Théorie de l'information :
Modèles, Algorithmes, Analyse

Notes du Mini-Cours donné aux Journées 2010 du GT ALEA

Brigitte VALLÉE
(GREYC, CNRS et Université de Caen)

Tout étudiant d'un cours d'algorithmique de base apprend que la complexité moyenne de l'algorithme QuickSort est en $O(n \log n)$, celle de QuickSelect est en $O(n)$ et celle de RadixSort est en $O(n \log n)$. De tels énoncés ont le mérite d'être simples, mais leur simplicité est trompeuse, car ils sont fondés sur des hypothèses spécifiques à chaque algorithme: pour les deux premiers algorithmes, le coût unitaire est la comparaison entre clés, tandis que, pour le troisième, le coût unitaire est la comparaison entre symboles.

Ces études souffrent donc de deux inconvénients majeurs: il n'est pas possible de comparer réellement ces algorithmes entre eux, car les mesures de coût sont différentes. Ensuite, la mesure de coût adoptée pour analyser QuickSort ou QuickSelect est peu réaliste, dès que les clés ont une structure complexe, ce qui est le cas dans le contexte des bases de données ou de la langue naturelle, par exemple.

Pour effectuer une analyse réaliste, il faut donc d'abord travailler en théorie de l'information pour définir un cadre adapté. En théorie de l'information, une source est un mécanisme aléatoire qui produit des symboles d'un alphabet donné. On construit ici un modèle de source très général, qui peut prendre en compte des corrélations importantes entre symboles émis. Les clés considérées par l'algorithme sont alors des mots produits (indépendamment) par la même source.

Il faut ensuite considérer un coût unitaire qui soit le même pour tous les algorithmes: c'est la comparaison entre symboles, et le coût de l'algorithme est donc le nombre total de comparaisons effectuées entre symboles.

Nous revisitons ainsi, dans un tel modèle, à la fois unifié et réaliste, l'analyse probabiliste de trois principaux algorithmes: QuickSort, QuickSelect, et les algorithmes de dictionnaire fondés sur la structure de trie.

Ce mini-cours est fondé sur des travaux communs avec Julien CL ÉMENT, James FILL, et Philippe FLAJOLET et, essentiellement sur les deux articles suivants, cités dans la bibliographie en [7] et [40]:

Julien Clément, Philippe Flajolet, et Brigitte Vallée, *Dynamical sources in Information Theory: Analysis of general tries*, Algorithmica (2001), vol 29 (1/2) pp 307–369

Brigitte Vallée, Julien Clément, James Fill, et Philippe Flajolet, *The number of symbol comparisons in QuickSort and QuickSelect*, Proceedings of ICALP 09, LNCS 5555, pages 750–763, 2009

Les notes qui suivent reprennent ainsi certains passages de ces deux articles, en uniformisant les notations.

Sources. In information theory contexts, the two simpler models of sources are memoryless sources, where symbols in words are each emitted independently of the previous ones, and Markov chains, where the probability of emitting a symbol depends solely on a bounded part of the past history.

However, as opposed to hashing, data on which tries are built or data which are sorted with `QuickSort` often arise from more realistic sources that are often more complex. We work here inside a quite general framework of sources, where a central idea is the modelling of the source via its *fundamental probabilities*, namely the probabilities that a word of the source begins with a given prefix. What we call a probabilistic *source* produces infinite words on the ordered alphabet Σ. The set of *keys* (words, data items) is then Σ^∞, endowed with the strict lexicographic order.
Tries. Digital trees, usually called tries, are both an abstract structure and a data structure that can be superimposed on a set of words produced by some source. We consider here a quite general situation, since we study *general* tries built on a set of words produced by a *general* source.

As an abstract structure, tries are based on a splitting according to symbols encountered in words: if $\mathcal{X}$ is a set of words, and $\Sigma = \{a_1, a_2, \ldots, a_r\}$ is the alphabet, then the trie associated to $\mathcal{X}$ is defined recursively by the rule:

$$\mathtt{Trie}(\mathcal{X}) = \langle \mathtt{Trie}(\mathcal{X} \setminus a_1), \ldots, \mathtt{Trie}(\mathcal{X} \setminus a_r) \rangle,$$

where $\mathcal{X} \setminus \alpha$ means the subset of $\mathcal{X}$ consisting of strings that start with α, stripped of their initial symbol α; recursion is halted as soon as $\mathcal{X}$ contains less than two elements. The advantage of the trie is that it only maintains the minimal prefix set of characters that is necessary to distinguish all the elements of $\mathcal{X}$.

Clearly the tree $\mathtt{Trie}(\mathcal{X})$ supports the search for any word X in the set $\mathcal{X}$ by following an access path dictated by the successive symbols of X. By similar means, the trie implements insertions and deletions, so that it is a fully dynamic dictionary data type. In addition, tries efficiently support set-theoretic operations like union and intersection, as well as partial match queries or interval search, and suitable adaptations make them a method of choice for complex text processing tasks. These various applications justify considering the trie structure as one of the central general-purpose data structures of computer science.

When it comes to implementation, several options are possible depending on the decision structure chosen to guide descent in each node to subtrees. Three major choices present themselves. The “array-trie” uses an *array* of pointers to access subtrees directly whereas the “list-trie” relies upon *linked lists* traversal. The “bst-trie” uses *binary search trees* (bst) as subtree access method. Each of these structures is then an hybridation between the trie structure and the node structure and it is called an “hybrid trie”.

Our motivation in considering hybrid trie structures comes in fact from a paper of Bentley and Sedgewick, who, following early ideas of Clampett, developed an elegant implementation of bst-tries, under the name of *ternary search trie*, or *tst* for short. The basic idea of these authors is to represent the bst-trie as a ternary tree where search on symbols is conducted like in a standard binary search tree over the alphabet set Σ, while trie descent is performed by following an escape pointer whenever equality of symbols of detected. In this way, the code is especially compact and, in simulations, the implementation constants appear to be particularly small. Bentley and Sedgewick report that, in practical situations, their data structure can be more efficient than hashing while offering considerably wider functionality. Our goal, as analysts, is to examine this claim and precisely quantify what goes on.

`QuickSort` and `QuickSelect`. We revisit the classical `QuickSort` and `QuickSelect` algorithms under a complexity model that fully takes into account the elementary comparisons between symbols composing the records to be processed.

Every student of a basic algorithms course is taught that, on average, the complexity of Quicksort [or the path-length of a Binary Search Tree (bst)] is $O(n \log n)$, and that of `QuickSelect` is $O(n)$. Such statements are based on specific assumptions—that the comparison of data items (for the first two) and the comparison of symbols (for the third one) have unit cost—and they have the obvious merit of offering an easy-to-grasp picture of the complexity landscape. However, as noted by Sedgewick, these simplifying assumptions suffer from limitations: they do not make possible a precise assessment of the relative merits of algorithms and data structures that resort to different methods (e.g., comparison-based versus radix-based sorting) in a way that would satisfy the requirements of either information theory or algorithms engineering. Indeed, computation is not reduced to its simplest terms, namely, the manipulation of totally elementary symbols, such as bits, bytes, characters. Furthermore, such simplified analyses say little about a great many application contexts, in databases or natural language processing, for instance, where information is highly "non-atomic", in the sense that it does not plainly reduce to a single machine word.

First, we observe that, for commonly used data models, the mean costs $S(n)$ and $K(n)$ of *any* algorithm under the symbol-comparison and the key-comparison model, respectively, are connected by the universal relation $S(n) = K(n) \cdot O(\log n)$. (This results from the fact that at most $O(\log n)$ symbols suffice, with high probability, to distinguish n keys; cf. the analysis of the height of tries The surprise is that there are cases where this upper bound is tight, as in `QuickSort`; others where both costs are of the same order, as in `QuickSelect`. In this work, we show that the expected cost of `QuickSort` is $O(n \log^2 n)$, *not* $O(n \log n)$, when *all* elementary operations—symbol comparisons—are taken into account. By contrast, the cost of `QuickSelect` turns out to be $O(n)$, in both the old and the new world, albeit, of course, with different implied constants.

Results. We consider a (finite) ordered *alphabet* Σ. Our main objects of study are the `QuickSort` and the `QuickSelect` algorithms, or the main parameters of a `Trie`, when the n keys are assumed to be *independently drawn* from the same source $\mathcal{S}$. Provided that the source $\mathcal{S}$ be tame (we will give later formal definitions of this notion), the following results are established for `Tries` `QuickSort`, (or `Bst` and `QuickSelect`. More precisely, we study the algorithm `QuickSelect` when it deals with n keys and searches the key of rank $\lfloor \alpha n \rfloor$: such an algorithm searches the α-quantile, and is called $\texttt{QuickQuant}_\alpha(n)$.

(*i*) Up to possible small fluctuations which occur in the case of a Λ–periodic source, the average size of the trie is well approximated by a quantity of order n

$$R(n) \sim \frac{1}{h(S)}\, n.$$

(*ii*) The average path length of the trie depends on the hybrid structure used (C for array-trie, L for list-trie, A for bst-trie). For any source built on a finite alphabet, all the average path lengths are of order $n \log n$,

$$C(n) \sim \frac{1}{h(S)} n \log n, \quad L(n) \sim \frac{K_L(S)}{h(S)} n \log n, \quad A(n) \sim \frac{K_A(S)}{h(S)} n \log n,$$

with explicit constants $K_L(S), K_A(S)$. However, for an infinite (denumerable) alphabet, the array-trie does not exist anymore, and the two average path lengths, relative to list-tries or bst-tries, can be of different order.

(*iii*) The mean number $B(n)$ of symbol comparisons of $\texttt{QuickSort}(n)$ – or the symbol-path-length of the BST built on n keys – involves the entropy $h(\mathcal{S})$ of the source:

$$B(n) \sim \frac{1}{h(\mathcal{S})} n \log^2 n.$$

(*iv*) The mean number of symbol comparisons $Q^{(\alpha)}(n)$ performed by the algorithm $\texttt{QuickQuant}_\alpha(n)$ satisfies

$$Q^{(\alpha)}(n) \sim \rho_{\mathcal{S}}(\alpha) n$$

The mean number of symbol comparisons, $M^{(-)}(n)$ for $\texttt{QuickMin}(n)$ and $M^{(+)}(n)$ for $\texttt{QuickMax}(n)$, satisfies with $\epsilon = \pm$,

$$M^{(\epsilon)}(n) = \rho_{\mathcal{S}}^{(\epsilon)} n, \qquad \text{with} \quad \rho_{\mathcal{S}}^{(+)} = \rho_{\mathcal{S}}(1), \qquad \rho_{\mathcal{S}}^{(-)} = \rho_{\mathcal{S}}(0).$$

The mean number $M(n)$ of symbol comparisons performed by $\texttt{QuickRand}(n)$ satisfies

$$M(n) = \gamma_{\mathcal{S}}\, n, \qquad \text{with} \quad \gamma_{\mathcal{S}} = \int_0^1 \rho(\alpha) d\alpha.$$

The first three results involve the entropy $h(\mathcal{S})$ of the source which always exists for a tame source. For a periodic source, a term $nP(\log n)$ is to be added to estimates of $R(n), C(n)$ and $B(n)$, where $P(u)$ is a (computable) continuous periodic function. This term adds to the dominant term for $R(n)$ and to subdominant terms for $C(n)$ and $B(n)$.

The results (*iii*) and (*iv*) about QuickSort andQuickSelect constitute broad extensions of earlier ones by Fill and Janson and Fill and Nakama, whose analysis is relative to data composed of random uniform bits. The results on Tries are the first results which are obtained on a general source, and take into account realistic implementations of a trie. However, the analysis of the main parameters of a trie is quite long. See the bibliography .....

Methods. We operate under a general model of source, parametrized by the unit interval $\mathcal{I}$. Our strategy comprises three main steps. The first two are essentially *algebraic*, while the last one relies on complex *analysis*.

Step (*a*)**.** We first work with the *P*oisson model where the number of keys, instead of being fixed, follows a Poisson law of parameter Z. We obtain expressions for the mean costs of interest which involve the fundamental probabilities of the source.

Step (*b*)**.** Then, simple algebra yield expressions relative to the model where the number n of keys is fixed. The exact representations of the mean costs is an alternating sum which involves two kinds of quantities, the size n of the set of data to be analyzed (which tends to infinity), and the fundamental probabilities (which tend to 0).

Step (c)**.** We approach the corresponding asymptotic analysis by means of *complex integral representations* of the Nörlund–Rice type. For each algorithm–source pair (or data structure-source), a series of Dirichlet type encapsulates both the properties of the source and the characteristics of the algorithm—this is the *mixed Dirichlet series*, denoted by $\varpi(s)$, whose singularity structure in the complex plane is proved to condition our final asymptotic estimates.

Plan of the paper. Section 1 describes the general framework of sources, and defines fundamental probabilities. It also focuses on a subclass of sources, related to dynamical systems of the interval, the class of dynamical sources. This class encompasses the simple sources (memoryless sources, and Markov chains), but it may also possess a high degree of correlation between symbols. Section 2 is devoted to the description of the various hybrid tries, and describes their algebraic analysis. It is shown that all of the analyses of various parameters involve the Dirichlet series of fundamental probabilities $\Lambda(s)$, or some of its close variants. Section 3 performs the two first steps of the analysis of the mean number of symbol comparisons of `QuickSort` and a dual algorithm of $\texttt{QuickQuant}_\alpha(n)$. These analyses involve $\Lambda(s)$ (for `QuickSort`) and another Dirichlet series (for `QuickQuant`). Section 4 provides a description of possible analytic behaviours of a source, which will be used in the final step of the analysis, the analytic step, performed in Section 5.

1. A general source model.

Throughout this paper, a totally ordered (finite) alphabet $\Sigma := \{a_1, a_2, \dots, a_r\}$ of "symbols" or "letters" is fixed.

1.1. The general model.

We first describe a general source.

Definition 1. *A* probabilistic source*, which produces infinite words of* Σ^∞*, is specified by the set* $\{p_w, w \in \Sigma^\star\}$ *of* fundamental probabilities p_w*, where* p_w *is the probability that an infinite word begins with the finite prefix* w*. It is furthermore assumed that* $\pi_k := \sup\{p_w : w \in \Sigma^k\}$ *tends to* 0*, as* $k \to \infty$*.*

For any prefix $w \in \Sigma^\star$, we denote by $|w|$ the length of w (i.e., the number of the symbols that it contains) and $p_w^{(-)}$, $p_w^{(+)}$, p_w the probabilities that a word produced by the source begins with a prefix α of the same length as w, which satisfies $\alpha < w$, $\alpha > w$, or $\alpha = w$, respectively:

$$p_w^{(-)} := \sum_{\substack{\alpha, |\alpha] = |w|, \\ \alpha < w}} p_\alpha, \qquad p_w^{(+)} := \sum_{\substack{\alpha, |\alpha] = |w|, \\ \alpha > w}} p_\alpha$$

Since the sum of these three probabilities equals 1, this defines two real numbers $b_w, c_w \in [0, 1]$ for which

$$b_w = p_w^{(-)}, \quad 1 - c_w = p_w^{(+)}, \quad c_w - b_w = p_w,$$

as it is explained in Figure 1. Denote by $\mathcal{L}(\mathcal{S})$ the set of (infinite) words produced by the source $\mathcal{S}$. Given an infinite word $X \in \mathcal{L}(\mathcal{S})$, denote by w_k its prefix of length k. The sequence (b_{w_k}) is increasing, the sequence (c_{w_k}) is decreasing, and $c_{w_k} - b_{w_k} = p_{w_k}$ tends to 0. Thus a unique real $P(X) \in [0, 1]$ is defined as common

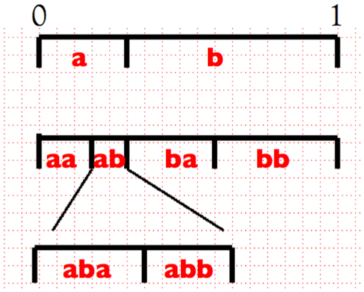


FIGURE 1. The parametrization of a source.

limit of (b_{w_k}) and (c_{w_k}), and $P(X)$ can be viewed as the probability that an infinite word Y be smaller than X.The mapping $P : \mathcal{L}(\mathcal{S}) \to [0,1]$ is strictly increasing outside the exceptional set formed with words which end with an infinite sequence of the smallest letter a_1 or with an infinite sequence of the largest letter a_r.

Conversely, almost everywhere, except on the set $\{b_w, w \in \Sigma^\star\}$, a mapping $M : [0,1] \to \mathcal{L}(\mathcal{S})$ associates, to a number u of the interval $\mathcal{I} := [0,1]$, a word

$$M(u) := (m_1(u),\ m_2(u),\ m_3(u), \ldots) \in \mathcal{L}(\mathcal{S}).$$

In this way, the probabilty that a word Y will be smaller than $M(u)$ equals u. The lexicographic order on words is compatible with the natural order on the interval $\mathcal{I}$. The interval $\mathcal{I}_w := [b_w, c_w]$ gathers (up to a denumerable set of points) all the reals u for which $M(u)$ begins with the prefix w. Its length equals p_w. This is the fundamental interval of the prefix w.

Our analyses involve the Dirichlet series of the source, defined as

$$(1) \qquad \Lambda(s) := \sum_{w \in \Sigma^\star} p_w^s, \qquad \Lambda^{(k)}(s) := \sum_{w \in \Sigma^k} p_w^s. \qquad \Pi(s) := \sum_{k \geq 0} \pi_k^s;$$

Since the equalities $\Lambda^{(k)}(1) = 1$ hold, the series $\Lambda(s)$ is divergent at $s = 1$, and the probabilistic properties of the source can be expressed in terms of the regularity of Λ near $s = 1$, as we will see it later.

For instance, the entropy $h(\mathcal{S})$ relative to a probabilistic source $\mathcal{S}$ is defined as the limit (if it exists) of a quantity that involve the fundamental probabilities

$$(2) \qquad h(\mathcal{S}) := \lim_{k\to\infty} \frac{-1}{k} \sum_{w \in \Sigma^k} p_w \log p_w = -\lim_{k\to\infty} \frac{-1}{k} \frac{d}{ds} \Lambda^{(k)}(s)_{|s=1}$$

1.2. **Simple sources: memoryless sources and Markov chains.** A memoryless source associated to the alphabet Σ (possibly infinite), is defined by the set $(p_j)_{j\in\Sigma}$ of probabilities, and the Dirichlet series $\Lambda, \Lambda^{(k)}$ are expressed with

$$(3) \quad \lambda(s) = \sum_{i \in \Sigma} p_i^s, \qquad \text{under the form} \qquad \Lambda^{(k)}(s) = \lambda(s)^k, \qquad \Lambda(s) = \frac{1}{1 - \lambda(s)}.$$

In this case, the entropy equals $h(\mathcal{S}) = -\sum_i p_i \log p_i = -\lambda'(1)$.

A Markov chain associated to the finite alphabet Σ, is defined by the vector R of initial probabilities $(r_i)_{i\in\Sigma}$ together with the transition matrix $P := [(p_{i|j})_{(i,j)\in\Sigma\times\Sigma}]$, whose each column has a sum equal to 1. We denote by $P(s)$ the matrix with general coefficient $p^s_{i|j}$, and by $R(s)$ the vector of components r^s_i. Then

$$(4) \qquad \Lambda^{(k)}(s) = {}^t\mathbf{1} \cdot P(s)^{k-1} \cdot R(s), \qquad \Lambda(s) = 1 + {}^t\mathbf{1} \cdot (I - P(s))^{-1} \cdot R(s).$$

If, moreover, the matrix P is irreducible and aperiodic, then, for any real s, the matrix $P(s)$ has a unique dominant eigenvalue $\lambda(s)$. For $s = 1$, the matrix $P = P(1)$ has a unique fixed vector with positive components π_i, whose sum equals 1. The entropy of the source is then equal to

$$h(\mathcal{S}) = -\lambda'(1) = -\sum_{(i,j)\in\Sigma^2} \pi_j \, p_{i|j} \, \log p_{i|j}.$$

1.3. **Dynamical sources.** An important subclass of sources is formed by *dynamical sources*, which are closely related to dynamical systems on the interval.

Definition 2. [Dynamical System of of the interval] *A dynamical system of the interval $\mathcal{I} := [0,1]$ is defined by a mapping $T : \mathcal{I} \to \mathcal{I}$ (called the shift) for which*

(a) there exists a finite alphabet Σ, and a topological partition of $\mathcal{I}$ with disjoint open intervals $\mathcal{I}_m$, $m \in \Sigma$, i.e. $\bar{\mathcal{I}} = \cup_{m\in\Sigma}\bar{\mathcal{I}}_m$.

(b) The restriction of T to each $\mathcal{I}_m$ is a $\mathcal{C}^2$ bijection from $\mathcal{I}_m$ to $T(\mathcal{I}_m)$.

The system is complete when each restriction is surjective, i.e., $T(\mathcal{I}_m) = \mathcal{I}$,

The system is Markovian when each interval $T(\mathcal{I}_m)$ is a union of intervals $\mathcal{I}_j$, of the form $T(\mathcal{I}_m) = \bigcup_{j\in\mathcal{K}_m} \mathcal{I}_j$.

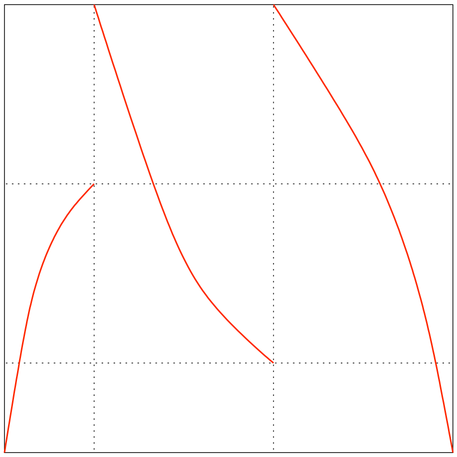


FIGURE 2. An instance of a Markovian source

Figure 2 shows an instance of a Markovian source.

A dynamical system, together with a distribution G on the unit interval $\mathcal{I}$, defines a probabilistic source, which is called a dynamical source and is now described (See also Figure 3). The map T is used as a shift mapping, and the mapping τ whose restriction to each $\mathcal{I}_m$ is equal to m, is used for coding. The words are emitted as follows: To each real x, (except for a denumerable set), one associates the trajectory $\mathcal{T}(x) = (x, T(x), T^2(x), \dots T^j(x), \dots)$, which gives rise, via the mapping τ to the word $M(x) \in \Sigma^\infty$ formed with the symbols

$$M(x) = (m_1(x), m_2(x), \dots, m_n(x), \dots) \qquad \text{with} \quad m_j(x) = \tau(T^{j-1}(x)).$$

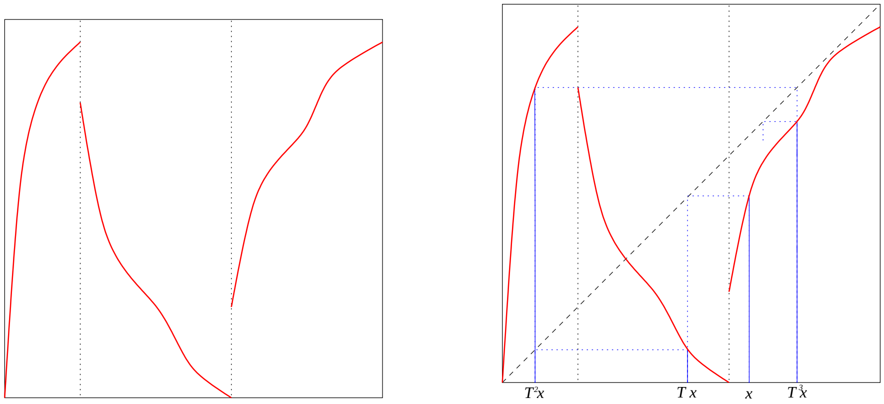


FIGURE 3. A dynamical system, with $\Sigma = \{a, b, c\}$ and a word $M(x) = (c, b, a, c \dots)$.

Given a prefix $w \in \Sigma^\star$, the set $\mathcal{I}_w$ of all reals x for which the word $M(x)$ begins with the prefix w is an interval, of the form $[b_w, c_w]$, the fundamental interval associated to w, and the measure of this interval (with respect to distribution G), is the fundamental probability p_w of the source, here equal to

$$p_w := G(c_w) - G(b_w).$$

In the case of a complete system, one denotes by $h_{[m]}$ the local inverse of T restricted to $\mathcal{I}_m$ and by $\mathcal{H}$ the set $\mathcal{H} := \{h_{[m]}, m \in \Sigma\}$ of all local inverses. All the local inverses of the k–th iterate T^k are then associated to a word $w = m_1 m_2 \dots m_k \in \Sigma^k$ are of the form $h_{[w]} := h_{[m_1]} \circ h_{[m_2]} \dots h_{[m_k]}$. Then, the set of all the inverse branches of T^k is $\mathcal{H}^k = \{h_{[w]};\quad w \in \Sigma^k\}$. Then, each fundamental interval $\mathcal{I}_w$ is just $\mathcal{I}_w = h_{[w]}(\mathcal{I})$ and the fundamental probability is just

$$p_w = |G(h_{[w]}(1)) - G(h_{[w]}(0))|. \tag{5}$$

For $h \in \mathcal{H}^k$, the number k is called the *depth* of h and it is denoted by $|h|$. We denote by $\mathcal{H}^\star := \cup_{k \geq 0} \mathcal{H}^k$ the set of inverse branches of any depth.

Such sources may possess a high degree of correlations, due to the *geometry* of the branches and also to the *shape* of branches.

The geometry of the branches is defined by the respective positions of "horizontal" intervals $\mathcal{I}_m$ with respect to "vertical" intervals $\mathcal{J}_\ell := T(\mathcal{I}_\ell)$ and allows to describe the set $\mathcal{S}_m$ formed with symbols which can be possibly emitted after symbol m. The geometry of the system then provides a first access to the correlation between successive symbols.

In a *complete* system, any symbol of Σ can be emitted after any symbol m, and thus the equality $\mathcal{S}_m = \Sigma$ always holds. In a *markovian* system, the set $\mathcal{S}_m$ equals $\mathcal{K}_m$, defined in Definition 2. In the case when the system is not Markovian, it is sometimes possible to obtain a refinement of the partition, for which the new system becomes Markovian. But, this is not always possible, and, in the case when it is not possible, the set S_m cannot be characterized when considering only bounded parts of the previous history. In all the cases, the following property [always true for a complete system] is essential:

[Topologically mixing] *For any pair of symbols (b, e), there exists $n_0 \geq 1$ such that, for any $n \geq n_0$, there is a word of length n which begins with symbol b and finishes with symbol e [i.e., $\mathcal{I}_b \cap T^{-n}(\mathcal{I}_e) \neq \emptyset$.*

The shape of the branches, and more precisely, the behavior of derivatives h'_m has also a great influence on correlations between symbols. For a fixed geometry of the branches, a system with affine branches is "less correlated" than the other systems with the same geometry.

1.4. **Simple sources seen as dynamical sources.** All memoryless sources and all Markov chain sources belong to the general framework of dynamical sources and correspond to a piecewise linear shift, under this angle of dynamical sources. For instance, the standard binary system is obtained by $T(x) = \{2x\}$ ($\{\cdot\}$ is the fractional part). Here, Figure 4 provides a representation of two memoryless sources and Markov chains. More precisely

– A complete dynamical source, with affine branches and a uniform initial distribution, defines a memoryless source.

– A Markovian dynamical source, with affine branches and a family of uniform initial distributions on each $\mathcal{J}_j$, defines a Markov chain.

Figure 4 shows two instances of memoryless sources and one instance of a Markov chain, viewed as dynamical sources.

However, as soon as the derivatives h' of the branches are not constant, there exist correlations between successive symbols, and the dynamical source is no longer simple. Dynamical sources with a non-linear shift allow for correlations that depend on the entire past. A main instance is the dynamical source relative to the Gauss map, which underlyes the Euclid Algorithm and is defined on the unit interval via the shift T

$$T(0) = 0, \qquad T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor \quad (x \neq 0). \tag{6}$$

The graph of this dynamical system is represented in Figure 4.

1.5. **Transfer operators.** One of the main tools in dynamical system theory is the transfer operator introduced by Ruelle, denoted by H_s. It generalizes the density transformer H that describes the evolution of the density. Here, as in [38], we describe a generalized version of the transfer operator –the secant operator– which gives rise to an expression of the Dirichlet series $\Lambda(s)$ as a quasi–inverse, in a way that generalises expressions obtained in (3) or in (4).

We first consider the case of a complete dynamical system: if $f = f_0$ denotes the initial density on $\mathcal{I}$, and f_1 the density on $\mathcal{I}$ after one iteration of T, then f_1 can be written as $f_1 = H[f_0]$, where H is defined by

$$H := \sum_{h \in \mathcal{H}} H_{(h)} \qquad \text{with} \quad H_{(h)}[f](x) := |h'(x)|\, f \circ h(x).$$

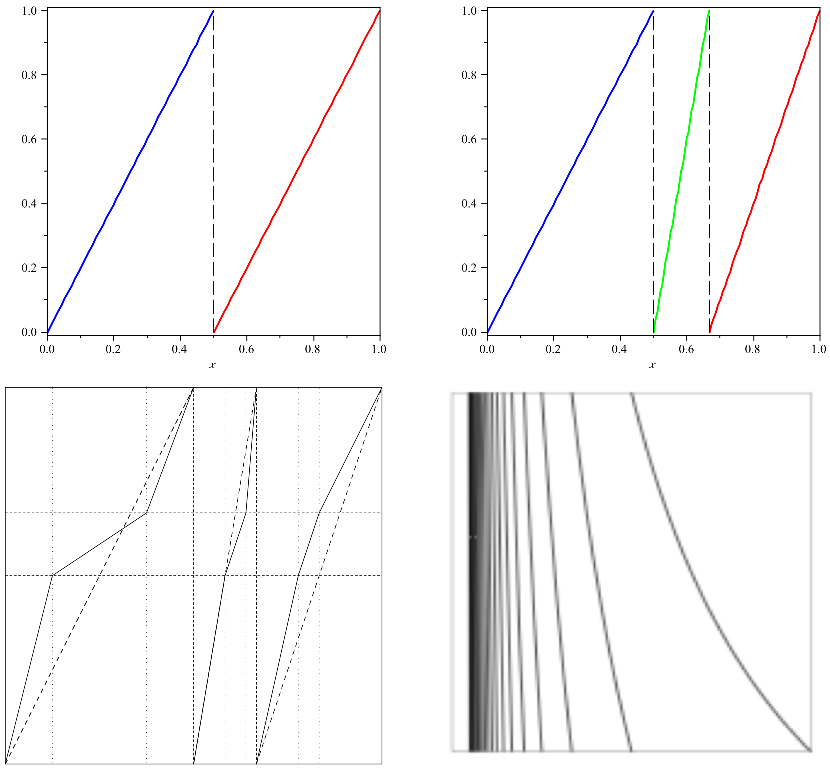


FIGURE 4. Two memoryless sources and a Markov chain, viewed as dynamical sources. The Continued fraction source.

The transfer operator extends the density transformer; it depends on a complex parameter s, and coincides with H when $s = 1$,

$$H_s = \sum_{h\in\mathcal{H}} H_{(h),s} \qquad \text{with} \quad H_{(h),s}[f](x) := |h'(x)|^s \cdot f \circ h(x). \tag{7}$$

Multiplicative properties of derivatives then proves that the k–th iterate of the transfer operator involves the set $\mathcal{H}^k$ under the form

$$H_s^k = \sum_{h\in\mathcal{H}^k} H_{(h),s}.$$

In the case of a Markovian system, one denotes by $\mathcal{H}_{[k|\ell]}$ the set of inverse branches of the shift T for which $h(\mathcal{I}_k) \subset \mathcal{I}_\ell$, and by $H_{[k|\ell],s}$ the operator defined as

$$H_{[k|\ell],s} := \sum_{h\in\mathcal{H}_{[k|\ell]}} H_{(h),s}\,.$$

The transfer operator H_s can be viewed as the matrix of operators

$$H_s = \left(H_{[k|\ell],s}\right)_{(k,\ell)\in\mathcal{L}^2}\,.$$

Here, we are interested by the fundamental probabilities, whose expression is provided in (5) in the case of a complete dynamical system. We now introduce the main tool for generating these probabilities, namely, the secant transfer operator. This operator involves the secant function of inverse branches (instead of their derivatives), it acts on functions F of two variables; for $s \in \mathbb{C}$, and $h \in \mathcal{H}$, we first define the component secant operator $\mathbf{H}_{(h),s}$ as

$$\mathbf{H}_{(h),s}[F](x,y) := \left|\frac{h(x)-h(y)}{x-y}\right|^s F(h(x),h(y)). \tag{8}$$

Then, the transfer operator of a complete system is defined as

$$\mathbf{H}_s := \sum_{h\in\mathcal{H}} \mathbf{H}_{(h),s}, \tag{9}$$

and, for a Markovian system, as a matrix of operators,

$$\mathbf{H}_s = \left(\mathbf{H}_{[k|\ell],s}\right)_{(k,\ell)\in\mathcal{L}^2}, \qquad \text{with} \quad \mathbf{H}_{[k|\ell],s} := \sum_{h\in\mathcal{H}_{[k|\ell]}} \mathbf{H}_{(h),s}. \tag{10}$$

The secant operator is then an extension of the plain transfer operator, since, on the diagonal $x = y$, one has

$$\mathbf{H}_s[F](x,x) = H_s[\operatorname{diag} F](x), \tag{11}$$

where the function $\operatorname{diag} F$ is defined by $\operatorname{diag} F(x) := F(x,x)$. In the complete case, and in the same vein as for tangent operators, multiplicative properties of secants then entail the relation

$$\mathbf{H}_s^k = \sum_{h\in\mathcal{H}^k} \mathbf{H}_{(h),s} \qquad \text{so that} \quad \mathbf{H}_s^k[F](x,y) = \sum_{h\in\mathcal{H}^k} \left|\frac{h(x)-h(y)}{x-y}\right|^s F(h(x),h(y)).$$

Then, for $w \in \Sigma^k$, the probability p_w^s can be written as a function of the component $\mathbf{H}_{(h_{[w]}),s}$ of the operator $\mathbf{H}_s^k$ relative to the inverse branch $h_{[w]}$ under the form

$$p_w^s = |G(h_{[w]}(1)) - G(h_{[w]}(0))|^s = \left|\frac{h_{[w]}(1)-h_{[w]}(0)}{1-0}\right|^s \cdot \left|\frac{G(h_{[w]}(1))-G(h_{[w]}(0))}{h_{[w]}(1)-h_{[w]}(0)}\right|^s.$$

Then, if L is the secant of the distribution G, defined by

$$L(x,y) := \frac{G(x)-G(y)}{x-y}, \tag{12}$$

then the series $\Lambda_k(s)$ and $\Lambda(s)$ are expressed as follows:

$$\Lambda_k(s) := \sum_{w\in\Sigma^k} p_w^s = \mathbf{H}_s^k[L^s](1,0), \qquad \Lambda(s) = (1-\mathbf{H}_s)^{-1}[L^s](1,0).$$

The formula extends to the Markovian case, when one replaces (12) by

$$L = (L_\ell)_{\ell\in\mathcal{L}}, \qquad \text{with} \quad L_\ell(x,y) := \frac{G_\ell(x)-G_\ell(y)}{x-y}, \tag{13}$$

where G_ℓ is the initial distribution on the interval $\mathcal{I}_\ell$. Finally, we have proven:

Proposition 1. *For a complete or a markovian dynamical source, relative to a shift T and a distribution G, the Dirichlet series of the source admits an alternative expression which involves the quasi–inverse of the secant operator, defined in (9) (for complete case), and in (10) (for markovian case), applied to the function L^s, where L is the secant of the distribution G, described in (12). More precisely, one has*

$$\Lambda^{(k)}(s) = \mathbf{H}_s^k[L^s](0,1), \qquad \Lambda(s) = (I-\mathbf{H}_s)^{-1}[L^s](0,1).$$

1.6. **Average-case analysis: various models.** The purpose of average–case analysis of structures (or algorithms) is to characterize the mean value of their parameters under a well-defined probabilistic model that describes the initial distribution of its inputs. Here, we adopt the following quite general model: we work with a finite sequence $\mathcal{X}$ of infinite words independently produced by the same source, of cardinality n. Such a sequence $\mathcal{X} := (X_1, X_2, \ldots, X_n)$ is obtained by n independent drawings $x_1, x_2, \ldots, x_n$ in the interval $\mathcal{I}$. We then set $X_i := M(x_i)$. This model is called the Bernoulli model and is denoted by $(\mathcal{B}_n, S)$ when it is relative to cardinality n and probabilistic source S.

Rather than fixing the cardinality n of the sequence $\mathcal{X}$, it proves technically convenient to consider that the sequence $\mathcal{X}$ has a variable number N of elements that obeys a Poisson law of parameter z,

$$\Pr\{N = k\} = e^{-z}\frac{z^k}{k!}. \tag{14}$$

In this model, N is narrowly concentrated near its mean z with a high probability so that the rate Z plays a rôle much similar to cardinality of $\mathcal{X}$. This model is called the Poisson model of rate z. When it is relative to probabilistic source source S, it is denoted by $(\mathcal{P}_z, S)$ and is composed with two main steps:

– The number N of words is drawn according to the Poisson law

– Then, the N words are independently drawn from the source $\mathcal{S}$, i.e., there are N real numbers x_i that are uniformly and independently drawn in the unit interval, and the words X_i are chosen as $X_i := M(x_i)$.

The interest of the Poisson model is that there is complete independence on what happens in disjoint subintervals of $\mathcal{I}$. Moreover, the number of elements that fall into any interval of measure p is itself distributed as a Poisson variable of rate zp. More precisely:

Lemma 1. *In the Poisson model $(\mathcal{P}_z, S)$, denote by $N_{[b,c]}$ the number of words $M(x)$ of the source $\mathcal{S}$ whose parameter x belongs to the interval $[b, c]$. Then:*

(i) $N_{[b,c]}$ follows a Poisson law of parameter $z(c - b)$.

(ii) For $[b, c] \cap [b', c'] = \emptyset$ the variables $N_{[b,c]}$ and $N_{[b',c']}$ are independent.

In particular, when the total number N of words drawn from a source follows a Poisson law $\mathcal{P}_z$, the number N_w of words which begin with the prefix w follows a Poisson law of parameter zp_w, where p_w is the fundamental probability of prefix w, and the two variables N_w and $N_{w'}$, relative to two prefixes w and w' which have no common prefix, are independent.

These two properties will give an easy access to the expectation of basic parameters in the model $(\mathcal{P}_z, S)$. It is then possible to go back from the Poisson model $\mathcal{P}_z$ of rate z to a Bernoulli model $\mathcal{B}_n$ where n is fixed. The following formula

$$\mathbb{E}[\gamma, \mathcal{P}_z, S] = e^{-z}\sum_{n=0}^{\infty}\frac{z^n}{n!}\mathbb{E}[\gamma, \mathcal{B}_n, S], \qquad \mathbb{E}[\gamma, \mathcal{B}_n, S] = n![z^n]\ e^z\mathbb{E}[\gamma, \mathcal{P}_z, S] \tag{15}$$

induces a formal dictionary

$$e^{-az} \mapsto (1-a)^n, \quad ze^{-az} \mapsto n(1-a)^{n-1}.$$

2. General tries

We first describe the trie and hybrid trie data structures along with their parameters. We show that the expectations of studied parameters (size and various path lengths) can be expressed as sums that involve the fundamental probabilities.

2.1. **Definition of tries.**

We deal with two basic mappings σ and T, defined as follows: For any (infinite) word X produced by the source, $\sigma(X)$ is the first symbol of X, and $T(X)$ is the suffix of X, i.e., the word X stripped from its first symbol. The (partial) map $T_{[m]}$ is a refinement of the map T: it is only defined on words X which begin with the symbol m (i.e., for which $\sigma(X) = m$) and, in this case, one has $T_{[m]}(X) := T(X)$.

The reader may think that there already exists a mapping T, namely the shift that defines a dynamical source. However, we explicitly choose to use the same name for the suffix map T (defined on words Σ^∞) and for the shift T of the dynamical system (defined on the unit interval). In the case of a dynamical source, nice relations indeed exist, between these two versions of T (at the one hand), and between the encoding map τ of the dynamical source and the map σ (at the other hand)

$$T(M(x)) = M(T(x)), \qquad \sigma(M(x)) = \tau(x).$$

We consider the problem of comparing n infinite words independently produced by the same source, we proceed by elementary comparisons between symbols, and we use the two maps σ and T. The underlying structure is a tree, called a trie. Let $\mathcal{X}$ be a sequence of words $(X_1, X_2, \ldots, X_n) \in (\Sigma^\infty)^n$. We deal with the sequence $\sigma(\mathcal{X})$ formed with the first symbol $\sigma(X_i)$ of each word X_i; it is called the first "slice" of $\mathcal{X}$,

$$\sigma(\mathcal{X}) := (\sigma(X_1), \sigma(X_2), \ldots, \sigma(X_n)).$$

Thus, two distinct kinds of collections of symbols will intervene in the analysis: the infinite words produced by the source (seen as vertical words) and also finite sequences seen as horizontal slices). See Figure 5.

To build the tree structure, we start from the root. First one groups together all the words which begin with the same first symbol m, along a branch labelled by m, so that the corresponding subtree groups all the words beginning by symbol m and stripped from their initial symbol, namely the shifted sequence

$$T_{[m]}(\mathcal{X}) := (T_{[m]}(X_1), T_{[m]}(X_2), \ldots, T_{[m]}(X_n)).$$

This process of splitting will continue until all words have been separated. More formally:

Definition 3. *One associates to a sequence $\mathcal{X}$ of words produced by the same source, a digital tree, called a trie, denoted by* `Trie`$(\mathcal{X})$*, and defined by the following recursive rules:*

- (R_0) *If $\mathcal{X} = \emptyset$, then* `Trie`$(\mathcal{X})$ *is the empty tree.*
- (R_1) *If $\mathcal{X} = (X)$ has cardinality equal to 1, then* `Trie`$(\mathcal{X})$ *consists of a single leaf node that contains the word X.*

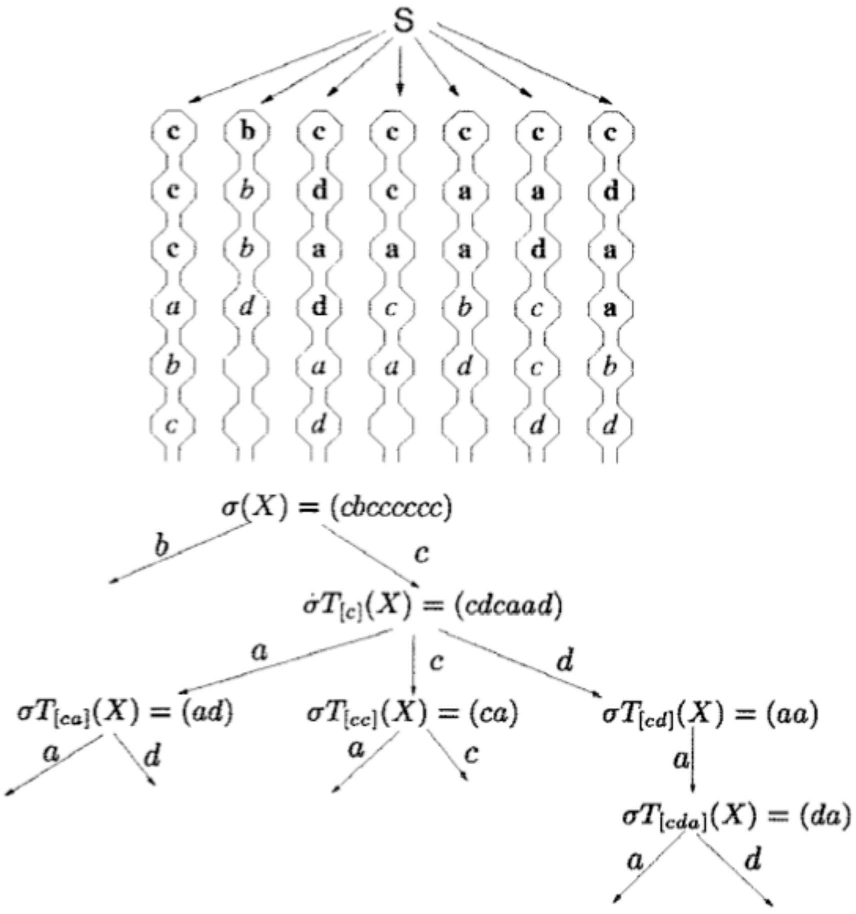


FIGURE 5. The infinite words ("vertical words") and the finite slices ("horizontal words")

(R_2) *If $\mathcal{X}$ has cardinality $|\mathcal{X}|$ at least equal to 2, then* $\texttt{Trie}(\mathcal{X})$ *is an internal node represented by $\circ$ to which are attached all the subtries built on the set $T_{[m]}(\mathcal{X})$. Then* $\texttt{Trie}(\mathcal{X})$ *is defined by*

$$\texttt{Trie}(\mathcal{X}) = \Big\langle \sigma(\mathcal{X}), \big\{\texttt{Trie}(T_{[m]}(\mathcal{X}))\big\}_{m\in\Sigma} \Big\rangle. \tag{16}$$

Such a tree structure underlies classical radix sorting methods. It can be built by following the recursive rules R_0, R_1, R_2. There, any prefix w which is common to at least two words of $\mathcal{X}$ is associated to an internal node of the trie. This internal node, which can be labelled by this common prefix w, is the root of the subtrie relative to the shifted sequence $T_{[w]}(\mathcal{X})$. Here, for any prefix $w = m_1 m_2 \dots m_k \in \Sigma^\star$, the mapping

$$T_{[w]} := T_{[m_k]} \circ T_{[m_{k-1}]} \circ \dots \circ T_{[m_1]} \tag{17}$$

is only defined on words which begin with the prefix w, and it associates to any such word the word stripped from its prefix w. See an instance of a trie in Figure 6.

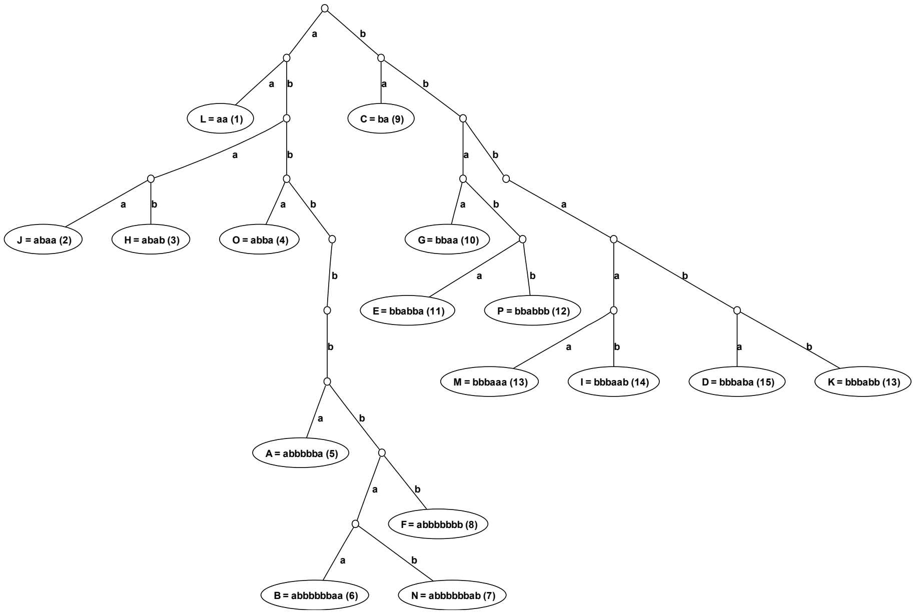


FIGURE 6. A trie built from the memoryless source $p_a = 1/3, p_b = 2/3$ on a set of words of sixteen words
A = $\underline{\text{abbbbba}}$aabab B = $\underline{\text{abbbbbbaa}}$baa C = $\underline{\text{ba}}$abbbabbbba
D = $\underline{\text{bbbaba}}$bbbaab E = $\underline{\text{bbabba}}$ababbb F = $\underline{\text{abbbbbbb}}$babb
G = $\underline{\text{bbaa}}$bbabbaba H = $\underline{\text{abab}}$bbabbbab I = $\underline{\text{bbbaab}}$bbbbbb
J = $\underline{\text{abaa}}$bbbbaabb K = $\underline{\text{bbbabb}}$bbbbaa L = $\underline{\text{aa}}$aabbabaaba
M = $\underline{\text{bbbaaa}}$bbbbbb N = $\underline{\text{abbbbbbab}}$baa O = $\underline{\text{abba}}$bababbbb
P = $\underline{\text{bbabbb}}$aaaabb
The underlined symbols are those which are used for building the trie.

2.2. **Various implementations of tries.** In the abstract tree structure representing the trie, each internal node has links to its children. They are three main implementations of such a node, each of them is based on classical data structures like arrays, ordered lists and binary search trees. There result "hybrid tries" which are the hybridation of an overall trie structure with another data structure to access children of a node. Then, three kinds of hybrid tries are considered depending on the method chosen to access children of a node. Figure 7 shows these three representations for the trie built on a given set of words.

(*i*) The more natural implementation uses arrays whose cardinality equals the size of the alphabet. Then one accesses children directly through an array of pointers; note that it is impossible for infinite alphabets, and space-wasting for large alphabets (too much null pointers are allocated). This classical solution which gives raise to "array-tries" is adequate only when the cardinality of the alphabet is small (typically for binary words).

(*ii*) The “list-trie” structure remedies the high storage cost of array-tries by linking sister subtrees at the expense of replacing direct array access by a *linked list* traversal.

(*iii*) The “bst-trie” uses *binary search trees* (bst) as subtree access method, with the goal of combining advantages of array-tries in terms of time cost, and list-tries in terms of storage cost. As stated in the introduction, this hybrid trie obtained is strictly equivalent to the *ternary search trie* structure proposed recently. Indeed, the bst-trie can be viewed as a ternary tree where search on symbols is conducted like in a standard binary search tree over the alphabet set Σ, while trie descent is performed by following an escape pointer whenever equality of symbols of detected.

The structure of a classical trie does not depend on the order of insertion of the words *i.e.* the order inside the sequence $\mathcal{X}$, but only on the set of words itself. However, when mixing the trie structure with node structures for which it is not the case (like binary search trees), it is no longer true.

The structure of the classical trie does not depend either on the order relation between symbols. Here, since structure nodes need an ordered alphabet, we consider an order on the symbols of the alphabet.

2.3. **Parameters.** The *level* of a node in a trie is the number of edges that connect it to the root. The *height* of the trie is the maximum level of any leaf. It is then a measure of distance between the two closest elements of $\mathcal{X}$. It is the minimum number of comparisons to separate without question any pair (X_i, X_j) of elements of $\mathcal{X}$. The *path length* of the trie is the sum of the levels of all leaves. The path length equals the total number of symbols that need to be examined in order to distinguish all elements of $\mathcal{X}$. Divided by the number of elements, it is also by definition the cost of a positive search (i.e. searching for a word that is present in the trie). The *size* of the tree is the number of its internal nodes. Adding to the size, the cardinality of $\mathcal{X}$ gives the number of prefixes necessary to isolate all elements of $\mathcal{X}$). It gives also a precise estimate of the place needed in memory, concerning elementary node structures to allocate, to store the trie in a real-life implementation.

In an hybrid trie, the external length path decomposes in two parts: the first is linked to the trie structure, the second is the extra-cost due to the traversal of internal node structures. It is this overhead which is analysed here for hybrid tries. If we are interested in a global external path length, we just have to combine our results with those about tries. interest for the hybrid trie.

2.4. **Additive parameters.** Let us consider now an “additive” parameter γ on $\texttt{Trie}(\mathcal{X})$ that decomposes recursively into a parameter δ over the root that contains the slice $\sigma(\mathcal{X})$ and the parameter γ on all the possible subtries relative to sequences $T_{[m]}(\mathcal{X})$. Such a parameter has a recursive definition quite similar as the definition of the structure itself (16)

$$\gamma[\texttt{Trie}(\mathcal{X})] = 0 \quad \text{if} \quad |\mathcal{X}| \leq 1,$$

$$\gamma[\texttt{Trie}(\mathcal{X})] = \delta(\sigma(\mathcal{X})) + \sum_{m\in\Sigma} \gamma[\texttt{Trie}(T_{[m]}(\mathcal{X}))] \quad \text{if} \quad |\mathcal{X}| \leq 2.$$

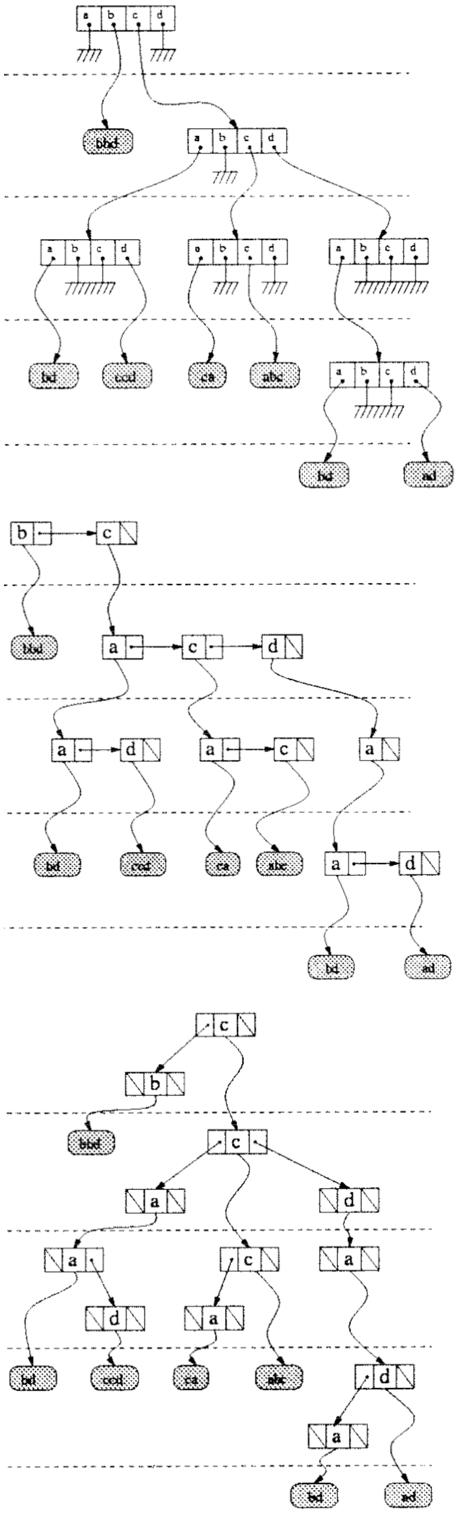


Figure 7. The three possible implementations of a trie: array trie on the top – list-trie on the middle– bst-trie on the bottom.

The recurrence relation can be unwinded, by using the mappings $T_{[w]}$ defined in (17), and the parameter γ on $\texttt{Trie}(X)$ is expressed as

$$\gamma[\texttt{Trie}(\mathcal{X})] = \sum_{w \in \Sigma^*} \delta[\sigma T_{[w]}(\mathcal{X})],$$

provided that $\delta(s)$ is zero on slices s that contain at most one symbol.
Our analysis needs keeping track of the process of construction of internal nodes. Even if the elements contained in each internal node labelled by prefix w are only the different symbols of the slice

$$\sigma T_{[w]}(\mathcal{X}) := (\sigma T_{[w]}(X_1), \sigma T_{[w]}(X_2), \ldots, \sigma T_{[w]}(X_n)),$$

we need (solely for the analysis) remembering the totality of the information contained in the slice $\sigma T_{[w]}(\mathcal{X})$, in particular the order in which the symbols are inserted (defined by the function $\texttt{ord}$, the number of occurences N_i of each symbol a_i, and the total number N of the symbols in the slice.

2.5. **Probabilistic model at nodes.** We describe now the probabilistic model that is induced by the Poisson model on each possible node of the trie relative to a prefix w. Since the probability that a word begins with prefix w is equal to the fundamental probability p_w, the probability that the following symbol m is emitted is equal to

$$p_{m|w} = \frac{p_{w \cdot m}}{p_w}. \tag{18}$$

Here, the notation $w \cdot m$ denotes the string obtained by concatenation of string w and symbol m. Furthermore, since all elements of $\mathcal{X}$ are independently drawn, the symbols that compose each slice are independent. Thus, at the internal node labelled by w, symbols are then emitted by a memoryless source B_w associated to probabilities $\{p_{m|w}\}_{m \in \Sigma}$. The (horizontal) slices are then very simple objects, even if the vertical (infinite words) may be emitted by a complex source.

Moreover, if the length of $\mathcal{X}$ is a random Poisson variable of rate z, the length of the string $\sigma T_{[w]}(\mathcal{X})$ is also a random Poisson variable of rate zp_w. This implies that the expectation of parameter γ can be expressed as a sum of expectations of parameter δ,

$$\mathbb{E}[\gamma, \mathcal{P}_z, S] = \sum_{w \in \Sigma^*} \mathbb{E}[\delta, \mathcal{P}_{zp_w}, B_w], \tag{19}$$

where w ranges over all finite prefixes and B_w denotes a memoryless source relying on the set of probabilities $\{p_{m|w}\}_{m \in \Sigma}$ defined in (18).

2.6. **Node search costs.** We consider here four parameters of interest that lie in the scope of additive parameters: the size, and the various path lengths (relative to each kind of hybrid trie). For example, the toll δ_R associated to size equals 1 provided that the internal node indexed by w exists or equivalently that the slice $\sigma T_{[w]}(\mathcal{X})$ has at least two symbols. The toll δ_C for path length of an array-trie is simply the number of symbols in the slice, provided that the node exists. Then, if we denote by $N(s)$ the cardinality of the slice s,

$$\delta_R = \mathbf{1}_{[N \geq 2]}, \qquad \delta_C = N \, \mathbf{1}_{[N \geq 2]}.$$

The parameters $\delta_L(s)$ and $\delta_A(s)$ (relative respectively to path lengths for list-tries and bst-tries) are exactly traversal costs through associated node-structures built over a slice s. The symbols of s (with repetitions allowed) are inserted in order in a structure (a list or a binary search tree), and then the toll is the cost due to retrieve in the structure each occurence of each symbol of s. Then, if we denote by $N_i(s)$ the number of elements of s whose value equals a_i, and by $N_{[i,j]}$ the number of elements of s whose value equals a_k, with $k \in [i,j]$, we have

$$\delta_L = \sum_{i\in\Sigma} N_i \sum_{j<i} \mathbf{1}_{[N_j \geq 1]}, \qquad \delta_A = \sum_{i\in\Sigma} N_i \sum_{j\neq i} \mathbf{1}_{[a_j \text{ ancestor of } a_i \text{ in bst}]}$$

Remark that a_j is an ancestor of a_i in bst(s) if and only if there exists x in s that satisfies

$$\texttt{val}(x) = a_j, \qquad \texttt{ord}(x) = \min\{\texttt{ord}(z);\ \ z \in s,\ \texttt{val}(z) = a_k,\ k \in [i,j]\}.$$

Then, when one considers all the slices $\tau(s)$ obtained with a permutation τ of $[1..N(s)]$, the probability of the event $[a_j$ is an ancestor of a_i in bst$(s)]$ is equal to the ratio $N_j/N_{[i,j]}$. Then, the average value $\underline{\delta}_A$ of δ_A on the set of all the slices $\tau(s)$ obtained with a permutation τ of $[1..N(s)]$ is

$$\underline{\delta}_A = 2\sum_{i\in\Sigma} N_i \sum_{j>i} \frac{N_j}{N_{[i,j]}} = 2\sum_{i\in\Sigma}\sum_{j>i} \frac{N_i N_j}{N_i + N_j + N_{]i,j[}}.$$

Remark that the two parameters δ_A and $\underline{\delta}_A$ have the same expectations in the Bernoulli model (and thus in the Poisson model). Remark also that the formulae which express δ_L and $\underline{\delta}_A$ involve independent variables which all follow Poisson laws. This allows easy computations and we obtain:

Proposition 2. [Toll parameters] *Let B be a memoryless source relying on a set of probabilities $\{p_i\}_{i\in\Sigma}$ and $\mathcal{P}_z$ the Poisson model of rate z. Then, in the model $(\mathcal{P}_z, B)$, expectations of the toll parameters relative to the size of a trie and the path length of an array-trie are respectively*

$$\mathbb{E}[\delta_R, \mathcal{P}_z, B] = 1 - (1+z)e^{-z}, \qquad \mathbb{E}[\delta_C, \mathcal{P}_z, B] = z(1 - e^{-z}).$$

In the model $(\mathcal{P}_z, B)$, the expectations of traversal costs for ordered lists and binary search trees are respectively

$$\begin{aligned}
\mathbb{E}[\delta_L, \mathcal{P}_z, B] &= \sum_{j\in\Sigma} P_{[>j]}\ z\ (1 - e^{-p_j z}), \\
\mathbb{E}[\delta_A, \mathcal{P}_z, B] &= 2 \sum_{\substack{(i,j)\in\Sigma^2 \\ i<j}} \frac{p_i\, p_j}{{P_{[i,j]}}^2} \left[e^{-zP_{[i,j]}} - 1 + zP_{[i,j]}\right],
\end{aligned}$$

where $P_{[i,j]} = \sum_{k=i}^{j} p_k$ and $P_{[>j]} = \sum_{k>j} p_k$.

Remark: the expression of the expectation of parameter δ_A is useful to recover in a simple way the result of Burge about the average path length in a `Bst` with repeated keys.

2.7. **Size and path length in the Poisson model.** The form of the recurrence (19), the form of the probabilities at each node (18) and the expressions obtained in Theorem 1 insures great simplification on formulae during the unwinding process of the recursion, so that the expectations of the four additive parameters can be solely expressed with fundamental measures.

Theorem 1. [Mean trie costs in the Poisson Model] *Let S be a source and $\mathcal{P}_z$ the Poisson model of rate z. Then expectations in the model $(\mathcal{P}_z, S)$ of the toll parameters respectively relative to the size of a trie, path length of an array-trie trie, path length of an ordered-list trie, path length of a bst-trie are*

$$\begin{aligned}
\widetilde{R}(z) &= \sum_{w\in\Sigma^*} \left[1-(1+zp_w)e^{-zp_w}\right],\\
\widetilde{C}(z) &= \sum_{w\in\Sigma^*} zp_w\left[1-e^{-zp_w}\right]\\
\widetilde{L}(z) &= \sum_{w\in\Sigma^*}\sum_{i\in\Sigma} z\,P_{w\cdot[>i]}\,(1-e^{-zp_{w\cdot i}})\\
\widetilde{A}(z) &= 2\sum_{w\in\Sigma^*}\sum_{\substack{(i,j)\in\Sigma^2\\ i<j}} \frac{p_{w\cdot i}\,p_{w\cdot j}}{{P_{w\cdot[i,j]}}^2}\left[e^{-zP_{w\cdot[i,j]}}-1+zP_{w\cdot[i,j]}\right],
\end{aligned}$$

where $P_{w\cdot[i,j]} = \sum_{k=i}^{j} p_{w\cdot k}$, and $P_{w\cdot[>j]} = \sum_{k>j} p_{w\cdot k}$.

2.8. **Size and path lengths in the Bernoulli model.** We can now return to the Bernoulli model with Equation (15). So, the expectations of the four additive parameters in the Bernoulli model $(\mathcal{B}_n, S)$ can be also be expressed with fundamental probabilities:

Let $(\mathcal{B}_n, S)$ be the Bernoulli model relative to a fixed number n of words independently drawn from a source S. Then the expectations for the size of a trie, path length of an array-trie, path length of an ordered-list trie, path length of a bst trie are

$$\begin{aligned}
R(n) &= \sum_{w\in\Sigma^*}\left[1-(1-p_w)^n - np_w\,(1-p_w)^{n-1}\right]\\
C(n) &= \sum_{w\in\Sigma^*} np_w[1-(1-p_w)^{n-1}]\\
L(n) &= \sum_{w\in\Sigma^*}\sum_{i\in\Sigma} nP_{w\cdot[>i]}(1-(1-p_{w\cdot i})^{n-1})\\
A(n) &= 2\sum_{w\in\Sigma^*}\sum_{\substack{(i,j)\in\Sigma^2\\ i<j}} \frac{p_{w\cdot i}\,p_{w\cdot j}}{{P_{w\cdot[i,j]}}^2}\left[(1-P_{w\cdot[i,j]})^n-1+nP_{w\cdot[i,j]}\right],
\end{aligned}$$

where $P_{w\cdot[i,j]} = \sum_{k=i}^{j} p_{w\cdot k}$ and $P_{w\cdot[>j]} = \sum_{k>j} p_{w\cdot k}$ as before.

With the use of binomial expansions, the expectations in the Bernoulli model with n fixed all follow the same pattern:

Theorem 2. [Mean trie costs in the Bernoulli model] *Let $(\mathcal{B}_n, S)$ be the Bernoulli model relative to a fixed number n of words independently drawn from a source S. Then the expectations for the size of a trie, path length of an array-trie, path length* *of an ordered-list trie, path length of a bst trie are all expressed as an expression of the form*

$$T(n) = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \varpi_T(k)$$

where the function $\varpi_T(s)$ is a Dirichlet series which depends on the parameter T and the source $\mathcal{S}$ and is defined as

$$\varpi_R(s) = (s-1) \sum_{w \in \Sigma^*} p_w^s \qquad \varpi_C(s) = s \sum_{w \in \Sigma^*} p_w^s$$

$$\begin{aligned} \varpi_L(s) &= \sum_{w \in \Sigma^*} p_w^s \; K_L(s,w) \qquad \textit{with} \quad K_L(s,w) = \sum_{i \in \Sigma} P_{[>i]|w} \; p_{i|w}^{s-1} \\ \varpi_A(s) &= 2 \sum_{w \in \Sigma^*} p_w^s \; K_A(s,w) \qquad \textit{with} \quad K_A(s,w) = \sum_{\substack{(i,j) \in \Sigma^2 \\ i<j}} p_{i|w} \, p_{j|w} \; P_{[i,j]|w}^{s-2} \end{aligned}$$

where $p_{i|w}$ is the probability of emitting the symbol a_i when the prefix w was previously emitted, $P_{[i,j]|w}$ is the probability of emitting a symbol a_k with $k \in [i,j]$ when the prefix w was previously emitted, and $P_{[>j]|w}$ is the probability of emitting a symbol a_k with $k > j$ when the prefix w was previously emitted.

Remark that the Dirichlet series ϖ_R and ϖ_C are expressed with the Dirichlet series of the source as

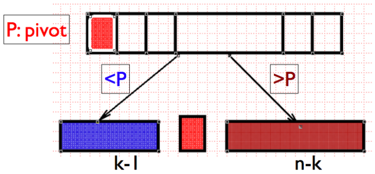
$$\varpi_R(s) = (s-1)\,\Lambda(s), \quad \varpi_C(s) = s\,\Lambda(s).$$

3. Analysis of QuickSort and QuickSelect

We first recall the (informal) recursive definitions of the algorithms `QuickSort` and `QuickSelect`, when they operate on a sequence $\mathcal{X}$ formed with distinct elements, placed in a array B of size n. The algorithm `QuickSort`(B, n) sorts the array B, whereas the algorithm `QuickSelect` (B, m, n) finds the value of the key of rank m in the array B. We also describe a third algorithm named `QuickVal`(B, b, n) which finds the rank of the key b inside the array B of size n. The algoritthm `QuickVal` is dual of `QuickSelect` since, for a given key b, the roles of the rank and of the value are exchanged. We study a randomized version of the algorithms, when the pivot is chosen at random in B.

The execution of the `QuickSort` algorithm builds a Binary Search Tree [`BST` in shorthand], the Binary Search Tree formed with the pivots chosen in the recursive calls. This is why the `QuickSort` algorithm is closely related to the `BST` structure.

```
QuickSort $(n, B)$:   sorts the array $B$
          Randomly choose a pivot in $B$;
          $(k, B_-, B_+) :=$ Partition$(B)$;
           QuickSort $(k-1, B_-)$;
           QuickSort $(n-k, B_+)$.
```



```
QuickSelect $(n, m, B)$: returns the value of the element of rank $m$ in $B$.
        Randomly choose a pivot in $B$;
        $(k, B_-, B_+) :=$ Partition$(B)$;
        If $m = k$ then QuickSelect := pivot
                    else if $m < k$ then QuickSelect $(k-1, m, B_-)$
                                else QuickSelect $(n-k, m-k, B_+)$;
```

```
QuickVal $(n, b, B)$. : returns the rank of the element $a$ in $C = B \cup \{b\}$
    $C := B \cup \{b\}$
    QV $(n, b, C)$;

QV $(n, b, C)$.
    Choose a pivot in $C$;
    $(k, C_-, C_+) :=$ Partition$(C)$;
    If $b =$ pivot then QV := $k$
                else if $b <$ pivot then QV := QV $(k-1, b, C_-)$
                            else QV := $k+$ QV $(n-k, b, C_+)$;
```

3.1. **Mean number of key-comparisons.** These algorithms deal with a sequence $\mathcal{X}$ formed with n distinct keys $X_1, X_2, \ldots, X_n$ of the same ordered set Ω. They perform comparisons and exchanges between keys, and the (usual) unit cost is the key–comparison. The behaviour of the algorithm (wrt to key–comparisons) only depends on the relative order between the keys. It is then sufficient to restrict to the case when $\Omega = [1..n]$. The input set is then the permutation group $\mathfrak{S}_n$, endowed with uniform probability.

Then, the analysis of all these algorithms is very well known, with respect to the number of key–comparisons performed in the worst-case, or in the average case. Figure 8 recalls these results in the average-case, for various values of rank m.

3.2. **Coincidence and fundamental triangles.** Here, we are interested by a more realistic cost, the number of symbol comparaisons performed by these algorithms, when the keys are words independently produced by the same source. The keys are ordered with respect to the lexicographic order, and the cost for comparing two words (measured as the number of symbol comparisons needed) is equal to 1+ the length of the longest common prefix of the two words. For instance, for the associated BST, the path-length of interest is now a weighted path-length, called in the following the symbol-path-length, which measures the total number of comparisons needed to build this BST. See Figure 9 for an example.

This is why the following definition will be useful.

Definition 4. *The* coincidence function $\gamma(u,t)$ *is the length of the largest common prefix of* $M(u)$ *and* $M(t)$*, namely,*

$$\gamma(u,t) := \max\{\ell : m_j(u) = m_j(t),\ \forall j \le \ell\}.$$

The coincidence $\gamma(u,t)$ is at least ℓ if and only if the two words $M(u)$ and $M(t)$ have the same common prefix w of length ℓ: The parameters u and t belong to the same

`QuickSort` (n) or Path-Length-`BST`(n)	sorts		$K_n \sim 2n \log n$
`QuickMin`(n) `QuickMax`(n) `QuickRand`(n)	minimum maximum	$m = 1$ $m = n$ $m \in [1..n]_{\mathcal{R}}$	$K_n \sim 2n$ $K_n \sim 2n$ $K_n \sim 3n$
`QuickQuant`$_\alpha(n)$ `QuickMed`(n)	α–quantile median	$m = \lfloor \alpha n \rfloor$ $m = \lfloor n/2 \rfloor$	$K_n \sim \kappa(\alpha)\, n$ $K_n \sim 2(1 + \log 2)n$

On the right, the function

$$\kappa : \alpha \mapsto 2\,[1 + h(\alpha)]$$

where $h(\cdot)$ is the entropy function

$$h(\alpha) = \alpha|\log \alpha| + (1-\alpha)|\log(1-\alpha)|$$

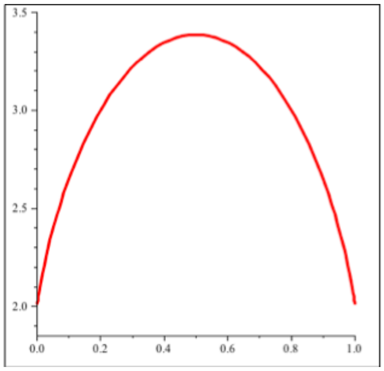


FIGURE 8. The mean number of key-comparisons for the main algorithms of interest.

fundamental interval $\mathcal{I}_w$ relative to a prefix of length ℓ, as it is defined in Section 1.1. Since any pair of words of the source is of the form $(M(u), M(t))$ with $0 \le u \le t \le 1$, it is convenient to introduce the triangle $\mathcal{T} := \{(u,t) : 0 \le u \le t \le 1\}$, and the domain $\mathcal{T} \cap [\gamma \ge \ell]$ is written as a union

$$(20) \qquad [\gamma \ge \ell] = \bigcup_{w \in \Sigma^\ell} (\mathcal{I}_w \times \mathcal{I}_w) \qquad \text{so that} \quad \mathcal{T} \cap [\gamma \ge \ell] = \bigcup_{w \in \Sigma^\ell} \mathcal{T}_w,$$

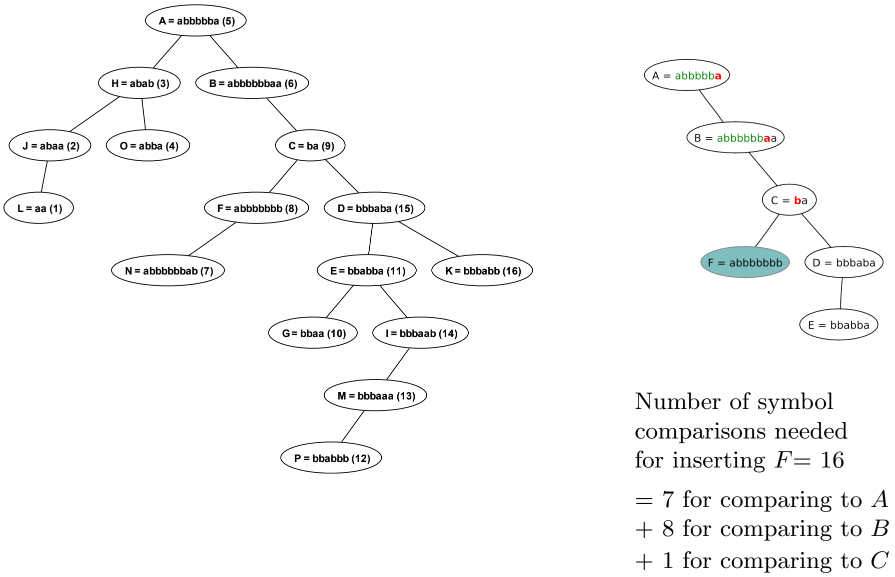
where $\mathcal{T}_w = (\mathcal{I}_w \times \mathcal{I}_w) \cap \mathcal{T}$. This motivates the definition:

Definition 5. *We set $\mathcal{T} := \{(u,t) : 0 \le u \le t \le 1\}$. For each $w \in \Sigma^\star$, the* fundamental triangle *of prefix w, denoted by $\mathcal{T}_w$, is the triangle built on the fundamental interval $\mathcal{I}_w := [b_w, c_w]$ corresponding to w,*

$$\mathcal{T}_w := \{(u,t) : b_w \le u \le t \le c_w\} = (\mathcal{I}_w \times \mathcal{I}_w) \cap \mathcal{T}.$$

Figure 10 represents the fundamental triangles for two memoryless sources.

3.3. **An expression for the mean number of symbol comparisons.** The second object of our analysis is the *density* of the algorithm $\mathcal{A}$, which measures the number of key–comparisons performed by the algorithm.



Number of symbol comparisons needed for inserting F= 16

= 7 for comparing to A

+ 8 for comparing to B

+ 1 for comparing to C

FIGURE 9. The BST built on the set described in Figure 6 – The cost for inserting key F.

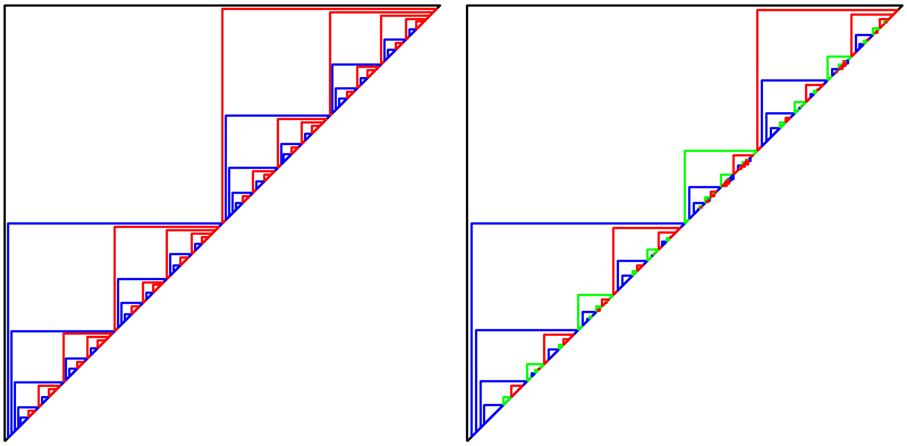


FIGURE 10. The fundamental triangles for two memoryless sources- on the left the unbiased source on $\{a, b\}$– on the right, the memoryless source with $p_a = 1/2, p_B = 1/6, p_c = 1/3$.

Definition 6. *The* density *of an algorithm $\mathcal{A}$ which compares words from the same probabilistic source $\mathcal{S}$ is defined as follows, in each probabilistic model of interest:*

(i) In the model $\mathcal{B}_n$,
$\phi_n(u,t)\,du\,dt$:= *"the mean number of key–comparisons performed by $\mathcal{A}$ between words $M(u'), M(t')$, with $u' \in [u, u+du]$ and $t' \in [t, t+dt]$, when the input sequence $\mathcal{X}$ is any sequence of the model $(\mathcal{B}_n, \mathcal{S})$ which contains two words $M(u'), M(t')$."*

(ii) In the model $\mathcal{P}_z$,
$\tilde\phi_z(u,t)\,du\,dt$:= *"the mean number of key–comparisons performed by $\mathcal{A}$ between words $M(u'), M(t')$, with $u' \in [u, u+du]$ and $t' \in [t, t+dt]$, when the input sequence $\mathcal{X}$ is any sequence of the model $(\mathcal{P}_z, \mathcal{S})$ which contains two words $M(u'), M(t')$."*

Our first result establishes a relation between the mean numbers of symbol-comparisons, the densities and the fundamental triangles

Proposition 3. *Fix a source on alphabet Σ, with fundamental triangles $\mathcal{T}_w$. For any integrable function g on the unit triangle $\mathcal{T}$, define the* integral transform

$$\mathcal{J}[g] := \sum_{w\in\Sigma^\star} \int_{\mathcal{T}_w} g(u,t)\,du\,dt.$$

Then the mean numbers $S(n), \tilde S(z)$ of symbol comparisons performed by $\mathcal{A}$ in the Bernouli model and in the Poisson model are equal to

$$S(n) = \mathcal{J}[\phi_n], \qquad \tilde S(z) = \mathcal{J}[\tilde\phi_z]$$

where $\phi_n, \tilde\phi_z$ are the densities of algorithm $\mathcal{A}$.

Proof. The number of symbol comparisons needed to compare two words $M(u)$ and $M(t)$, is $\gamma(u,t)+1$ and the mean numbers $S(n), \tilde S(z)$ of symbol comparisons performed by $\mathcal{A}$ satisfy

$$S(n) = \int_{\mathcal{T}} [\gamma(u,t)+1]\,\phi_n(u,t)\,du\,dt, \qquad \tilde S(z) = \int_{\mathcal{T}} [\gamma(u,t)+1]\,\tilde\phi_z(u,t)\,du\,dt$$

where $\phi_n(u,t), \tilde\phi_Z(u,t)$ are the densities of the algorithm. The useful identity

$$\sum_{\ell\ge 0} (\ell+1)\mathbf{1}_{[\gamma=\ell]} = \sum_{\ell\ge 0} \mathbf{1}_{[\gamma\ge\ell]}$$

holds for *any* integer-valued random variable γ ($\mathbf{1}_A$ is the indicator of A). With (20), this ends the proof. □

3.4. **Computation in the Poisson model.** It is then essential to characterize the words which are compared by each algorithm. For a given algorithm, we define the level of an element b of the array B (denoted by $\texttt{lev}(b)$), as the level of the recursion where it is chosen as the pivot. This is thus the level of the pivot in the `BST` associated. If it is never chosen as a pivot, we let $\texttt{lev}(b) = +\infty$. The following property will be important in our study:

Lemma 2. *Two elements a, c (with $a < c$) of the array B are compared by* $\texttt{QuickSort}(B)$ *if and only if a or c is the pivot of smallest level in the set $\{x \in B;\ \ x \in [a,c]\}$. Two elements a, c (with $a < c$) of the array B are compared by* $\texttt{QuickVal}(B, b)$ *if and only if a or c is the pivot of smallest level in the set $\{x \in B;\ \ x \in [\min(a,b), \max(b,c)]\}$.*

There is no such characterization for `QuickSelect`. But we will "replace" the algorithm `QuickSelect` by the algorithm `QuickVal` previously described. More precisely, we call $\texttt{QuickVal}_\alpha$ the `QuickVal` algorithm, when used to seek the rank of the word $M(\alpha)$. As the *α–quantile* of a random set of words of large enough cardinality is, with high probability, close to the word $M(\alpha)$, the behaviours of $\texttt{QuickVal}_\alpha(n)$ and $\texttt{QuickQuant}_\alpha(n)$ should be asymptotically similar. This is indeed the case, as we will see it later.

The following statement shows that the Poissonized densities relative to `QuickSort` and $\texttt{QuickVal}_\alpha$ admit simple expressions, which in turn entail nice expressions for the mean value $\tilde{S}(Z)$ in this Poisson model, via the equality $\tilde{S}(Z) = \mathcal{J}[\tilde{\phi}_Z]$.

Theorem 3. [Mean costs for `QuckSort` and `QuickVal` in the Poisson model] *Set $f_1(\theta) := \theta^{-2}[e^{-\theta} - 1 + \theta]$. The mean numbers of comparisons of* `QuickSort` *and* $\texttt{QuickVal}_\alpha$ *in the Poisson model $\mathcal{P}_z$ satisfy*

$$\tilde{B}(z) = 2z^2 \mathcal{J}[f_1(z(t-u))], \qquad \tilde{V}^{(\alpha)}(z) = 2z^2\, \mathcal{J}[f_1(z(\max(t,\alpha) - \min(u,\alpha)))]$$

Proof. We begin with the case of `QuickSort`. The probability that $M(u')$ and $M(t')$ are both keys for some $u' \in [u, u+du]$ and $v' \in [v+dv]$ is $zdu \cdot zdt$, since the two intervals are disjoint. Conditionnally, given that $M(u')$ and $M(t')$ are both keys of a fixed sequence $\mathcal{X}$, they are compared if and only if $M(u')$ or $M(t')$ is chosen as the first pivot (with respect to the recursion level) amongst the set $\mathcal{M} := \{M(z) \in \mathcal{X};\ z \in [u', t']\}$. The cardinality of the "good" set $\{M(u'), M(t')\}$ is 2, while the total cardinality of $\mathcal{M}$ equals $2 + N[u', t'](\mathcal{X})$, where $N[u', t'](\mathcal{X})$ is the number of keys of the sequence $\mathcal{X}$ strictly between $M(u')$ and $M(t')$. Then, for any fixed sequence $\mathcal{X}$ of words which contains the words $M(u'), M(t')$, the probability that $M(u')$ and $M(t')$ are compared is

$$\frac{2}{2 + N[u', t'](\mathcal{X})} \approx \frac{2}{2 + N[u, t](\mathcal{X})}.$$

To evaluate the mean value of this ratio in the Poisson model (when the sequence $\mathcal{X}$ now varies), Lemma 1 states that, if we draw $\mathcal{P}_z$ i.i.d. random variables uniformly distributed over $[0,1]$, the number $N(\lambda)$ of those that fall in an interval of (Lebesgue) measure λ is $\mathcal{P}_{\lambda z}$ distributed, so that

$$\mathbb{E}\left[\frac{2}{N(\lambda)+2}\right] = \sum_{k \geq 0} \frac{2}{k+2}\, e^{-\lambda z}\, \frac{(\lambda z)^k}{k!} = \frac{2}{\lambda^2 z^2} f_1(\lambda z).$$

In the case of $\texttt{QuickVal}_\alpha$, the proof is in the same vein. But, now, given that $M(u')$ and $M(t')$ are both keys of a fixed sequence $\mathcal{X}$, they are compared if and only if $M(u')$ or $M(t')$ is chosen as the first pivot (with respect to the recursion level) amongst the set

$$\mathcal{M} := \{M(z) \in \mathcal{X};\ z \in [x', y']\}, \qquad \text{with} \quad x' = \min(u', \alpha),\ y' := \max(t', \alpha)$$

Then, the proof is the same as for `QuickSort`, when one replaces u by $\min(u, \alpha)$ and t by $\max(t, \alpha)$. □

3.5. **Exact formula in the Bernoulli model.** We now return to the model of prime interest, where the number of keys is a fixed number n.

Theorem 4. [Mean costs of `QuickSort` and `QuickVal` in the Bernoulli model] *Assume that the Dirichlet series* $\Lambda(s)$ *converges at* $s = 2$. *Then the mean values associated with* `QuickSort` *and* $\texttt{QuickVal}_\alpha$ *can be expressed as*

$$S(n) = \sum_{k=2}^{n} \binom{n}{k} (-1)^k \varpi_S(k), \qquad \textit{for } n \geq 2,$$

where $\varpi_S(s)$ *is a series of Dirichlet type, defined for* $\Re s \geq 2$. *It depends on the algorithm and the source and is called the* mixed Dirichlet series. *It is given by*

$$\varpi_B(s) = 2\mathcal{J}[(t-u)^{s-2}] = 2\frac{\Lambda(s)}{s(s-1)}, \qquad \varpi_V^{(\alpha)}(s) = 2\mathcal{J}[(\max(t,\alpha) - \min(\alpha,u))^{s-2}].$$

Proof. We let

$$x(u,t) := u, \quad y(u,t) := t \qquad \text{for } \texttt{QuickSort}$$

$$x(u,t) := \min(u,\alpha), \quad y(u,t) := \max(t,\alpha) \qquad \text{for } \texttt{QuickVal}_\alpha.$$

By expanding f_1, then exchanging the order of summation and integration, one obtains

$$\tilde{S}(z) = \sum_{k=2}^{\infty} (-1)^k \varpi(k) \frac{z^k}{k!}, \quad \text{with } \varpi(k) := 2 \sum_{w \in \Sigma^\star} \int_{\mathcal{T}_w} (y(u,t) - x(u,t))^{k-2} du dt.$$

Analytically, the previous form is justified as soon as the integral defining $\varpi(2)$ is convergent. Remark that, in the case of `Quicksort`, each integral on each fundamental triangle can be easily easily computed

$$\int_{\mathcal{T}_w} (t-u)^{k-2} du dt = \frac{1}{k(k-1)} p_w^k \qquad \text{so that} \quad \varpi_B(k) = \frac{2}{k(k-1)} \Lambda(k).$$

Then, in both cases, since $S(n)$ is related to $\tilde{S}(z)$ via the relation of algebraic depoissonisation

$$S(n) = n! \text{ Coeff of } z^n \text{in } \left(e^z \tilde{S}(z)\right),$$

it can be easily recovered by a binomial convolution. □

4. General scheme for the analytical step. Different classes of sources.

4.1. **Two main ways towards the asymptotics.** In the previous two sections, we have obtained exact expressions for the mean values of parameters of interest, in the Poisson model [Theorems 1 and 3] or in Bernoulli model [Theorems 2 and 4]. We wish to obtain now an asymptotic form for these mean values in the Bernoulli model. There are two possible ways described in Figure 11. We have already performed the Algebraic DePoissonisation (AlgDePo in shorthand) in Theorems 2 and 4, and we choose in this paper the way which uses the Rice Formula. If we have chosen the alternative way, we should perform an analytic Depoissonisation (AnDePo in shorthand).

$$\begin{tikzcd}
\text{Mean Value in the Poisson model} \arrow[r, Rightarrow, "\text{AlgDePo}"] \arrow[d, Rightarrow, "\text{Mellin}"'] & \text{Mean Value in the Bernoulli model} \arrow[d, Rightarrow, "\text{Rice}"] \\
\text{Asymptotic Mean Value in the Poisson model} \arrow[r, Rightarrow, "\text{AnDePo}"] & \text{Asymptotic Mean Value in the Bernoulli model}
\end{tikzcd}$$

FIGURE 11. Possible ways to obtain the asymptotic mean value in the Bernoulli model from the exact mean value in the Poisson model.

4.2. **Rice formula.** It transforms an alternate sum into an integral of the complex plane.

Proposition 4. *Let $S(n)$ be a numerical sequence which can be written as*

$$S(n) = \sum_{k=2}^{n} \binom{n}{k} (-1)^k \varpi(k), \qquad \text{for } n \geq 2.$$

Assume that the function $\varpi(s)$ is analytic in $\Re(s) > C$, with $1 < C < 2$, and is there of polynomial growth with order at most r. Then the sequence $S(n)$ admits a Nörlund–Rice representation, *for $n > r+1$ and any $C < d < 2$.*

$$\begin{aligned} S(n) &= \frac{1}{2i\pi} \int_{-d-i\infty}^{-d+i\infty} \varpi(-s) \frac{n!}{s(s+1)\cdots(s+n)}\, ds \\ &= \frac{1}{2i\pi} \int_{-d-i\infty}^{-d+i\infty} \varpi(-s) \frac{\Gamma(n+1)\Gamma(s)}{\Gamma(n+1+s)}\, ds. \end{aligned} \tag{21}$$

Proof. The residue theorem justifies the form

$$S(n) = \frac{1}{2i\pi} \int_{\mathcal{R}} \varpi(-s) \frac{n!}{s(s+1)\cdots(s+n)}\, ds, \tag{22}$$

where $\mathcal{R}$ is a rectangle enclosing the points $-2, \ldots, -n$, whose right vertical line $\Re s = -d$ satisfies $d > C$. This representation is *a priori* valid for $n \geq 2$. For n large enough, i.e., $n > r+1$, it is legitimate to push first the horizontal boundaries of $\mathcal{R}$ to $\pm i\infty$, then the left-most boundary to $-\infty$. We obtain in this way an integral representation, now along the vertical line $\Re s = -d$. This concludes the proof. □

4.3. **Possible behaviours for $\varpi(s)$.** The idea is now to push the contour of integration in (21) to the right, past -1. This is why we have to consider the possible behaviours for the function $\varpi(s)$ near $\Re s = 1$; more precisely on the left of the line $\Re s = 1$. We will later show why the behaviours that are described in the following definition arise in a natural way in the present study related to sources.

Definition 7. *A function $\varpi(s)$ is*

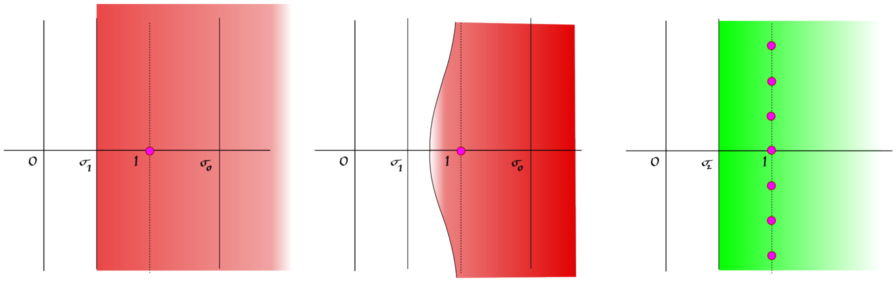


FIGURE 12. Three possible domains where the function $\varpi(s)$ is analytic and of polynomial growth.

(a) *strongly–tame (S–tame in shorthand) if $\varpi(s)$ is meromorphic in $\Re(s) > 1-\delta$ for some $\delta > 0$, has only a pole (of order $k_0 \geq 0$) at $s = 1$ and is of polynomial growth in $\Re(s) > 1-\delta$ as $|s| \to +\infty$.*

(b) *Hyperbolically tame (H–tame in shorthand) if there exists an hyperbolic region $\mathcal{R}$, defined as, for some $A, B, \beta > 0$*

$$\mathcal{R} := \{s = \sigma + it; \quad |t| \geq B, \quad \sigma > 1 - \frac{A}{t^\beta}\} \bigcup \{s = \sigma + it; \quad \sigma > 1 - \frac{A}{B^\beta}, |t| \leq B\},$$

where $\varpi(s)$ has only a pole (of order $k_0 \geq 0$) at $s = 1$ and is of polynomial growth in $\mathcal{R}$ as $|s| \to +\infty$.

(c) *periodic if $\varpi(s)$ is meromorphic in $\Re(s) > 1-\delta$ for some $\delta > 0$, has a pole (of order $k_0 \geq 0$) at $s = 1$ and a family (s_k) (for $k \in \mathbb{Z}, k \neq 0$) of simple poles at points $s_k = 1 + 2ki\pi t$ with $t \neq 0$ and is of polynomial growth in $\Re(s) > 1-\delta$ as $|s| \to +\infty$*[1].

In all the cases, the integer k_0 is called the order, and, when they exist, the real δ is the abscissa, and the real β is the exponent. Figure 12 shows the three possible behaviours.

4.4. **Possible asymptotics behaviours for the mean costs.** The sequence of numerical values $\varpi(k)$ lifts into an analytic function $\varpi(s)$, whose singularities essentially determine the asymptotic behaviour of the mean costs of interest. We describe now a dictionary which transfers the analytical properties of $\varpi(s)$ near $\Re s = 1$ into asymptotic properties of the mean cost.

Proposition 5. *The following asymptotics hold for the sequence $S(n)$, when it is related to $\varpi(s)$ by the Rice formula (21):*

(a) *If $\varpi(s)$ is S–tame with order k_0 and abscissa δ_0, then, for any $\delta < \delta_0$, one has*

$$S(n) = -\operatorname{Res}\left(\frac{n!\,\varpi(-s)}{s(s-1)\cdots(s-n)}; s = -1\right) + O(n^{1-\delta})$$
$$= n\left(\sum_{k=0}^{k_0} a_k \log^k n\right) + O(n^{1-\delta}) \quad (n \to +\infty),$$

[1] More precisely, this means that $\varpi(s)$ is of polynomial growth on a family of horizontal lines $t = t_k$ with $t_k \to \infty$, and on vertical lines $\Re(s) = 1 - \delta'$ with some $\delta' < \delta$

(*b*) *If $\varpi(s)$ is H–tame with order k_0 and exponent β_0, then, for any β with $\beta < 1/(\beta_0+1)$, one has:*

$$\begin{aligned} S(n) &= -\operatorname{Res}\left(\frac{n!\,\varpi(-s)}{s(s-1)\cdots(s-n)}; s=-1\right) + n\cdot O(\exp[-(\log n)^{\beta}]) \\ &= n\left(\sum_{k=0}^{k_0} a_k \log^k n\right) + n\cdot O(\exp[-(\log n)^{\beta}]) \end{aligned}$$

(*c*) *If $\varpi(s)$ is periodic with abscissa δ_0 and order k_0, then, for any $\delta < \delta_0$, one has:*

$$\begin{aligned} S(n) &= -\sum_{k=-\infty}^{k=+\infty} \operatorname{Res}\left(\frac{n!\,\varpi(-s)}{s(s-1)\cdots(s-n)}; s=-s_k\right) + O(n^{1-\delta}) \\ &= n\left(\sum_{k=0}^{k_0} a_k \log^k n\right) + n\cdot\Phi(n) + O(n^{1-\delta}), \end{aligned}$$

where $n\cdot\Phi(n)$ is the part of the expansion brought by the family of the non real poles located on the vertical line $\Re s = 1$.

Proof. The proof relies on the residue theorem, upon pushing the contour of integration in (21) to the right, past -1. More precisely, if $\varpi(s)$ is of moderate growth in a region $\mathcal{U}$, the line of integration $\Re(s) = -d$ can be moved to the right until a curve ρ, which lies inside the region $\mathcal{U}$, with residues inside the region $\mathcal{U}$ taken into account. If $\varpi(s)$ has a pôle of order k_0 at $s = 1$, then $\varpi(-s)/(s+1)$ has a pôle of order $k_0 + 1$, and this pole contributes with a quantity of the form

$$= n\left(\sum_{k=0}^{k_0} a_k \log^k n\right).$$

In cases (*a*) or (*c*), the curve ρ can be chosen as a vertical line of equation $\sigma + 1 = \delta$ with $\delta < \delta_0$. In case (*b*), the curve ρ can be chosen as the curve of equation $\sigma + 1 = (A/2)t^{-\beta_0}$. The remainder of the proof is devoted to the computation of the integral

$$\int_\rho \varpi(-s)\frac{n!}{s(s+1)\cdots(s+n)}\,ds = \int_\rho \varpi(-s)\frac{\Gamma(n+1)\Gamma(s)}{\Gamma(n+1+s)}\,ds$$

More precisely, we wish to prove that, if $\varpi(-s)$ is of polynomial growth on the curve ρ as $|s| \to \infty$, this integral is $O(n^{1-\delta})$ in cases (*a*) and (*c*) and of order $O(n\exp[-(\log n)^{\beta}])$ with $\beta < 1/(1+\beta_0)$ in case (*b*). This part of the proof can be found in the appendix. □

In our analyses, the main Dirichlet series of interest are closely related to two Dirichlet series, $\Lambda(s)$ or $\Pi(s)$ which are both expressed with the fundamental probabilities p_w of the source:

$$\Lambda(s) := \sum_{w\in\Sigma^\star} p_w^s,$$

$$\Pi(s) := \sum_{k=0}^{\infty} \pi_k^s, \qquad \text{with} \quad \pi_k := \sup\{p_w;\;\; w\in\Sigma^k\}..$$

There are two types of problems: for `QuickSelect` and `QuickVal`, the last analytic step will be based on analytical properties of the function $\Pi(s)$, whereas, for all the other analyses, this step will be based on analytical properties of the function $\Lambda(s)$. Many probabilistic properties of the source can be expressed as regularity properties of these functions $\Lambda(s)$ or $\Pi(s)$, when $\Re s$ is close to 1.

4.5. **Π-tame sources.** This type of sources will intervene in the analysis of the algorithms `QuickSelect` and `QuickVal`.

Definition 8. *A source is Π-tame of abscissa δ (with $\delta > 0$) if the function $\Pi(s)$ is S–tame, of order $k_0 = 0$ and of abscissa δ.*

A sufficient condition for a source to be Π-tame is : There exist constants (A, γ) with $A > 0, \gamma > 1$ for which $\pi_k \leq Ak^{-\gamma}$. In this case, the abscissa δ satisfies $\delta < 1 - (1/\gamma)$. This condition is weak, so that most of the natural sources are Π–tame.
The following result describes the nice analytic behaviour of the mixed Dirichlet series of `QuickVal` as soon as the source is Π–tame:

Proposition 6. *The mixed series $\varpi(s)$ of $\texttt{QuickVal}_\alpha$ relative to a Π–tame source with abscissa δ_0 is analytic and bounded in $\Re(s) \geq 1 - \delta$ where δ is any strictly positive real which satisfies $\delta < \delta_0$*

4.6. **Λ-tame sources.** This type of sources will intervene in all the other analyses, where the function $\varpi(s)$ is closely related to $\Lambda(s)$.
In the following definition, we describe possible behaviours of this function Λ.

Definition 9. [Tame Sources.] *A source is*

(*a*) *tame if the function $s \mapsto \Lambda(s)$ is analytic on $\Re s > 1$ and meromorphic on $\Re s \geq 1$.*
(*b*) *entropic, if the function $s \mapsto \Lambda(s)$ admits at $s = 1$ a simple pole, with a residue equal to $1/h(\mathcal{S})$, where $h(\mathcal{S})$ is the entropy of the source already defined in (2)*

$$h(\mathcal{S}) := \lim_{k\to\infty} \frac{-1}{k} \sum_{w\in\Sigma^k} p_w \log p_w = -\lim_{k\to\infty} \frac{-1}{k} \frac{d}{ds} \Lambda^{(k)}(s)_{|s=1}$$

A source, which is tame and entropic is

(*c*) *S–tame if $\Lambda(s)$ is S–tame.*
(*d*) *H–tame if $\Lambda(s)$ is H–tame.*
(*e*) *Λ–periodic if $\Lambda(s)$ is periodic*

The following two propositions show that all the simple sources fulfill properties (*a*) and (*b*). They also discuss the case of a subclass of dynamical sources, the dynamical sources of the Good Class. See the appendix.

Proposition 7. *Any simple source (memoryless source or irreducible aperiodic Markov chain) or any dynamical source of the Good Class is tame and entropic.*

Proof. We study the first two properties of the previous definition for simple sources

(a) *Tame.* First, for any $\gamma > 0$, and for any s with $\Re s \geq 1 + \gamma$, one has

$$|\lambda(s)| \leq \lambda(1+\gamma) < 1, \qquad \|P^k(s)\| \leq \|P^k(1+\gamma)\| \qquad \text{so that} \quad r(s) \leq r(1+\gamma)$$

where $r(s)$ denotes the spectral radius of the matrix $P(s)$. Furthermore, the spectral radius $r(1+\gamma)$ is equal to the dominant eigenvalue $\lambda(1+\gamma)$, since the matrix $P(1+\gamma)$ is positive, irreducible and aperiodic. Since the matrix $P(1+\gamma)$ satisfies $P(1+\gamma) \leq p^\gamma P(1)$, with $p := \min\{p_{ij}, p_{i,j} > 0\}$, the inequality $r(1+\gamma) \leq p^\gamma < 1$ holds.

Since the functions $s \mapsto \lambda(s)$ (for memoryless sources) or $s \mapsto P(s)$ (for Markov chain) define analytic functions, the function $\Lambda(s)$ is meromorphic, with a set of poles, equal to

$$\mathcal{Z} = \{s; \quad 1 - \lambda(s) = 0\} \quad \text{(memoryless case)}$$

$$\mathcal{Z} = \{s; \quad \det(I - P(s)) = 0\} \quad \text{(Markov chain).}$$

(b) *Entropic.* For any real s, the matrix $P(s)$ is positive, irreducible and aperiodic, and it has a dominant eigenvalue denoted by $\lambda(s)$. Furthermore, in both cases (memoryless source or Markov chain), the derivative map $\lambda'(1)$ is not zero, and the equality $\lambda(1) = 1$ holds.

□

4.7. **Tameness of simple sources.** Always for simple sources, we now describe the possible location of poles of the function Λ and we are more precisely interested in the possible existence of regions (of hyperbolic shape) which contain no other poles than the pole $s = 1$.

Proposition 8. *The simple sources may have three different possible behaviours. They are never S–tame, but they may be Λ–periodic, or H–tame. There exists a precise characterization for each situation.*

Proof. To a family of probabilities $\mathfrak{P} = (p_1, p_2, \ldots, p_r)$, we associate the *logarithms* $w_i := |\log p_i|$ and, for each pair (k, j) which satisfies $1 \leq k, j \leq r$, the *ratio* $\alpha_{k,j} := w_j/w_k$. The following classical result proves that a memoryless source of probabilities $\mathfrak{P}$ is Λ-periodic if and only all the real numbers $\alpha_{k,j}$ are rational. More precisely

The following conditions are equivalent:

(a) *All the real numbers $\alpha_{k,j}$ are rational*
(b) *The intersection $\mathcal{Z} \bigcap \{\Re s = 1\}$ contains a point $s \neq 1$.*
(c) *There exists $\tau > 0$ for which the following equality holds*

$$\mathcal{Z} \bigcap \{\Re s = 1\} = 1 + i\tau\mathbb{Z}$$

(d) *The function $\lambda(s)$ is periodic of period $i\tau$.*

The classification periodic/aperiodic depends on the arithmetic properties of $\mathfrak{P}$.

- A memoryless source is periodic if and only there exists an algebraic integer $a < 1$ for which all the probabilities p_i belong to the semi-group generated by a.

- A Markov chain irreducible and aperiodic, whose transition matrix is P, is periodic if there exists an algebraic integer a and a vector of positive reals $(\nu_1, \nu_2, \ldots \nu_r)$ for which the matrix P is written as $P = D^{-1}QD$, where D is the matrix whose diagonal is $(\nu_1, \nu_2, \ldots \nu_r)$ and all the nonzero coefficients of the matrix Q belong to the group generated by a.

We focus now on memoryless sources, defined by the vector $\mathfrak{P} = (p_1, p_2, \ldots, p_r)$, which are not Λ-periodic. The intersection $\mathcal{Z} \cap \{s;\ \Re s = 1\}$ only contains the point $s = 1$, but there exist points of $\mathcal{Z}$ which are arbitrary close to the vertical line $\Re s = 1$. This entails that a simple source is never strongly tame. In the aperiodic case, there is, amongst all the reals $\alpha_{k,j}$, at least one real $\alpha_{k,j}$ which is irrational, and it is then possible to define the irrationnality exponent of the family $\alpha_{k,j}$, which is denoted by $\mu(\mathfrak{P})$. Such an exponent measures the degree of approximability of the family α by rationals. A source for which $\mu(\mathfrak{P})$ is finite is called diophantine. In particular, an irrational number α is diophantine if there exist $\eta > 0$ and $M > 0$ for which

$$\forall (p, q) \in \mathbb{Z} \times \mathbb{N}^\star, \qquad \left| \alpha - \frac{p}{q} \right| > \frac{M}{q^{2+\eta}}.$$

The second characterisation is as follows:

A memoryless source is H–tame if and only it is diophantine. Moreover, there is a relation between the exponent β of H–tameness and the irrationnality exponent $\mu(\mathfrak{P})$: one can choose as β any real strictly greater than $2\mu(\mathfrak{P}) - 2$. □

4.8. **Tameness of dynamical sources.** There exist natural instances of dynamical sources which are S–tame, or H–tame. A dynamical source can be Λ–periodic only if it "resembles" a memoryless sources. See the appendix for definitions of these notions.

Theorem 5. *All the sources of the Good-UNI Class are S–tame. All the sources of the Good-DIOP Class are H–tame. The only sources of the Good Class which are periodic are the sources which are conjugated to sources with affine branches.*

5. Precise statements and final proofs of the main results.

We return to the initial problems –obtaining the asymptotic mean values of the main parameters of interest. We conclude the study and state the precise results.

5.1. **Analysis of the `QuickSelect` Algorithm.** We begin by analysing `QuickVal`, then we return to `QuickQuant`.

Analysis of $\texttt{QuickVal}_\alpha$. In this case, the Dirichlet series of interest is

$$\frac{\varpi(s)}{s-1} = \frac{2}{s-1} \mathcal{J}[(\max(\alpha, t) - \min(\alpha, u))^{s-2)}]\,.$$

Proposition 6 entails that $\varpi(s)$ is analytic at $s = 1$. Then, the integrand in (21) has a simple pole at $s = 1$, brought by the factor $1/(s+1)$ and Proposition 4 applies as soon as the source $\mathcal{S}$ is Π–tame. Thus, for $\delta < \delta_0$ (where δ_0 is the abscissa of Π–tameness), one has :

$$V_n^{(\alpha)} = \rho_{\mathcal{S}}(\alpha) n + O(n^{1-\delta}). \tag{23}$$

The three possible expressions of the function $(u,t) \mapsto \max(\alpha,t) - \min(\alpha,u)$ on the unit triangle give rise to three intervals of definition for the function H defined in Theorem 6 (respectively, $]-\infty,-1/2]$, $[-1/2,+1/2]$, $[1/2,\infty[$).

Analysis of $\texttt{QuickQuant}_\alpha$. The main chain of arguments connecting the asymptotic behaviours of $\texttt{QuickVal}_\alpha(n)$ and $\texttt{QuickQuant}_\alpha(n)$ is the following.

(a) The algorithms are asymptotically "similar enough": if $X_1,\dots,X_n$ are n i.i.d random variables uniform over $[0,1]$, then the *α–quantile* of set X is with high probability close to α. For instance, it is at distance at most $(\log^2 n)/\sqrt{n}$ from α with an exponentially small probability (about $\exp(-\log^2 n)$).
(b) The function $\alpha \mapsto \rho_{\mathcal{S}}(\alpha)$ is Hölder with exponent $c > \delta$.
(c) The error term in the expansion (23) is uniform with respect to α.

Theorem 6. [Analysis of QuickQuant [40] (2009)] *For any Π–tame source of abscissa δ_0, the mean number of symbol comparisons used by* $\texttt{QuickQuant}_\alpha(n)$ *satisfies, with any $\delta < \delta_0$,*

$$Q_n^{(\alpha)} = \rho_{\mathcal{S}}(\alpha)\, n + O(n^{1-\delta}), \qquad \textit{with} \quad \rho_{\mathcal{S}}(\alpha) = \sum_{w\in\Sigma^\star} p_w\, L\left(\frac{|\alpha-\mu_w|}{p_w}\right).$$

The real μ_w is the middle of the fundamental interval $\mu_w = (1/2)(p_w^{(+)} + p_w^{(-)})$.
The function L is an even function given by $L(y) = 2[1+H(y)]$, which involves a modified entropy function H expressed with $y^+ := (1/2)+y$, $y^- = (1/2)-y$ under the form

$$H(y) = \begin{cases} -(y^+\log y^+ + \ y^-\log y^-), & \textit{if } 0 \le y < 1/2 \\ 0, & \textit{if } y = 1/2 \\ y^+(\log|y^+| - \log|y^-|), & \textit{if } y > 1/2. \end{cases}$$

There are some particular cases for the constant $\rho_{\mathcal{S}}(\alpha)$.

Constants for QuickMin *$(\alpha = 0 \to \epsilon = +)$ and* QuickMax *$(\alpha = 1 \to \epsilon = -)$*

$$c_{\mathcal{S}}^{(\epsilon)} := 2\sum_{w\in\Sigma^\star} p_w\left[1 - \frac{p_w^{(\epsilon)}}{p_w}\log\left(1+\frac{p_w}{p_w^{(\epsilon)}}\right)\right].$$

Constant for QuickRand $\underline{c}_{\mathcal{S}} = \displaystyle\int_0^1 \rho_{\mathcal{S}}(\alpha)d\alpha$

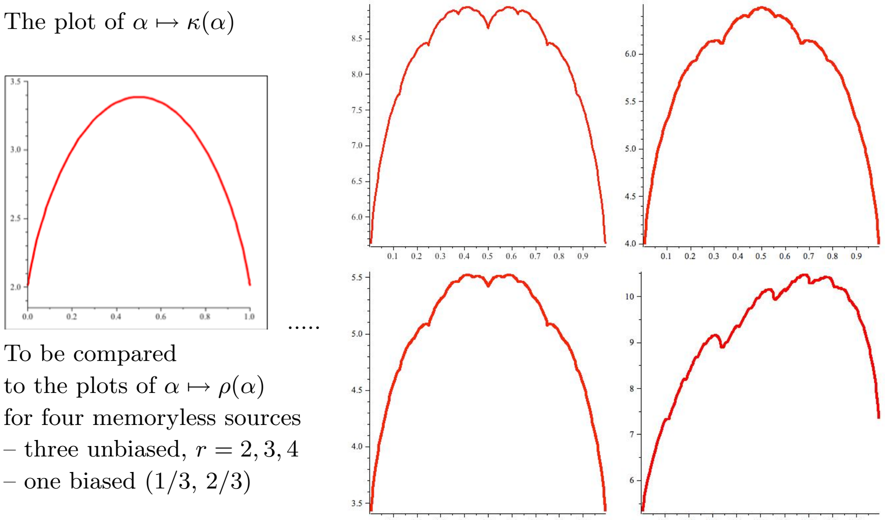
$$\underline{c}_{\mathcal{S}} = \sum_{w\in\Sigma^\star} p_w^2\left[2 + \frac{1}{p_w} + \sum_{\epsilon=\pm}\left[\log\left(1+\frac{p_w^{(\epsilon)}}{p_w}\right) - \left(\frac{p_w^{(\epsilon)}}{p_w}\right)^2\log\left(1+\frac{p_w}{p_w^{(\epsilon)}}\right)\right]\right].$$

The constants of the analysis for the binary source are

$$c_{\mathcal{B}}^{(+)} = c_{\mathcal{B}}^{(-)} = c_{\mathcal{B}}^{(\epsilon)}$$

$$c_{\mathcal{B}}^{(\epsilon)} = 4 + 2\sum_{\ell\ge 0}\frac{1}{2^\ell} + 2\sum_{\ell\ge 0}\frac{1}{2^\ell}\sum_{k=1}^{2^\ell-1}\left[1 - k\log\left(1+\frac{1}{k}\right)\right]$$

$$\underline{c}_{\mathcal{B}} = \frac{14}{3} + 2\sum_{\ell=0}^{\infty}\frac{1}{2^{2\ell}}\sum_{k=1}^{2^\ell-1}\left[k+1+\log(k+1) - k^2\log\left(1+\frac{1}{k}\right)\right]$$



The plot of $\alpha \mapsto \kappa(\alpha)$

.....

To be compared
to the plots of $\alpha \mapsto \rho(\alpha)$
for four memoryless sources
– three unbiased, $r = 2, 3, 4$
– one biased (1/3, 2/3)

FIGURE 13. The curve $\alpha \mapsto \rho(\alpha)$ is a fractal deformation of the curve $\alpha \mapsto \kappa(\alpha)$, where $\kappa(\alpha)$ is the constant relative to the number of key–comparisons in $\texttt{QuickQuant}_\alpha$.

$$\text{Numerically,} \qquad c_{\mathcal{B}}^{(\epsilon)} = 5.27937......, \qquad c_{\mathcal{B}} = 8.20731......$$

To be compared to the constants of the number of key–comparisons $\kappa = 2$ or $\kappa = 3$.

We now conclude our discussion for tree parameters : size of a `Trie`, and various path lengths for Trie and the path length for `Bst`. We first begin by the three parameters –size, path-length of an array–trie, path-length of a Bst – whose series $\varpi(s)$ directly expresses with the Dirichlet series $\Lambda(s)$.

5.2. **Analysis of the parameters for `Trie` and `Bst`: Size and (plain) path-lengths.** Recall that in these cases, the Dirichlet series $\varpi(s)$ of interest are

$$(s-1)\Lambda(s), \qquad s\Lambda(s), \qquad 2\frac{\Lambda(s)}{s(s-1)}.$$

Then, if the source is tame and entropic, the integer k_0 equal 0 for the size R, 1 for the path length C of the array–trie, and finally 2 for the path-length B of the BST. Moreover, the dominant term in the asymptotic expansion equals

$$\frac{1}{h(\mathcal{S})}n, \qquad \frac{1}{h(\mathcal{S})}\, n\log n, \qquad \frac{1}{h(\mathcal{S})}\, n\log^2 n.$$

If the source is Λ-periodic, the other poles located on $\Re s = 1$ provide the periodic term $nP(\log n)$, in the form of a Fourier series. This term adds to the dominant term for the size R, whereas it intervenes in a subdominant term for the two path-lengths C(for the array-trie) and B (for the BST). This leads to the following result:

Theorem 7. [Analysis of the size of a trie, the path-length of the array–trie, the path-length for the BST, [7] (2001) – [40] (2009)] *Consider n words independently drawn from a Λ-tame source $\mathcal{S}$ with entropy $h(\mathcal{S})$. Then, the mean size $R(n)$ of* *the trie, the mean path-length $C(n)$ of the trie, the mean symbol-path-length $B(n)$ of the BST satisfy, for some constants a, b, c which depend on the source*

$$R(n) = \frac{1}{h(\mathcal{S})}\, n + R_1(n) \qquad C(n) = \frac{1}{h(\mathcal{S})}\, n \log n + a\, n + C_1(n)$$

$$B(n) = \frac{1}{h(\mathcal{S})}\, n \log^2 n + b\, n \log n + c\, n + B_1(n).$$

and the remainder terms $R_1(n), C_1(n), B_1(n)$ satisfy the following:

- (a) *If the source is S-tame, with abcsissa δ_0, then, they are of order $O(n^{1-\delta})$, for any $\delta < \delta_0$.*
- (b) *If the source is H-tame, with exponent β_0, then they are are of order $O(n \cdot \exp[-(\log n)^{\beta}])$, for any $\beta < 1/(1+\beta_0)$.*
- (c) *If the source is Λ-periodic, with abcsissa δ_0, then, they are of the form $n \cdot P(\log n) + O(n^{1-\delta})$, for any $\delta < \delta_0$. Here $P(u)$ is a periodic function of small amplitude.*

5.3. **Analysis of the `Trie` parameters: path-lengths of the list–trie and the bst–trie.** Here, Theorem 4 provides the exact expressions of the mixed Dirichlet series

$$\varpi_L(s) \;=\; \sum_{w \in \Sigma^*} p_w^s\; K_L(s,w) \qquad \text{with} \quad K_L(s,w) = \sum_{i \in \Sigma} P_{[>i]|w}\; p_{i|w}^{s-1}$$

$$\varpi_A(s) \;=\; 2 \sum_{w \in \Sigma^*} p_w^s\; K_A(s,w) \qquad \text{with} \quad K_A(s,w) = \sum_{\substack{(i,j) \in \Sigma^2 \\ i<j}} p_{i|w}\, p_{j|w}\; P_{[i,j]|w}^{s-2}$$

An interesting (and easy) case arises when the source is stationary.

Definition 10. *A source is stationary if for any prefixes (w, w') the probabilities $p_{w'|w}$ do not depend on w and are equal to $p_{w'}$.*

A memoryless case, and a Markov chain where the initial probability is chosen as the eigenvector of the matrix P are stationary. More generally, a dynamical source of the Good Class is stationary if one chooses as initial density the unique density which is fixed by the density transformer.

In this case, the expressions $K_L(s,w)$ and $K_A(s,w)$ do not depend on w, are denoted by $K_L(s)$ and $K_A(s)$ and at $s = 1$, one has

$$K_L(1) = \sum_{i \in \Sigma} P_{[>i]}, \qquad K_A(1) = 2\frac{p_i p_j}{P_{[i,j]}}$$

and the mixed Dirichlet series $\varpi_L(s)$ and $\varpi_A(s)$ satisfy near $s = 1$

$$\varpi_L(s) \sim\; K_L(1) \cdot \Lambda(s), \qquad \varpi_A(s) \sim K_A(1) \cdot \Lambda(s).$$

In the case when the dynamical source belongs to the Good Class, and, even if the initial density is not chosen as the invariant density φ, if one lets

$$\widehat{K}_L(1) := \sum_{i \in \Sigma} \widehat{P}_{[>i]}, \qquad \widehat{K}_A(1) = 2\frac{\widehat{p}_i \widehat{p}_j}{\widehat{P}_{[i,j]}} \qquad \text{with} \quad \widehat{p}_w = \int_{\mathcal{I}_w} \varphi(t) dt,$$

the mixed Dirichlet series $\varpi_L(s)$ and $\varpi_A(s)$ satisfy near $s = 1$

$$\varpi_L(s) \sim \widehat{K}_L(1) \cdot \Lambda(s), \qquad \varpi_A(s) \sim \widehat{K}_A(1) \cdot \Lambda(s).$$

Theorem 8. [Analysis of path-lengths of the hybrid tries : list-trie and bst-trie [7](2001)] *Consider n words independently drawn from a Λ-tame source $\mathcal{S}$ with entropy $h(\mathcal{S})$. Assume moreover that the source is stationary. Then, the mean path-lengths of the hybrid tries, $L(n)$ for the list-trie, $A(n)$ for the bst-trie satisfy*

$$L(n) = \frac{K_L(\mathcal{S})}{h(\mathcal{S})}\, n\log n + b_L n + L_1(n) \qquad A(n) = \frac{K_A(\mathcal{S})}{h(\mathcal{S})}\, n\log n + b_A n + C_1(n).$$

Here, the constants $K_L(\mathcal{S})$ and $K_A(\mathcal{S})$ are defined as

$$K_L(\mathcal{S}) = \sum_{i\in\Sigma} P_{[>i]}, \qquad K_A(\mathcal{S}) = 2\frac{p_i p_j}{P_{[i,j]}}$$

the constants b_L, b_A depend both on the source and the structure, and the remainder terms $L_1(n), A_1(n)$ satisfy the following:

- (a) *If the source is S-tame, with abcsissa δ_0, then, they are of order $O(n^{1-\delta})$, for any $\delta < \delta_0$.*
- (b) *If the source is H-tame, with exponent β_0, then they are are of order $O(n \cdot \exp[-(\log n)^{\beta}])$, for any $\beta < 1/(1+\beta_0)$.*
- (c) *If the source is Λ-periodic, with abcsissa δ_0, then, they are of the form $n \cdot P(\log n) + O(n^{1-\delta})$, for any $\delta < \delta_0$. Here $P(u)$ is a periodic function of small amplitude.*

6. Appendix

We gather in the appendix some technical results about the Rice integral (useful for the proof of Proposition 5. We also describe results about dynamical sources of the Good Class.

6.1. **Estimates for the Rice integral.**

This section is concerned with two estimates of the quantity

$$\frac{n!}{s(s+1)\cdots(s+n)}, \tag{24}$$

which are central in the analysis of Nörlund–Rice integrals[2].

Lemma 3. *For s outside of a fixed sector containing the negative real axis in its interior, and under the condition*

$$|s| \le \sqrt{n},$$

one has, as $n \to \infty$:

$$\frac{n!}{s(s+1)\cdots(s+n)} = n^{-s}\Gamma(s)\left(1 + O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{s^2}{n}\right)\right). \tag{25}$$

Also. for any s fixed with $s \notin \mathbb{Z}_{\ge 0}$, one has

$$\frac{n!}{s(s+1)\cdots(s+n)} = n^{-s}\Gamma(s)\left(1 + O\left(\frac{1}{n}\right)\right). \tag{26}$$

[2] Thanks to Philippe Flajolet for this subsection.

Proof. As it is well known, Stirling's formula holds in the complex plane, provided a sector around the negative real axis is avoided.Under this condition, one has

$$\Gamma(w+1) = w^w e^{-w}\sqrt{2\pi w}\left(1+O\left(\frac{1}{n}\right)\right), \qquad |w| \to +\infty. \tag{27}$$

Regarding (25), we have

$$\frac{n!}{s(s+1)\cdots(s+n)} = \frac{\Gamma(n+1)}{\Gamma(s+n+1)}\Gamma(s), \tag{28}$$

so that it suffices to study the first factor of (28). By Stirling:

$$\begin{aligned}\frac{\Gamma(n+1)}{\Gamma(s+n+1)} &= \frac{n^n e^{-n}\sqrt{2\pi n}}{(s+n)^{s+n}e^{-s-n}\sqrt{2\pi(s+n)}}\left(1+O\left(\frac{1}{n}\right)\right)\\ &= \exp\left[n\log n-(s+n)\log(s+n)-s\right]\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)\\ &= \exp\left[-s\log n-(s+n)\log(1+s/n)-s\right]\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right).\end{aligned}$$

In the region under consideration, we have $s/n = O(1/\sqrt{n})$, which is a small quantity, so that $\log(1+s/n) = s/n + O(s^2/n^2)$. Consequently,

$$\begin{aligned}\frac{\Gamma(n+1)}{\Gamma(s+n+1)} &= n^{-s}\exp\left[O\left(\frac{s^2}{n}\right)\right]\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)\\ &= n^{-s}\left(1+O\left(\frac{1}{\sqrt{n}}\right)+O\left(\frac{s^2}{n}\right)\right),\end{aligned}$$

and the estimate (25) results. The proof of (26) is similar, even simpler, via the relation $s/n = O(1/n)$. □

The asymptotic study requires estimates valid for *both* large n and large $|s|$, this in a way that should exhibit a suitable decay in both variables. We state:

Lemma 4. *Fix any number $m > 0$. Then there exists a computable constant $K_m > 0$ such that for n large enough, $s = b + it$, b fixed and $t \geq \sqrt{n}$, one has*

$$\frac{n!}{s(s+1)\cdots(s+n)} \leq \frac{K_m}{t^m}e^{-L\sqrt{n}},$$

with $L = \log(2/\sqrt{3}) \doteq 0.14384$.

Proof. The proof is done for $b = 0$, but it extends to any value b fixed. Choose an integer $m > 0$ and set

$$A = \lfloor\sqrt{n}\rfloor.$$

We write

$$\frac{n!}{s(s+1)\cdots(s+m)} = \frac{1}{s}\prod_{a=1}^{m}\frac{a}{s+a}\prod_{a=m+1}^{m+A}\frac{a}{s+a}\prod_{a=m+A+1}^{n}\frac{a}{s+a}. \tag{29}$$

For the first product in (29), we have by trivial bounds:

$$\left|\prod_{a=1}^{m}\frac{a}{s+a}\right| \leq \frac{m!}{t^m}. \tag{30}$$

For the second product in (29), we consider the right triangle with vertices at $-a, 0, s$. The angle at a varies from nearly $\pi/2$ when a is close to the lower bound $m+1$ to nearly $\pi/4$ when a is close to its upper limit $m+A$; at any rate, this angle is, for n large enough, at least $\pi/6$. Consequently, we have

$$\left|\frac{a}{s+a}\right| \leq \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2},$$

resulting in

$$\left|\prod_{a=m+1}^{m+A} \frac{a}{s+a}\right| < \left(\frac{\sqrt{3}}{2}\right)^{A}. \tag{31}$$

For the third product in (29), we plainly use the triangle inequality, which gives $|a/(s+a)| < 1$ and

$$\left|\prod_{a=m+A+1}^{n} \frac{a}{s+a}\right| < 1. \tag{32}$$

The collection of the three bounds (30), (31), and (32) yields the statement with, additionally, $K_m = \frac{3}{2}m!$. □

Here is finally a consequence of this estimates regarding Nörlund–Rice integrals applied to functions of at most polynomial growth.

Proposition 9. (*i*) *Consider a vertical line $\Re(s) = \alpha$ with $\alpha \notin \mathbb{Z}_{\leq 0}$ and assume that $\omega(s)$ be continuous on $\Re(s) = \alpha$ and be of at most polynomial growth there: $\omega(s) = O(s^r)$ as $|s| \to \infty$ on $\Re(s) = \alpha$. Then, the integral on the vertical $\Re s = \alpha$ admits the following estimate, as $n \to \infty$,*

$$\int_{\Re s=\alpha} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds = O\left(n^{-\alpha}\right).$$

(*ii*) *Consider a curve ρ of hyperbolic type, namely of the form*

$$\rho := \{s = \sigma + it;\quad |t| \geq B,\quad \sigma + 1 = \frac{A}{t^{\beta_0}}\}\bigcup\{s = \sigma + it;\quad \sigma + 1 = \frac{A}{B^{\beta_0}}, |t| \leq B\},$$

for some strictly positive constants (A, B, β_0) and assume that $\omega(s)$ be continuous on ρ and be of at most polynomial growth there: $\omega(s) = O(s^r)$ as $|s| \to \infty$. Then the integral on the curve ρ admits the following estimate, as $n \to \infty$,

$$\int_\rho \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds = n \cdot O\left(\exp[-(\log n)^\beta\right), \qquad \textit{with} \quad \beta < \frac{1}{1+\beta_0}$$

Proof. We only need to consider the upper half-plane. We use $T = \sqrt{n}$ as a cut-off point and decompose earch of the curves –the vertical line or the hyperbolic curve ρ– into two parts.

Let us begin with the case of the vertical line $\Re s = \alpha$, and decompose

$$\int_\alpha^{\alpha+i\infty} = \int_\alpha^{\alpha+iT} + \int_{\alpha+iT}^{\alpha+i\infty}.$$

First, near the real axis, Lemma 3 applies to give

$$\int_{\alpha}^{\alpha+iT} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds = \int_{\alpha}^{\alpha+iT} n^{-s}\Gamma(s)\omega(s)\left(1+O(n^{-1})\right)ds. \tag{33}$$

Taking into account the fact that $|n^{-s}| = n^{-\alpha}$, the last integral is $O(n^{-\alpha})$, given the fast decay of $\Gamma(s)$ which more than compensates for the polynomial growth of $\omega(s)$.

Second, near imaginary infinity, Lemma 4 gives

$$\begin{aligned} \left|\int_{\alpha+iT}^{\alpha+i\infty} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds\right| &< K_m\int_T^{\infty} O(t^r)\cdot O(t^{-m})\cdot e^{-L\sqrt{n}}dt \\ &= O\left(e^{-L\sqrt{n}}\right), \end{aligned} \tag{34}$$

for n large enough, provided m has been chosen such that $m > r+2$. The combination of (36) and (35) yields the claimed estimate in the case of the vertical line.

Consider now the case of an hyperbolic curve, and consider the two parts of the curve ρ_T^- (near the real axis) and ρ_T^+ (near imaginary infinity). In the case of the curve ρ_T^+, which can be compared to a vertical line, Lemma 4 gives

$$\begin{aligned} \left|\int_{\rho_T^+} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds\right| &< K_m\int_T^{\infty} O(t^r)\cdot O(t^{-m})\cdot e^{-L\sqrt{n}}dt \\ &= O\left(e^{-L\sqrt{n}}\right), \end{aligned} \tag{35}$$

for n large enough, provided m has been chosen such that $m > r+2$.

Now, near the real axis, Lemma 3 applies to give

$$\int_{\rho_T^-} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds = \int_{\rho_T^-} n^{-s}\Gamma(s)\omega(s)\left(1+O(n^{-1})\right)ds. \tag{36}$$

Taking into account the fact that (we let $s := \sigma + it$)

$$|n^{-s}| = n^{-\sigma} = n\exp[-At^{-\beta_0}], \qquad |\varpi(s)\Gamma(s)| \le \exp[-Kt],$$

for some $K > 0$, given the fast decay of $\Gamma(s)$ which more than compensates for the polynomial growth of $\omega(s)$. If we let $L := \log n$, the modulus of the integrand is at most $n\exp[-Kt - ALt^{-\beta_0}]$. When n (and then L) is fixed, the maximum of this function is reached for $t = O(L^{1/(1+\beta_0)})$ and the maximum is of order $\exp[-(\log n)^{\beta}]$ with $\beta < 1/(1+\beta_0)$. Using the same principles as in the Laplace method, we obtain the estimate

$$\int_{\rho_T^-} \omega(s)\frac{n!}{s(s+1)\cdots(s+n)}ds = nO(\exp[-(\log n)^{\beta}]) \qquad \text{with} \quad \beta < 1/(1+\beta_0)$$

□

6.2. **Tameness of dynamical sources of the Good Class. Properties of the Good-UNI Class and Good-DIOP Class.** Here, we consider particular *complete* dynamical systems, for which it will be possible to prove that the quasi-inverse has nice spectral properties. This will entail nice properties on the function $\Lambda(s)$, from which one deduces that the associated dynamical source will be S– tame.

We first define the Good Class:

Definition 11. [Good Class] *A dynamical system of the interval $(\mathcal{I}, T)$ belongs to the good class if it is complete, with a set $\mathcal{H}$ of inverse branches which satisfies the following:*

$(G1)$ *The set $\mathcal{H}$ is uniformly contracting, i.e., there exists a constant $\rho < 1$, for which*

$$\forall h \in \mathcal{H}, \quad \forall x \in \mathcal{I}, \qquad |h'(x)| \le \rho.$$

$(G2)$ *There is a constant $A > 0$, so that every inverse branch $h \in \mathcal{H}$ satisfies $|h''| \le A|h'|$.*

$(G3)$ *There exists $\sigma_0 < 1$ for which the series $\sum_{h\in\mathcal{H}} \beta_h^s$ converges on $\Re s > \sigma_0$.*

The bounded distortion property $(G2)$ and the property $(G3)$ are always fulfilled for a finite alphabet Σ. Properties $(G1)$ and $(G2)$ together imply the existence of a constant $K > 0$ for which the following inequalities are true for all $x, y \in \mathcal{I}$ and all $h \in \mathcal{H}^\star$:

$$|h''(x)| \le K|h'(x)| \qquad |h'(x)| \le K|h'(y)| \qquad \left|\frac{h(x) - h(y)}{x - y}\right| \le K|h'(x)|. \tag{37}$$

When the dynamical system belongs to the Good Class, these operators admit dominant spectral properties for s near the real axis, together with a spectral gap. This implies that, for s near 1, the function $\Lambda(s)$ is meromorphic for s with a small imaginary part, and admits a simple pôle at $s = 1$. This proves that any system of the Good Class gives rise to a source, which is both tame and entropic.

The UNI Condition. We first define a probability $\Pr_n$ on each set $\mathcal{H}^n \times \mathcal{H}^n$, in a natural way: we let

$$\Pr_n\{(h, k)\} := |h(\mathcal{I})| \cdot |k(\mathcal{I})|,$$

where $|\mathcal{J}|$ denotes the length of the interval $\mathcal{J}$. Furthermore, $\Delta(h, k)$ denotes the "distance" between two inverse branches h and k of same depth, defined as

$$\Delta(h, k) = \inf_{x\in\mathcal{I}} |\Psi'_{h,k}(x)| \qquad \text{with} \quad \Psi_{h,k}(x) = \log\left|\frac{h'(x)}{k'(x)}\right|. \tag{38}$$

The distance $\Delta(h, k)$ is a measure of the difference between the "form" of the two branches h, k. The UNI Condition, stated as follows, is a geometric condition which expresses that the probability that two inverse branches have almost the same form is very small:

Definition 12. [Condition *UNI*] *A dynamical system $(\mathcal{I}, T)$ satisfies the UNI condition if its set $\mathcal{H}$ of inverse branches satisfies the following*

$(U1)$ *For any $a \in]0, 1[$, and for any integer n, one has $\Pr_n[\ \Delta \le \rho^{an}] \ll \rho^{an}$.*

$(U2)$ *Each $h \in \mathcal{H}$ is of class $\mathcal{C}^3$ and for each integer n, there exists B_n for which $|h'''| \le B_n|h'|$ for any $h \in \mathcal{H}^n$.*

Dynamical sources with affine branches and the UNI Condition. A source with affine branches never satisfies the Condition *UNI* : in this case, the "distance" Δ is always zero, and the probabilities of Assertion $(U1)$ are all equal to 1. Conversely, a dynamical source of the Good Class which satisfies the condition *UNI* cannot be conjugated to a source with affine branches, as it is proven by Baladi and Vallée.

Then, the condition *UNI* excludes all the simple sources, which cannot be S–tame. The strength of the Condition *UNI* is due to the fact that this condition is sufficient to imply strong tameness :

Theorem 9. [Dolgopyat, Baladi-Vallée, Cesaratto-Vallée] *When the dynamical system of the Good Class satisfies the condition* UNI*, it gives rise to a source which is S–tame.*

There are natural instances of sources that belong to the *Good-UNI* Class, for instance the Euclidean dynamical system defined in (6), together with two other dynamical systems, of the Euclidean type.

The Diophantine condition. The *Good-UNI* Class gathers systems which are quite different from systems with affine branches. The *DIOP* Condition "copies" the behaviour of memoryless sources, when they are H–tame. In this case, we recall that the quotients

$$\alpha_{k,j} := \frac{\log p_i}{\log p_k}$$

define diophantine reals, namely reals whose irrationnality exponent is finite.

The *DIOP* condition is an arithmetical condition, which generalises this condition to a system of the Good Class. For an inverse branch $h^\star$, one denotes by $h^\star$ its unique fixed point (It is easy to prove that such a point exists and is unique for a system of the Good Class), by $p(h)$ its depth, and one lets, for h, k, ℓ in $\mathcal{H}^\star$,

$$a(h) := \log |h'(h^\star)|, \qquad b(h) = \frac{a(h)}{p(h)}, \quad c(h,k) = \frac{b(h)}{b(k)} \quad d(h,k,\ell) = \frac{b(h)-b(k)}{b(h)-b(\ell)}.$$

We can now state the definition of diophantine dynamical systems:

Definition 13. [2DIOP and 3DIOP] *An irrational number α is diophantine if there exist $\eta > 0$ and $M > 0$ for which*

$$\forall (p,q) \in \mathbb{Z} \times \mathbb{N}^\star, \qquad \left| \alpha - \frac{p}{q} \right| > \frac{M}{q^{2+\eta}}.$$

A dynamical source is 2– diophantine ([2DIOP] in shorthand) f there exist two branches h et k of $\mathcal{H}^\star$ for which the ratio $c(h,k)$ is diophantine .
A dynamical source is 3– diophantine ([3DIOP] in shorthand) f there exist three branches h, k and ℓ of $\mathcal{H}^\star$ for which the ratio $d(h,k,\ell)$ is diophantine

These conditions are sufficient to entail H–tameness of associated sources:

Theorem 10. [Dologopyat-Naud-Melbourne- Roux - Vallée] *A dynamic system of the Good Class, with a finite number of branches, which is moreover* 2DIOP *gives rise to a H–tame source. A dynamic system of the Good Class, with an infinite number of branches, which is moreover* 3DIOP *gives rise to a H–tame source.*

A little piece of history. Dolgopyat, in two seminal papers, introduces the Conditions *UNI* and *2DIOP*. He proves that, under these conditions, the quasi-inverse of the plain transfer operator has nice properties in a region on the left of the line $\{\Re s = 1\}$. When the *UNI* Condition holds, the region is a vertical strip, and when the *DIOP* Condition holds, the region is of hyperbolic type. However, he does not consider the case of an infinite number of branches, and his results are extended to this case by Baladi and Vallée for the *UNI* condition, and by Melbourne in the case of the *DIOP* condition, who introduces the *3DIOP* Condition. However, in order to deal with the Dirichlet series $\Lambda(s)$, one needs to extend the previous proofs to the secant operator. This habe been done by Cesaratto and Vallée for the *UNI* Condition, and there are works of progress of Roux and Vallée which extend the results of Dolgopyat and Melbourne to the secant operator.

References

[1] Baladi, V., and Vallée, B. Euclidean algorithms are Gaussian. *Journal of Number Theory 110* (2005), 331–386.

[2] Baladi, V., and Vallée, B. Exponential decay of correlations for surface semi-flows without finite Markov partitions. *Proceedings of the American Mathematical Society 133*, 3 (2005), 865–874 (electronic).

[3] Bentley, J., and Sedgewick, R. Fast algorithms for sorting and searching strings. In *Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* (January 1997), SIAM Press, pp. 360–369. New Orleans.

[4] Burge, W. H. An analysis of binary search trees formed from sequences of nondistinct keys. *JACM 23*, 3 (July 1976), 451–454.

[5] Clément, J. Arbres digitaux et sources dynamiques, PhD, University of Caen (2000)

[6] Cesaratto, E., and Vallée, B. Gaussian distribution of trie depth for dynamical sources. Manuscript,, 2010.

[7] Clément, J., Flajolet, P., and Vallée, B. Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica 29*, 1/2 (2001), 307–369.

[8] Devroye, L. A probabilistic analysis of the height of tries and of the complexity of triesort. *Acta Informatica*, vol 21, pp 229-237

[9] Dolgopyat, D. On decay of correlations in Anosov flows, *Annals of Mathematics* 147 (1998) 357-390.

[10] Dolgopyat, D. Prevalence of rapid mixing (I) *Ergodic Theory and Dynamical Systems* 18 (1998) 1097-1114.

[11] Delange, H. Généralisation du théorème de Ikehara. *Annales scientifiques de l'École Normale Supérieure Sér. 3 71*, 3 (1954), 213–142.

[12] Fayolle, G., Flajolet, P., and Hofri, M. On a functional equation arising in the analysis of a protocol for a multi-accessbroadcast channel. *Adv. Appl. Prob.*, 18 (1986), 441–472.

[13] Fill, J. A., and Janson, S. The number of bit comparisons used by Quicksort: An average-case analysis. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA04)* (2001), pp. 293–300.

[14] Fill, J. A., and Nakama, T. Analysis of the expected number of bit comparisons required by Quickselect. ArXiv, 2007.

[15] Flajolet, P. The ubiquitous digital tree. In *STACS 2006* (2006), B. Durand and W. Thomas, Eds., vol. 3884 of *Lecture Notes in Computer Science*, pp. 1–22. Proceedings of 23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, February 2006.

[16] Flajolet, P., Régnier, M., and Sedgewick, R. Some uses of the Mellin integral transform in the analysis of algorithms. In *Combinatorial Algorithms on Words* (1985), A. Apostolico and Z. Galil, Eds., vol. 12 of *NATO Advance Science Institute Series.* Series F: Computer and Systems Sciences, Springer Verlag, pp. 241–254. (Invited Lecture).

[17] Flajolet, P., Gourdon, X., and Dumas, P. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science 144*, 1–2 (June 1995), 3–58.

[18] Flajolet, P., and Sedgewick, R. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Computer Science 144*, 1–2 (June 1995), 101–124.

[19] Flajolet, P., and Sedgewick, R. *Analytic Combinatorics.* Cambridge University Press, 2009. Available electronically from the authors' home pages.

[20] Flajolet, P., Roux, M., and Vallée, B. Digital Trees and Memoryless Sources: from Arithmetics to Analysis, submitted

[21] Gonnet, G. H., and Baeza-Yates, R. *Handbook of Algorithms and Data Structures: in Pascal and C*, second ed. Addison–Wesley, 1991.

[22] Grabner, P., and Prodinger, H. On a constant arising in the analysis of bit comparisons in quickselect. Preprint, 2007.

[23] Jacquet, P., and Régnier, M. Trie partitioning process: Limiting distributions. In *CAAP'86* (1986), P. Franchi-Zanetacchi, Ed., vol. 214 of *Lecture Notes in Computer Science*, pp. 196–210. Proceedings of the 11th Colloquium on Trees in Algebra and Programming, Nice France, March 1986.

[24] Jacquet, P., and Szpankowski, W. Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory 37*, 5 (1991), 1470–1475.

[25] Jacquet, P., and Szpankowski, W. Analytical de-Poissonization and its applications. *Theoretical Computer Science 1-2*, 201 (1998), 1–62.

[26] Knuth, D. E. *The Art of Computer Programming*, 2nd ed., vol. 3: Sorting and Searching. Addison-Wesley, 1998.

[27] Lagarias, J. C. Best simultaneous Diophantine approximations I: Growth rates of best approximation denominators. *Transactions of the American Mathematical Society 272*, 2 (1982), 545–554.

[28] Lapidus, M. L., and van Frankenhuijsen, M. *Fractal Geometry, Complex Dimensions and Zeta Functions: Geometry and Spectra of Fractal Strings.* Springer, 2006.

[29] Mohamed, H., and Robert, P. A probabilistic analysis of some tree algorithms. *Annals of Applied Probability 15*, 4 (2005), 2445–2471.

[30] Mohamed, H., and Robert, P. Dynamic tree algorithms. *Annals of Applied Probability 20*, 1 (2010), 26–51.

[31] Nörlund, N. E. Leçons sur les équations linéaires aux différences finies. In *Collection de monographies sur la théorie des fonctions*. Gauthier-Villars, Paris, 1929.

[32] Nörlund, N. E. *Vorlesungen über Differenzenrechnung.* Chelsea Publishing Company, New York, 1954.

[33] Pittel, B. Paths in a random digital tree: limiting distributions. *Advances in Applied Probability 18*, 1 (1986), 139–155.

[34] Roux, M., and Vallée, B. Séries de Dirichlet, Théorie de l'information, et Analyse d'algorithmes, manuscript.

[35] Sedgewick, R. *Quicksort*. Garland Pub. Co., New York, 1980. Reprint of Ph.D. thesis, Stanford University, 1975.

[36] Sedgewick, R. *Algorithms in C, Parts 1–4*, third ed. Addison–Wesley, Reading, Mass., 1998.

[37] Szpankowski, W. *Average-Case Analysis of Algorithms on Sequences.* John Wiley, 2001.

[38] Vallée, B. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica 29*, 1/2 (2001), 262–306.

[39] Vallée, B. Euclidean dynamics. *Discrete and Continuous Dynamical Systems 15*, 1 (2006), 281–352.

[40] Vallée, B., Clément, J., Fill, J. A., and Flajolet, P. The number of symbol comparisons in QuickSort and QuickSelect. In *ICALP 2009, Part I* (2009), S. A. *et al.*, Ed., vol. 5555 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 750–763. Proceedings of the 36th International Colloquium on Automata, Languages and Programming.