{1}{c^s} - \left(1+\frac{1}{2^s}\right) \sum_{d\geq 2} \frac{1}{d^{2s}} \\
&= -\left(1+\frac{1}{2^s}\right)\zeta(2s) \sum_{d\geq 1} \frac{1}{d^s} + \sum_{d\leq c\leq 2d} \frac{1}{c^s}
\end{aligned}$$

En résumé, on a obtenu :

$$\begin{aligned}
\Lambda(s) &= 1+\frac{1}{2^s}+\frac{2}{\zeta(2s)}\left(\sum_{d\geq 1} \frac{1}{d^s} \sum_{d\leq c\leq 2d} \frac{1}{c^s} - \left(1+\frac{1}{2^s}\right)\zeta(2s)\right) \\
&= -1-\frac{1}{2^s}+\frac{2}{\zeta(2s)} \sum_{d\geq 1} \frac{1}{d^s} \sum_{d\leq c\leq 2d} \frac{1}{c^s}
\end{aligned}$$

La formule de sommation d'Euler-Maclaurin permet d'évaluer

$$\sum_{d\leq c\leq 2d}\frac{1}{c^s}.$$

Théorème 5. *Soit f une fonction à valeurs dans $\mathbb{C}$ définie sur l'intervalle $[a,b]$, où a et b sont des entiers. Si f est dérivable et à dérivée continue sur l'intervalle $[a,b]$, alors on a :*

$$\sum_{a\leq k\leq b} f(k)=\int_a^b f(x)\,dx+\frac{f(a)+f(b)}{2}+\int_a^b(\{x\}-\frac{1}{2})f'(x)\,dx,$$

où $\{x\}$ est la partie fractionnaire de x.

Ainsi, on a :

$$\begin{aligned}\sum_{d\leq c\leq 2d}\frac{1}{c^s} &= \int_d^{2d}\frac{dx}{x^s}+\frac{1}{2}(\frac{1}{d^s}+\frac{1}{(2d)^s})+\int_d^{2d}(\{x\}-\frac{1}{2})\frac{-s}{x^{s+1}}\,dx\\ &= \frac{1}{s-1}(1-2^{1-s})\frac{1}{d^{s-1}}+\frac{1}{2}(1+2^{-s})\frac{1}{d^s}-s\int_d^{2d}(\{x\}-\frac{1}{2})\frac{1}{x^{s+1}}\,dx\end{aligned}$$

On obtient finalement

$$\begin{aligned}\Lambda(s) &= -1-2^{-s}+\frac{2}{\zeta(2s)}\left(\frac{1-2^{1-s}}{s-1}\zeta(2s-1)+\frac{1+2^{-s}}{2}\zeta(2s)\right)-R(s)\\ &= 2\,\frac{1-2^{1-s}}{s-1}\frac{\zeta(2s-1)}{\zeta(2s)}-R(s)\end{aligned}$$

avec :

$$R(s)=\frac{2s}{\zeta(2s)}\sum_{d\geq 1}\frac{1}{d^s}\int_d^{2d}(\{x\}-\frac{1}{2})\frac{1}{x^{s+1}}\;dx.$$

Posant $\sigma=\Re(s)$, on remarque

$$\left|\int_d^{2d}(\{x\}-\frac{1}{2})\frac{1}{x^{s+1}}\,dx\right|\leq\int_d^{2d}\left|\frac{1}{x^{s+1}}\right|\,dx=\int_d^{2d}\frac{1}{x^{\sigma+1}}\,dx=\frac{1}{\sigma}(2^{-\sigma}-1)\frac{1}{d^\sigma}$$

d'où :

$$|R(s)|\leq\left|\frac{2s}{\zeta(2s)}\right|\sum_{d\geq 1}\left|\frac{1}{d^s}\right|\frac{1}{\sigma}(2^{-\sigma}-1)\frac{1}{d^\sigma}=2\,\frac{|s|}{\sigma}(2^{-\sigma}-1)\frac{\zeta(|2\sigma|)}{|\zeta(2s)|}$$

Ainsi $R(s)$ est analytique pour $\Re(s)>\frac{1}{2}$. On voit donc que $\Lambda(s)$ n'a pas de pôles pour $\frac{1}{2}<\Re(s)<1$.

Remarque – De plus, on peut calculer le résidu de $\Lambda(s)$ en $s=1$. Sachant que

$$\begin{array}{l}\zeta(2s-1)=\frac{1}{2(s-1)}+O(1)\quad(s\to 1)\\ 2^{1-s}-1=(s-1)\log 2+\frac{1}{2}+O((s-1)^2)\quad(s\to 1)\end{array}\ ,$$

le résidu de $\Lambda(s)$ en $s=1$ est donc égal à :

$$\mathrm{R}es(\Lambda(s),s=1)=\frac{\log 2}{\zeta(2)}=\frac{6\log 2}{\pi^2}.$$

La deuxième partie de la preuve peut être faite grâce à des majorations sur un domaine approprié de $\zeta(2s-1)^{-1}$ et de $\zeta(2s)$ (d'après [3]).

8.1.2. *Longueur de cheminement externe.* L'étude est un peu moins précise que pour les tries binaires. En effet , on a déterminé le résidu de Λ en $s=1$. Dans le cas où $g^\star$ admet aussi un pôle en $s=1$, il faudrait pour obtenir *la partie singulière en $s=1$, connaître le terme constant de la série de Laurent. On peut donc introduire la constante K telle que

$$\Lambda(s) = \frac{-1}{2\lambda'(2)}\frac{1}{s-1} + K + O(s-1) \text{ quand } x \to 1.$$

Le mode de raisonnement est le même que pour les tries binaires. On cherche le développement asymptotique de la somme harmonique

$$F(x) = \sum_{h\in H} f(u_h x),$$

avec

$$f(x) = x(1-e^{-x}).$$

De même que précédemment, on va appliquer le théorème des sommes harmoniques à $G(x) = F(1/x)$ pour obtenir le développement asymptotique de G en 0^+. La transformée de Mellin de la fonction de base $g(x) = f(1/x)$, ayant pour bande fondamentale $< 1, 2 >$, est

$$g^\star(s) = -\Gamma(-s+1) \qquad (s \in < 1, 2 >).$$

La fonction $g^\star$ admet un unique pôle simple en $s=1$ sur $< -\infty, \frac{3}{2} >$. D'après l'étude de la fonction Λ du paragraphe précédent, $G^\star(s) = g^\star(s)\Lambda(s)$ a pour partie singulière en $s=1$,

$$G^\star(s) \asymp \frac{-1}{\lambda'(2)}\frac{1}{(s-1)^2} + (K + \frac{\gamma}{\log 2})\frac{1}{s-1} \qquad (s=1).$$

En appliquant le théorème des sommes harmoniques, puis en appliquant un changement de variable pour obtenir le comportement de $F(x)$ en $+\infty$, on obtient l'équivalent asymptotique de la longueur de cheminement externe $E[L_N] = F[n]$ dans le modèle de Poisson

$$\begin{aligned} \mathrm{E}[L_N] &= -\frac{1}{2\lambda'(2)} n\log n + \Theta(n) \\ &= \frac{6\log 2}{\pi^2} n\log n + \Theta(n). \end{aligned}$$

8.1.3. *Nombre de noeuds internes.* Le développement asymptotique du nombre de noeuds internes se calcule de la même façon que pour la longueur de cheminement externe, et fait intervenir lui aussi la fonction étudiée précédemment. On considère la somme harmonique

$$F(x) = \sum_{h\in H} f(u_h x),$$

avec

$$f(x) = (1 - e^{-x}(1+x)).$$

On passe à l'etude de $G(x) = F(1/x)$ quand $x \to 0^+$. La fonction $g(x) = f(1/x)$, fonction de base de G, a pour transformée de Mellin avec comme bande fondamentale $< 0, 2 >$

$$g^\star(s) = -(-s+1)\Gamma(-s) \qquad s \in < 0, 2 > .$$

La fonction $g^\star$ admet un pôle simple sur $< -\infty, \frac{3}{2} >$ en $s = 0$. C'est donc le pôle de Λ en $s = 1$ qui donne le terme asymptotique dominant. On trouve que l'espérance du nombre de noeuds internes d'un trie construit sur la représentation continue dans le modèle de Poisson est

$$\begin{aligned} \mathrm{E}[T_N] &= -\frac{1}{2\lambda'(2)} n + o(n) \\ &= \frac{6\log 2}{\pi^2} n + o(n) \end{aligned}$$

8.2. **Hauteur.** La hauteur ne se traite pas aussi simplement puisqu'elle ne s'exprime pas directement comme une somme harmonique. On a déjà montré la relation

$$\mathrm{E}[H_N] \leq \sum_{k\geq 0} (1 - e^{\frac{-n^2}{2} S_k^{(2)}}),$$

avec

$$S_k^{(2)} = \sum_{|h|=k} u_h^2 = \sum_{|h|=k} \frac{1}{{Q_k}^2 (1 + \frac{Q_{k-1}}{Q_k})^2} = G_4^k[\frac{1}{(1+x)^2}](0).$$

L'opérateur $\mathcal{G}_4$ possède une valeur propre dominante $\lambda(4)$. Il existe donc une constante c telle que

$$S_k^{(2)} \leq c\lambda(4)^k.$$

On en déduit

$$\mathrm{E}[H_N] \leq \sum_{k\geq 0} (1 - e^{-c\frac{n^2}{2}\lambda(4)^k}).$$

On va donc considérer la somme harmonique $F(x)$

$$F(x) = \sum_{k\geq 0} f(\mu_k x) \text{ avec } f(x) = 1 - e^{-x} \text{ et } \mu_k = \lambda(4)^k.$$

On a alors $\mathrm{E}[H_N] \leq F(c\frac{n^2}{2})$. Il reste à chercher le développement asymptotique de $G(x) = F(1/x)$ en 0^+ dont la fonction de Dirichlet associée à G est

$$\Lambda(s) = \frac{1}{1 - \lambda(4)^s}.$$

On voit que Λ possède une infinité de pôles sur la droite verticale $\Re(s) = 0$. La transformée de Mellin de la fonction de base g est $g^\star(s) = -\Gamma(-s)$ admet un unique pôle sur $< -\infty, \frac{1}{2} >$ en $s = 0$. On trouve ainsi le terme asymptotique de G en 0^+, dû au pole d'ordre 2 en $s = 0$. On obtient

$$G(x) = \frac{1}{\log \lambda(4)} \log x + O(1),$$

et donc

$$F(x) = -\frac{1}{\log \lambda(4)} \log x + O(1).$$

Finalement, on voit que l'espérance de la hauteur d'un trie de nombre baseé sur les fractions continues dans un modèle de Poisson est

$$\mathrm{E}[H_N] \leq F(cn^2/2) = -\frac{2}{\log \lambda(4)} \log n + O(1).$$

8.3. **Grandeurs fondamentales de l'analyse.**

8.3.1. *Entropie.* On définit l'entropie $H(S)$ d'une variable aléatoire discrète à valeurs dans $\{x_i\}_{i\in I}$ par

$$H(X) := -\sum_i -p_i \log p_i,$$

où $p_i = \Pr[X = x_i]$. Ainsi pour 2 variables aléatoires X et Y, à valeurs dans $\{x_i\}_{i\in I} et \{y_j\}_{j\in J}$, on a

$$H(X,Y) = \sum_{(i,j)\in I\times J} -p_{i,j} \log p_{i,j},$$

avec $p_{i,j} = \Pr[X = x_i \; et \; Y = y_j]$.

On peut généraliser l'entropie à n variables aléatoires. On définit l'entropie $H(S)$ d'une source S de variables aléatoires $(X_i)_{i\in\mathbb{N}}$ par

$$H(S) := \lim_{n\to+\infty} \frac{1}{n} H(X_1, \ldots, X_n), \text{quand elle existe.}$$

Si S est une source aléatoire de bits (X_i) tirés indépendamment avec une probabilité p de valoir 1 et une probabilité q de valoir 0, alors l'entropie de cette source est la limite quand $k \to \infty$ de

$$\frac{1}{k} H(x_1, \ldots, X_k) = \frac{1}{k} k H(X),$$

donc, $H(S) = -p \log p - q \log q$. Si S est une source de variables aléatoires donnant la suite des termes du développement en fraction continue d'un nombre tiré uniformément sur [0,1[, alors on a

$$\begin{aligned} H(X_1, \ldots, X_k) &= \frac{1}{k} \sum_{|h|=k} -u_h \log u_h \\ &= -\mathrm{E}[\log u_h] \\ &= -\mathrm{E}[\log \frac{1}{Q_k(Q_k + Q_{k-1})}] \\ &= 2\mathrm{E}[\log Q_k] + \mathrm{E}[\log(1 + \frac{Q_{k-1}}{Q_{k-1}})] \end{aligned}$$

Or $1 < 1 + \frac{Q_{k-1}}{Q_k} < 2$ donc l'espérance est inférieure ou gale à 2. En passant à la limite, on trouve

$$H(S) = \lim_{k\to\infty} \frac{1}{k} 2\mathrm{E}[\log Q_k] = -2\lambda'(2).$$

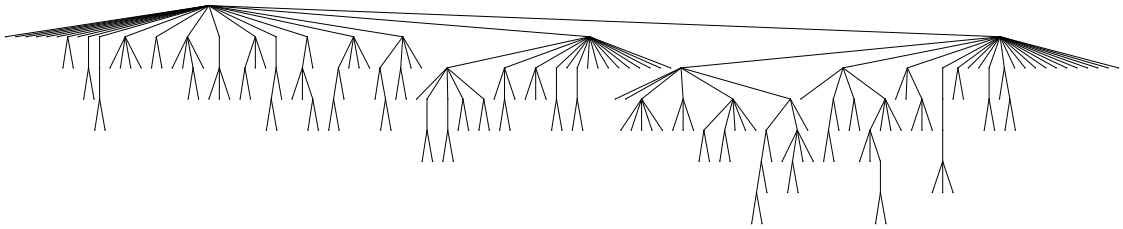

FIG. 7. Trie en fraction continue : nombre de feuilles (ou nombre de chaines stockées) 100, hauteur 7, longueur de cheminement externe 505, nombre de noeuds interne 75.

La dernière égalité est établie grâce à l'équivalent de $\mathrm{E}[\log Q_k]$ trouvé dans [Va2]. Ainsi, dans tous les cas, que ce soit pour les tries binaires biaisés où les tries en fraction continue, on arrive à

$$\mathrm{E}[L_N] = \frac{n \log n}{H(S)} + \Theta(n), \quad \mathrm{E}[T_N] = \frac{n}{H(S)} + o(n).$$

8.3.2. *Probabilité de coincïdence.* On considère 2 sources de variables aléatoires $(X_i)_{i\in\mathbb{N}}$ et $(Y_j)_{j\in\mathbb{N}}$. La probabilité de coincïdence est

$$P := \lim_{n\to\infty} P_n^{\frac{1}{n}} \; avec \; P_n = Pr[X_1 = Y_1 \; et \; X_2 = Y_2 \; et \; \ldots \; et \; X_n = Y_n].$$

Dans le cas binaire biaisé, les variables étant toutes indépendantes

$$P_n = \Pr[X_1 = Y_1] \Pr[X_2 = Y_2] \ldots \Pr[X_n = Y_n] = (p^2 + q^2)^n,$$

et par suite, $P = (p^2 + q^2)$. Dans le cas des fractions continues, de façon analogue, on voit que la probabilité que 2 séquences de chiffres coincident jusqu'au k-ième chiffre est

$$\sum_{|h|=k} {u_k}^2 \approx c\lambda(4)^k.$$

En prenant la racine k-ième, on a $P = \lambda(4)$. L'espérance de la hauteur d'un trie binaire biaisé et d'un trie en fraction continue s'exprime donc sous la forme

$$\mathrm{E}[H_N] \leq \frac{2 \log n}{|\log P|} + O(1).$$

9. CONCLUSION

A la suite de cette étude, on voit que le trie en fraction continue a un "bon" comportement algorithmique. Le tri de n rationnels, notamment, peut se faire sans problème de précision et efficacement.

Les résultats obtenus en base 2 sur un modèle uniforme de données sont généralisables pour n'importe quelle base b. Il est alors intéressant de noter que le fait de choisir un système de numération en fraction continue permet de s'affranchir du choix arbitraire d'une base, mais que, en ce qui concerne le comportement asymptotique sur les paramètres, tout se passe comme si l'on se trouvait dans un système de numération en base 5. Les formes d'arbres sont

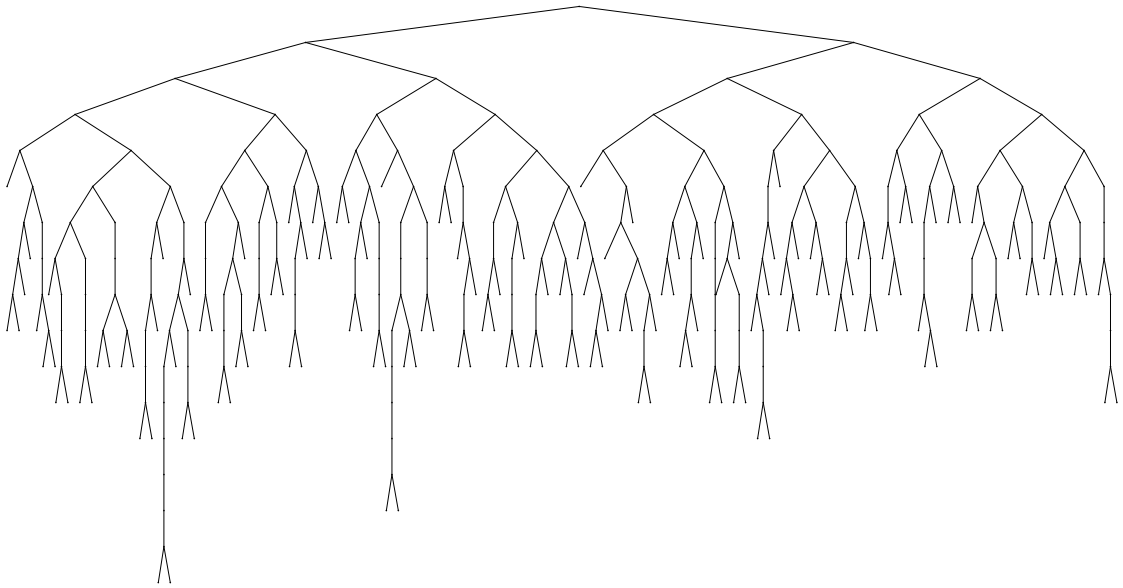

FIG. 8. Trie binaire correspondant à la figure 7 : nombre de feuilles (ou nombre de chaines stockées) 100, hauteur 16, longueur de cheminement externe 1323, nombre de noeuds interne 222.

assez caractéristiques et sont liées à la géométrie des intervalles fondamentaux (voir figure). L'étude des tries de nombres a mis en jeu des méthodes d'analyse fonctionnelle, et les valeurs moyennes obtenues font intervenir des constantes caractéristiques de l'algorithme d'Euclide. Il reste à généraliser ces études lorsque les données ne sont plus uniformes et/ou ne sont plus indépendantes. L'étude d'une distribution non uniforme a déjà été abordée par Devroye dans le cas de la numération binaire, mais les résultats ne sont pas optimaux et peuvent être revus à la lumière des méthodes d'analyse fonctionnelle, qui permettent alors de trouver un cadre commun pour les deux numérations –binaire et fractions continues. Ces mêmes méthodes permettent aussi sans doute de débuter l'étude dans le cas beaucoup plus difficile où les données ne sont plus indépendantes.

RÉFÉRENCES

[1] DAUDÉ, H., FLAJOLET, P., AND VALLÉE, B. An analysis of the Gaussian algorithm for lattice reduction. In *Algorithmic Number Theory Symposium* (1994), L. Adleman, Ed., no. 877 in Lecture Notes in Computer Science, pp. 144–158. Proceedings of *ANTS'94*.

[2] DEVROYE, L. Ma probabilistic analysis of the height of tries and the complexity of trie-sort. *Acta Arithmetica 21* (1974), 229–237.

[3] ELLISON, W., AND MENDÈS FRANCE, M. *Les nombre premiers*. Publications de l'institut de mathématique de l'université de NANCANGO, IX. Editions Hermann, 1975.

[4] FAIVRE, C. Distribution of Levy's constants for quadratic numbers. *Acta Arithmetica 61.1* (1992), 13–34.

[5] FLAJOLET, P., GOURDON, X., AND DUMAS, P. Mellin transforms and asymptotics : Harmonic sums. *Theoretical Computer Science 144*, 1–2 (June 1995), 3–58.

[6] Flajolet, P., and Vallée, B. Continued fraction algorithms, functional operators, and structure constants. Research Report 2931, Institut National de Recherche en Informatique et en Automatique, July 1996. 33 pages. (Invited lecture at the 7th Fibonacci Conference, Graz, July 1996.).

[7] Hensley, D. The number of steps in the euclidean algorithm. *Journal of Number Theory 49*, 2 (november 1994), 142–182.

[8] Lévy, P. Sur la loi de probabilité dont dépendent les quotients complets et incomplets d'une fraction continue. *Bull. Soc. Math. France 557* (1929), 178–194.

[9] Vallée, B. Algorithms for computing signs of 2×2 determinants : dynamics and average–case analysis. Les cahiers du GREYC, Université de Caen, 1997. 12p.

[10] Wirsing, E. On the theorem of Gauss–Kusmin–Lévy and a Frobenius–type theorem for function spaces. *Acta Arithmetica 24* (1974), 507–528.