HAL
open science

The Number of Symbol Comparisons in QuickSort and QuickSelect

Brigitte Vallée, Julien Clément, James Allen Fill, Philippe Flajolet

To cite this version:

Brigitte Vallée, Julien Clément, James Allen Fill, Philippe Flajolet. The Number of Symbol Comparisons in QuickSort and QuickSelect. 36th International Colloquium on Automata, Languages and Programming, Jul 2009, Rhodes, Greece. pp.750 - 763, 10.1007/978-3-642-02927-1_62 . hal-01082394

HAL Id: hal-01082394

https://hal.science/hal-01082394v1

Submitted on 13 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Number of Symbol Comparisons in QuickSort and QuickSelect

Brigitte Vallée[1], Julien Clément[1], James Allen Fill[2], and Philippe Flajolet[3]

[1] GREYC, CNRS and University of Caen, 14032 Caen Cedex (France)
[2] Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218-2682 (USA)
[3] Algorithms Project, INRIA-Rocquencourt, 78153 Le Chesnay (France)

Abstract 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. Our probabilistic models belong to a broad category of information sources that encompasses memoryless (i.e., independent-symbols) and Markov sources, as well as many unbounded-correlation sources. We establish that, under our conditions, the average-case complexity of `QuickSort` is $O(n \log^2 n)$ [rather than $O(n \log n)$, classically], whereas that of `QuickSelect` remains $O(n)$. Explicit expressions for the implied constants are provided by our combinatorial–analytic methods.

Introduction. Every student of a basic algorithms course is taught that, on average, the complexity of Quicksort is $O(n \log n)$, that of binary search is $O(\log n)$, and that of radix-exchange sort is $O(n \log n)$; see for instance [13,16]. 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 digital trees,

also known as "*tries*", in [3].) 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. Our results constitute broad extensions of earlier ones by Fill and Janson (`Quicksort`, [7]) and Fill and Nakama (`QuickSelect`, [8]).

The sources that we deal with include memoryless ("Bernoulli") sources with possibly *non-uniform* independent symbols and Markov sources, as well as many non-Markovian sources with unbounded memory from the past. 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. Our approach otherwise relies on methods from analytic combinatorics [10], as well as on information theory and the theory of dynamical systems. It relates to earlier analyses of digital trees ("tries") by Clément, Flajolet, and Vallée [3,19].

Results. We consider a totally ordered *alphabet* Σ. What we call a probabilistic *source* produces infinite words on the alphabet Σ (see Definition 1). The set of *keys* (words, data items) is then Σ^∞, endowed with the strict lexicographic order denoted '$\prec$'. Our main objects of study are two algorithms, applied to n keys assumed to be *independently drawn* from the same source $\mathcal{S}$, namely, the standard sorting algorithm $\texttt{QuickSort}(n)$ and the algorithm $\texttt{QuickSelect}(m, n)$, which selects the mth smallest element. In the latter case, we shall mostly focus our attention on situations where the rank m is proportional to n, being of the form $m = \alpha n$, so that the algorithm determines the αth quantile; it will then be denoted by $\texttt{QuickQuant}_\alpha(n)$. We also consider the cases where rank m equals 1 (which we call `QuickMin`), equals n (`QuickMax`), or is randomly chosen in $\{1, \ldots, n\}$ (`QuickRand`).

Our main results involve constants that depend on the source $\mathcal{S}$ (and possibly on the real α); these are displayed in Figures 1–3 and described in Section 1. They specialize nicely for a binary source $\mathcal{B}$ (under which keys are compared via their binary expansions, with uniform independent bits), in which case they admit pleasant expressions that simplify and extend those of Fill and Nakama [8] and Grabner and Prodinger [11] and lend themselves to precise numerical evaluations (Figure 2). The conditions Λ–tamed, Π–tamed, and "periodic" are made explicit in Subsection 2.1. Conditions of applicability to classical source models are discussed in Section 3.

Theorem 1. (i) *For a Λ–tamed source $\mathcal{S}$, the mean number T_n of symbol comparisons of* $\texttt{QuickSort}(n)$ *involves the entropy $h(\mathcal{S})$ of the source:*

$$T_n = \frac{1}{h(\mathcal{S})}\, n \log^2 n + a(\mathcal{S})\, n \log n + b(\mathcal{S})\, n + o(n), \quad \text{for some } a(\mathcal{S}), b(\mathcal{S}) \in \mathbb{R}. \quad (1)$$

(ii) *For a periodic source, a term $nP(\log n)$ is to be added to the estimate* (1), *where $P(u)$ is a (computable) continuous periodic function.*

Theorem 2. *For a Π–tamed source $\mathcal{S}$, one has the following, with $\delta = \delta(\mathcal{S}) > 0$.*
(i) The mean number of symbol comparisons $Q_n^{(\alpha)}$ for $\texttt{QuickQuant}_\alpha$ *satisfies*

$$Q_n^{(\alpha)} = \rho_{\mathcal{S}}(\alpha) n + O(n^{1-\delta}). \tag{2}$$

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

$$M_n^{(\epsilon)} = \rho_{\mathcal{S}}^{(\epsilon)} n + O(n^{1-\delta}), \qquad \text{with} \quad \rho_{\mathcal{S}}^{(+)} = \rho_{\mathcal{S}}(1), \qquad \rho_{\mathcal{S}}^{(-)} = \rho_{\mathcal{S}}(0). \tag{3}$$

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

$$M_n = \gamma_{\mathcal{S}}\, n + O(n^{1-\delta}), \qquad \text{with} \quad \gamma_{\mathcal{S}} = \int_0^1 \rho(\alpha) d\alpha. \tag{4}$$

These estimates are to be compared to the classical ones (see, e.g., [13, p.634] and [12], for extensions), relative to the *number of comparisons* in $\texttt{QuickQuant}_\alpha$:

$$K_n^{(\alpha)} = \kappa(\alpha) n + O(\log n), \qquad \kappa(\alpha) := 2(1 - \alpha \log(\alpha) - (1-\alpha)\log(1-\alpha)). \tag{5}$$

General strategy. 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 show (Proposition 1) that the mean number S_n of symbol comparisons performed by a (general) algorithm $\mathcal{A}(n)$ applied to n words from a source $\mathcal{S}$ can be expressed in terms of two objects: (i) the *density* ϕ_n of the algorithm, which uniquely depends on the algorithm and provides a measure of the mean number of key-comparisons performed near a specific point; (ii) the family of *fundamental triangles*, which solely depends on the source and describes the location of pairs of words which share a common prefix.

Step (b)**.** This step is devoted to computing the density of the algorithm. We first deal with the *Poisson model*, where the number of keys, instead of being fixed, follows a Poisson law of parameter Z. We provide an expression of the Poissonized density relative to each algorithm, from which we deduce the mean number of comparisons under this model. Then, simple algebra yield expressions relative to the model where the number n of keys is fixed.

Step (c)**.** 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 sorted (which tends to infinity), and the fundamental probabilities (which tend to 0). We approach the corresponding asymptotic analysis by means of *complex integral representations* of the Nörlund–Rice type. For each algorithm–source pair, 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.

1 Algebraic analysis

The developments of this section are essentially formal (algebraic) and *exact*; that is, no approximation is involved at this stage.

Entropy constant [`Quicksort`, Theorem 1]: $h(\mathcal{S}) := \lim_{k\to\infty} \Big[-\frac{1}{k} \sum_{w\in\Sigma^k} p_w \log p_w \Big]$.

Quantile constant $\rho_{\mathcal{S}}(\alpha)$ [`QuickQuant`, Theorem 2, Part (i)]

$$\rho_{\mathcal{S}}(\alpha) := \sum_{w\in\Sigma^\star} p_w\, L\left(\frac{|\alpha-\mu_w|}{p_w}\right).$$

Here, $\mu_w := (1/2)\,(p_w^{(+)} + p_w^{(-)})$ involves the probabilities $p_w^{(+)}$ and $p_w^{(-)}$ defined in Subsection 1.1; the function L is given by $L(y) := 2[1+H(y)]$, where $H(y)$ is expressed by $y^+ := (1/2)+y$, $y^- := (1/2)-y$, and

$$H(y) := \begin{cases} -(y^+\log y^+ + \ y^-\log y^-), & \text{if } 0\le y<1/2 \\ 0, & \text{if } y=1/2 \\ y^-(\log y^+ - \log|y^-|), & \text{if } y>1/2. \end{cases} \tag{6}$$

— **Min/max constants** [`QuickMin`, `QuickMax`, Theorem 2, Part (ii)]:

$$\rho_{\mathcal{S}}^{(\epsilon)} = \sum_{w\in\Sigma^\star} p_w \Big[1 - \frac{p_w^{(\epsilon)}}{p_w}\log\Big(1+\frac{p_w}{p_w^{(\epsilon)}}\Big)\Big].$$

— **Random selection constant** [`QuickRand`, Theorem 2, Part (iii)]:

$$\gamma_{\mathcal{S}} = \sum_{w\in\Sigma^\star} p_w^2 \Big[2 + \frac{1}{p_w} + \sum_{\epsilon=\pm}\Big[\log\Big(1+\frac{p_w^{(\epsilon)}}{p_w}\Big) - \Big(\frac{p_w^{(\epsilon)}}{p_w}\Big)^2 \log\Big(1+\frac{p_w}{p_w^{(\epsilon)}}\Big)\Big]\Big].$$

Figure 1. The main constants of Theorems 1 and 2, relatively to a general source $(\mathcal{S})$.

$$h(\mathcal{B}) = \log 2 \qquad \text{[entropy]}$$

$$\rho_{\mathcal{B}}^{(\epsilon)} = 4 + 2\sum_{\ell\ge 0}\frac{1}{2^\ell}\sum_{k=1}^{2^\ell-1}\Big[1 - k\log\Big(1+\frac{1}{k}\Big)\Big] \doteq 5.27937\,82410\,80958.$$

$$\gamma_{\mathcal{B}} = \frac{14}{3} + 2\sum_{\ell=0}^{\infty}\frac{1}{2^{2\ell}}\sum_{k=1}^{2^\ell-1}\Big[k+1+\log(k+1) - k^2\log\Big(1+\frac{1}{k}\Big)\Big] \doteq 8.20730\,88638.$$

Figure 2. The constants relative to a binary source.

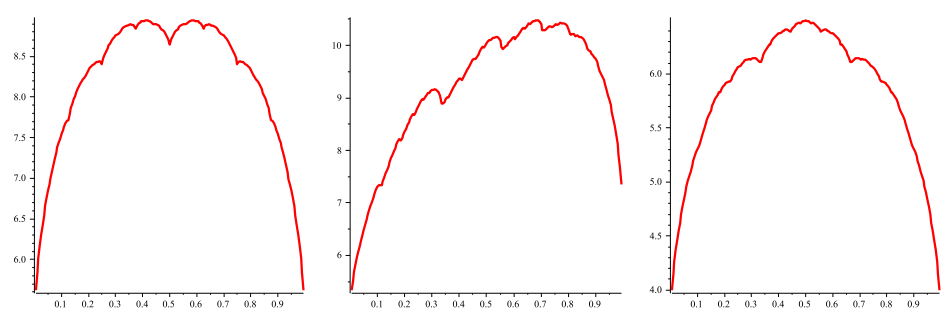

Figure 3. Plot of the function $\rho_{\mathcal{S}}(\alpha)$ for $\alpha\in[0,1]$ and the three sources: Bern$(\frac{1}{2},\frac{1}{2})$, Bern$(\frac{1}{3},\frac{2}{3})$, and Bern$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$. The curves illustrate the fractal character of the constants involved in `QuickSelect`. (Compare with $\kappa(\alpha)$ in Eq. (5), which appears as limit of $\rho_{\mathcal{S}}(\alpha)$, for an unbiased r–ary source $\mathcal{S}$, when $r\to\infty$.)

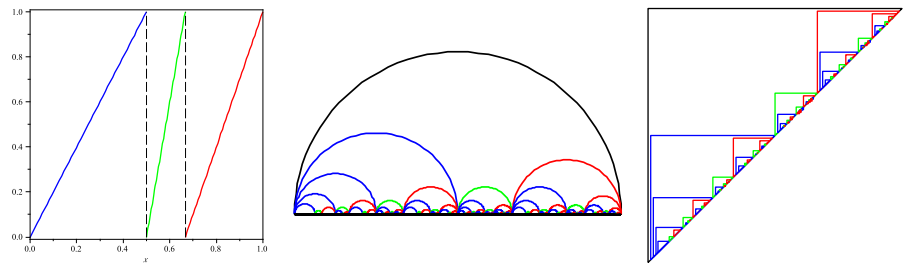

Figure 4. *Left*: the shift transformation T of a ternary Bernoulli source with $\mathbb{P}(0) = 1/2$, $\mathbb{P}(1) = 1/6$, $\mathbb{P}(2) = 1/3$. *Middle*: fundamental intervals $\mathcal{I}_w$ materialized by semicircles. *Right*: the corresponding fundamental triangles $\mathcal{T}_w$.

1.1. A general source model. Throughout this paper, a totally ordered (finite or denumerable) alphabet Σ of "symbols" or "letters" is fixed.

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 $p_w^{(-)}$, $p_w^{(+)}$, p_w the probabilities that a word produced by the source begins with a prefix w' of the same length as w, which satisfies $w' \prec w$, $w' \succ w$, or $w' = w$, respectively. Since the sum of these three probabilities equals 1, this defines two real numbers $a_w, b_w \in [0, 1]$ for which

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

Given an infinite word $X \in \Sigma^\infty$, denote by w_k its prefix of length k. The sequence (a_{w_k}) is increasing, the sequence (b_{w_k}) is decreasing, and $b_{w_k} - a_{w_k}$ tends to 0. Thus a unique real $\pi(X) \in [0, 1]$ is defined as common limit of (a_{w_k}) and (b_{w_k}). Conversely, a mapping $M : [0, 1] \to \Sigma^\infty$ associates, to a number u of the interval $\mathcal{I} := [0, 1]$, a word $M(u) := (m_1(u),\ m_2(u),\ m_3(u), \ldots) \in \Sigma^\infty$. In this way, the lexicographic order on words ('$\prec$') is compatible with the natural order on the interval $\mathcal{I}$; namely, $M(t) \prec M(u)$ if and only if $t < u$. Then, the fundamental interval $\mathcal{I}_w := [a_w, b_w]$ is the set of reals u for which $M(u)$ begins with the prefix w. Its length equals p_w.

Our analyses involve the two Dirichlet series of fundamental probabilities,

$$\Lambda(s) := \sum_{w \in \Sigma^\star} p_w^{-s}, \qquad \Pi(s) := \sum_{k \geq 0} \pi_k^{-s}; \tag{7}$$

these are central to our treatment. They satisfy, for $s < 0$, the relations $\Pi(s) \leq \Lambda(s) \leq \Pi(s + 1)$. The series $\Lambda(s)$ always has a singularity at $s = -1$, and $\Pi(s)$ always has a singularity at $s = 0$, The regularity properties of the source can be expressed in terms of Λ near $s = -1$, and Π near 0, as already shown in [19].

An important subclass is formed by *dynamical sources*, which are closely related to dynamical systems on the interval; see Figure 4 and [19]. One starts

with a partition $\{\mathcal{I}_\sigma\}$ indexed by symbols $\sigma \in \Sigma$, a coding map $\tau : \mathcal{I} \to \Sigma$ which equals σ on $\mathcal{I}_\sigma$, and a shift map $T : \mathcal{I} \to \mathcal{I}$ whose restriction to each $\mathcal{I}_\sigma$ is increasing, invertible, and of class $\mathcal{C}^2$. Then the word $M(u)$ is the word that encodes the trajectory $(u, Tu, T^2u, \ldots)$ via the coding map τ, namely, $M(u) := (\tau(u), \tau(Tu), \tau(T^2u), \ldots)$. All memoryless (Bernoulli) sources and all Markov chain sources belong to the general framework of Definition 1: they correspond to a piecewise linear shift, under this angle. For instance, the standard binary system is obtained by $T(x) = \{2x\}$ ($\{\cdot\}$ is the fractional part). Dynamical sources with a non-linear shift allow for correlations that depend on the entire past (e.g., sources related to continued fractions obtained by $T(x) = \{1/x\}$).

1.2. The algorithm $\texttt{QuickVal}_\alpha$. We consider an algorithm that is dual of `QuickSelect`: it takes as input a set of words $\mathcal{X}$ and a *value* V, and returns the *rank* of V inside the set $\mathcal{X} \cup \{V\}$. This algorithm is of independent interest and is easily implemented as a variant of `QuickSelect` by resorting to the usual partitioning phase, then doing a comparison between the *value* of the pivot and the input *value* V (rather than a comparison between their ranks). We call this modified `QuickSelect` algorithm $\texttt{QuickVal}_\alpha$ when it is used to seek the rank of the data item with value $M(\alpha)$. The main idea here is that the behaviors of $\texttt{QuickVal}_\alpha(n)$ and $\texttt{QuickQuant}_\alpha(n)$ should be asymptotically similar. Indeed, the *α–quantile* of a random set of words of large enough cardinality must be, with high probability, close to the word $M(\alpha)$.

1.3. An expression for the mean number of symbol comparisons. The first object of our analysis is the *density* of the algorithm $\mathcal{A}(n)$, which measures the number of key–comparisons performed by the algorithm. The second object is the collection of *fundamental triangles* (see Figure 4).

Definition 2. *The* density *of an algorithm $\mathcal{A}(n)$ which compares n words from the same probabilistic source $\mathcal{S}$ is defined as follows:*

$\phi_n(u,t)\,du\,dt$:= *"the mean number of key–comparisons performed by $\mathcal{A}(n)$ between words $M(u'), M(t')$, with $u' \in [u, u+du]$ and $t' \in [t, t+dt]$."*

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 $\mathcal{I}_w := [a_w, b_w]$ corresponding to w; i.e., $\mathcal{T}_w := \{(u,t) : a_w \le u \le t \le b_w\}$. We set $\mathcal{T} \equiv \mathcal{T}_\epsilon := \{(u,t) : 0 \le u \le t \le 1\}$.*

Our first result establishes a relation between the mean number S_n of symbol-comparisons, the density ϕ_n, and the fundamental triangles $\mathcal{T}_w$:

Proposition 1. *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 number S_n of symbol comparisons performed by $\mathcal{A}(n)$ is equal to

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

where ϕ_n is the density of algorithm $\mathcal{A}(n)$.

Proof. The *coincidence function* $\gamma(u,t)$ is the length of the largest common prefix of $M(u)$ and $M(t)$, namely, $\gamma(u,t) := \max\{\ell : m_j(u) = m_j(t),\ \forall j \le \ell\}$. Then, the number of symbol comparisons needed to compare two words $M(u)$ and $M(t)$, is $\gamma(u,t)+1$ and the mean number S_n of symbol comparisons performed by $\mathcal{A}(n)$ satisfies

$$S_n = \int_{\mathcal{T}} [\gamma(u,t)+1]\, \phi_n(u,t)\, du\, dt,$$

where $\phi_n(u,t)$ is the density of the algorithm. Two useful identities are

$$\sum_{\ell \ge 0} (\ell+1)\mathbf{1}_{[\gamma=\ell]} = \sum_{\ell \ge 0} \mathbf{1}_{[\gamma \ge \ell]} \qquad [\gamma \ge \ell] = \bigcup_{w \in \Sigma^\ell} (\mathcal{I}_w \times \mathcal{I}_w)\,.$$

The first one holds for *any* integer-valued random variable γ ($\mathbf{1}_A$ is the indicator of A). The second one follows from the definitions of the coincidence γ and of the fundamental intervals $\mathcal{I}_w$. Finally, the domain $\mathcal{T} \cap [\gamma \ge \ell]$ is a union of fundamental triangles. □

1.4. Computation of Poissonized densities. Under the *Poisson model* Poi_Z of parameter Z, the number N of keys is no longer fixed but rather follows a Poisson law of parameter Z:

$$\mathbb{P}[N = n] = e^{-Z} \frac{Z^n}{n!}$$

Then, the expected value $\widetilde{S}_Z$ in this model, namely,

$$\widetilde{S}_Z := e^{-Z} \sum_{n \ge 0} \frac{Z^n}{n!} S_n = e^{-Z} \sum_{n \ge 0} \frac{Z^n}{n!} \mathcal{J}[\phi_n] = \mathcal{J}\left[e^{-Z} \sum_{n \ge 0} \frac{Z^n}{n!} \phi_n \right] = \mathcal{J}[\widetilde{\phi}_Z],$$

involves the Poissonized density $\widetilde{\phi}_Z$, defined as

$$\widetilde{\phi}_Z(u,t) := e^{-Z} \sum_{n \ge 0} \frac{Z^n}{n!} \phi_n(u,t).$$

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

Proposition 2. *Set $f_1(\theta) := e^{-\theta} - 1 + \theta$. The Poissonized densities of* `QuickSort` *and* $\texttt{QuickVal}_\alpha$ *satisfy*

$$\begin{aligned} \widetilde{g}_Z(u,t) &= 2(t-u)^{-2} f_1(Z(t-u)), \\ \widetilde{g}_Z^{(\alpha)}(u,t) &= 2\left(\max(\alpha,t) - \min(\alpha,u)\right)^{-2} f_1\left(Z(\max(\alpha,t) - \min(\alpha,u)\right). \end{aligned}$$

The mean number of comparisons of `QuickSort` *and* $\texttt{QuickVal}_\alpha$ *in the Poisson model satisfy*

$$\begin{aligned} \widetilde{S}_Z &= 2\mathcal{J}[(t-u)^{-2} \cdot f_1(Z(t-u))] \\ \widetilde{S}_Z^{(\alpha)} &= 2\mathcal{J}\left[(\max(\alpha,t) - \min(\alpha,u))^{-2} \cdot f_1(Z(\max(\alpha,t) - \min(\alpha,u))\right]. \end{aligned}$$

Proof. The probability that $M(u')$ and $M(t')$ are both keys for some $u' \in [u, u+du]$ and $t' \in [t, t+dt]$ is $Z^2 du\, dt$. Denote by $[x,y] := [x(u,t), y(u,t)]$ the smallest closed interval that contains $\{u,t,\alpha\}$ ($\texttt{QuickVal}_\alpha$) or $\{u,t\}$ (`QuickSort`). Conditionally, given

that $M(u)$ and $M(t)$ are both keys, $M(u)$ and $M(t)$ are compared if and only if $M(u)$ or $M(t)$ is chosen as the pivot amongst the set $\mathcal{M} := \{M(z);\ z \in [x, y]\}$. The cardinality of the "good" set $\{M(u), M(t)\}$ is 2, while the cardinality of $\mathcal{M}$ equals $2 + N[x, y]$, where $N[x, y]$ is the number of keys strictly between $M(x)$ and $M(y)$. Then, for any fixed set of words, the probability that $M(u)$ and $M(t)$ are compared is $2/(2+N[x(u,t), y(u,t)])$. To evaluate the mean value of this ratio in the Poisson model, we remark that, if we draw Poi (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 Poi (λZ) distributed, so that

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

1.5. Exact formulae for S_n. We now return to the model of prime interest, where the number of keys is a fixed number n.

Proposition 3. *Assume that $\Lambda(s)$ of* (7) *converges at $s = -2$. Then the mean values S_n, associated with* `QuickSort` *and* $\texttt{QuickVal}_\alpha$ *can be expressed as*

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

where $\varpi(s)$ is a series of Dirichlet type given, respectively, by

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

The two functions $\varpi(s)$ are defined for $\Re s \leq -2$; they depend on the algorithm and *the source and are called the* mixed Dirichlet series.

For `QuickSort`, the series $\varpi(s)$ admits a simple form in terms of $\Lambda(s)$:

$$\varpi(s) = \frac{\Lambda(s)}{s(s+1)}, \qquad S_n = 2\sum_{k=2}^{n} \frac{(-1)^k}{k(k-1)} \binom{n}{k} \sum_{w\in\Sigma^*} p_w^k. \tag{8}$$

For `QuickVal`, similar but more complicated forms can be obtained.

Proof (Prop. 3). Consider any sequence S_n whose Poissonized version is of the form

$$\widetilde{S}_Z \equiv e^{-Z} \sum_{n\geq 0} S_n \frac{Z^n}{n!} = \int_D \lambda(x)\, f_1(Z\mu(x))\, dx, \tag{9}$$

for some domain $D \subset \mathbb{R}^d$ and some weights $\lambda(x), \mu(x) \geq 0$, with, additionally, $\mu(x) \leq 1$. By expanding f_1, then exchanging the order of summation and integration, one obtains

$$\widetilde{S}_Z = \sum_{k=2}^{\infty} (-1)^k \varpi(-k) \frac{Z^k}{k!}, \qquad \text{where} \quad \varpi(-k) := \int_D \lambda(x)\, \mu(x)^k\, dx. \tag{10}$$

Analytically, the form (10) is justified as soon as the integral defining $\varpi(-2)$ is convergent. Then, since S_n is related to $\widetilde{S}_Z$ via the relation $S_n = n![Z^n](e^Z \widetilde{S}_Z)$, it can be recovered by a binomial convolution. $\square$

2 Asymptotic analysis

For asymptotic purposes, the *singular structure* of the involved functions, most notably the mixed Dirichlet series $\varpi(s)$, is essential. With tameness assump-

tions described in Subsection 2.1, it becomes possible to develop *complex integral representations* (Subsection 2.2), themselves strongly tied to Mellin transforms [9,18]. We can finally conclude with the proofs of Theorems 1 and 2 in Subsection 2.3.

2.1. Tamed sources. We now introduce important regularity properties of a source, via its Dirichlet series $\Lambda(s), \Pi(s)$, as defined in (7). Recall that $\Lambda(s)$ always has a singularity at $s = -1$, and $\Pi(s)$ always has a singularity at $s = 0$.

Definition 3. *Consider a source $\mathcal{S}$, with $\Lambda(s), \Pi(s)$ its Dirichlet series.*

(*a*) *The source $\mathcal{S}$ is Π–*tamed *if the sequence π_k satisfies $\pi_k \leq Ak^{-\gamma}$ for some $A > 0, \gamma > 1$. In this case, $\Pi(s)$ is analytic for $\Re(s) < -1/\gamma$.*

(*b*) *The source $\mathcal{S}$ is Λ–*tamed *if $\Lambda(s)$ is meromorphic in a "hyperbolic domain"*

$$\mathcal{F} = \left\{ s \quad \middle| \quad \Re s \leq -1 + \frac{c}{(1 + |\Im s|)^d} \right\} \qquad (c, d > 0), \tag{11}$$

is of polynomial growth as $|s| \to \infty$ in $\mathcal{F}$, and has only a pole at $s = -1$, of order 1, with a residue equal to the inverse of the entropy $h(\mathcal{S})$.

(*c*) *The source $\mathcal{S}$ is Λ–*strongly tamed *if $\Lambda(s)$ is meromorphic in a half-plane*

$$\mathcal{F} = \{\Re(s) \leq \sigma_1\}, \tag{12}$$

for some $\sigma_1 > -1$, is of polynomial growth as $|s| \to \infty$ in $\mathcal{F}$, and has only a pole at $s = -1$, of order 1, with a residue equal to the inverse of the entropy $h(\mathcal{S})$.

(*d*) *The source $\mathcal{S}$ is* periodic *if $\Lambda(s)$ is analytic in $\Re(s) < -1$, has a pole of order 1 at $s = -1$, with a residue equal to the inverse of the entropy $h(\mathcal{S})$, and admits an imaginary period $i\Omega$, that is, $\omega(s) = \omega(s + i\Omega)$.*

Essentially all "reasonable" sources are Π– tamed and "most" classical sources are Λ–tamed; see Section 3 for a discussion. The properties of $\varpi(s)$ turn out to be related to properties of the source via the Dirichlet series $\Lambda(s), \Pi(s)$.

Proposition 4. *The mixed series $\varpi(s)$ of* $\texttt{QuickVal}_\alpha$ *relative to a Π–tamed source with exponent γ is analytic and bounded in $\Re(s) \leq -1+\delta$ with $\delta < 1-1/\gamma$. The mixed series $\varpi(s)$ of* `QuickSort` *is meromorphic and of polynomial growth in a hyperbolic domain* (11) *when the source is Λ–tamed, and in a vertical strip* (12) *when the source is Λ–strongly tamed.*

2.2. General asymptotic estimates. The sequence of numerical values $\varpi(-k)$ lifts into an analytic function $\varpi(s)$, whose singularities essentially determine the asymptotic behaviour of `QuickSort` and `QuickSelect`.

Proposition 5. *Let (S_n) be a numerical sequence which can be written as*

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

(*i*) *If the function $\varpi(s)$ is analytic in $\Re(s) < C$, with $-2 < C < -1$, and is of polynomial growth with order at most r, then the sequence S_n admits a*

Nörlund–Rice representation, *for* $n > r+1$ *and any* $-2 < d < C$*:*

$$S_n = \frac{1}{2i\pi} \int_{d-i\infty}^{d+i\infty} \varpi(s) \frac{n!}{s(s+1)\cdots(s+n)}\, ds. \tag{13}$$

(*ii*) *If, in addition,* $\varpi(s)$ *is meromorphic in* $\Re(s) < D$ *for some* $D > -1$*, has only a pole (of order* $k_0 \geq 0$*) at* $s = -1$ *and is of polynomial growth in* $\Re(s) < D$ *as* $|s| \to +\infty$*, then* S_n *satisfies* $S_n = A_n + O(n^{1-\delta})$*, for some* $\delta > 0$*, with*

$$A_n = -\mathrm{Res}\left(\frac{n!\,\varpi(s)}{s(s+1)\cdots(s+n)}; s=-1\right) = n\left(\sum_{k=0}^{k_0} a_k \log^k n\right).$$

2.3. Application to the analysis of the three algorithms. We conclude our discussion of algorithms `QuickSort` and `QuickVal`, then that of `QuickQuant`.

— *Analysis of* `QuickSort`. In this case, the Dirichlet series of interest is

$$\frac{\varpi(s)}{s+1} = \frac{2}{s+1}\mathcal{J}[(t-u)^{-(2+s)}] = 2\frac{\Lambda(s)}{s(s+1)^2}\,. \tag{14}$$

For a Λ–tamed source, there is a triple pole at $s = -1$: a pole of order 1 arising from $\Lambda(s)$ and a pole of order 2 brought by the factor $1/(s+1)^2$. For a periodic source, the other poles located on $\Re s = -1$ provide the periodic term $nP(\log n)$, in the form of a Fourier series. This concludes the proof of Theorem 1.

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

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

Proposition 4 entails that $\varpi(s)$ is analytic at $s = -1$. Then, the integrand in (13) has a simple pole at $s = -1$, brought by the factor $1/(s+1)$ and Proposition 5 applies as soon as the source $\mathcal{S}$ is Π–tamed. Thus, for $\delta < 1 - (1/\gamma)$:

$$V_n^{(\alpha)} = \rho_{\mathcal{S}}(\alpha)n + O(n^{1-\delta}). \tag{15}$$

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 Figure 1 (respectively, $]-\infty, -1/2]$, $[-1/2, +1/2]$, $[1/2, \infty[$).

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

(a) The algorithms are asymptotically "similar enough": If $X_1, \ldots, X_n$ 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 $c = \min(1/2, 1-(1/\gamma))$.
(c) The error term in the expansion (15) is uniform with respect to α.

3 Sources, `QuickSort`, and `QuickSelect`

We can now recapitulate and examine the situation of classical sources with respect to the tameness and periodicity conditions of Definition 3. *All* the sources listed below are Π–tamed, so that Theorem 2 (for `QuickQuant`) is systematically

applicable to them. The finer properties of being Λ–tamed, Λ–strongly tamed, or periodic only intervene in Theorem 1, for `Quicksort`.

Memoryless sources and Markov chains. The memoryless (or Bernoulli) sources correspond to an alphabet of some cardinality $r \geq 2$, where symbols are taken independently, with p_j the probability of the jth symbol. The Dirichlet series is

$$\Lambda(s) = (1 - p_1^{-s} - \cdots - p_r^{-s})^{-1}$$

Despite their simplicity, memoryless sources are *never* Λ–strongly tamed. They can be classified into three subcategories (Markov chain sources obey a similar trichotomy).

(i) A memoryless source is said to be *Diophantine* if at least one ratio of the form $(\log p_i)/(\log p_j)$ is irrational and poorly approximable by rationals. *All Diophantine sources are Λ-tamed in the sense of Definition 3(b);* i.e., the existence of a hyperbolic region (11) can be established. Note that, in a measure-theoretic sense, *almost all memoryless sources are Diophantine.*

(ii) A memoryless source is periodic when *all* ratios $(\log p_i)/(\log p_1)$ are rational (this includes cases where $p_1 = \cdots = p_r$, in particular, the model of uniform random bits). The function $\Lambda(s)$ then admits an imaginary period, and its poles are regularly spaced on $\Re(s) = -1$. This entails the presence of an additional periodic term in the asymptotic expansion of the cost of `QuickSort`, which corresponds to complex poles of $\Lambda(s)$; this is Case (ii) of Theorem 1.

(iii) Finally, *Liouvillean sources* are determined by the fact that the ratios are not all rational, but *all* the irrational ratios are very well approximated by rationals (i.e., have infinite irrationality measure); for their treatment, we can appeal to the Tauberian theorems of Delange [6]

Dynamical sources were described in Section 1. They are principally determined by a *shift* transformation T. As shown by Vallée [19] (see also [3]), the Dirichlet series of fundamental intervals can be analysed by means of *transfer operators*, specifically the ones of *secant type*. We discuss here the much researched case where T is a non-linear expanding map of the interval. (Such a source may be periodic only if the shift T is conjugated to a piecewise affine map of the type previously discussed.) One can then adapt deep results of Dolgopyat [4], to the secant operators. In this way, it appears that a dynamical source is strongly tamed as soon as the branches of the shift (i.e., the restrictions of T to the intervals $\mathcal{I}_\sigma$) are not "too often" of the "same form"—we say that such sources are of *type* $D1$. Such is the case for continued fraction sources, which satisfy a "Riemann hypothesis" [1,2]. The adaptation of other results of Dolgopyat [5] provides natural instances of tamed sources with a hyperbolic strip: this occurs as soon as the branches all have the same geometric form, but not the same arithmetic features—we say that such sources are of type $D2$.

Theorem 3. *The memoryless and Markov chain sources that are Diophantine are Λ–tamed. Dynamical sources of type $D2$ are Λ–tamed and dynamical sources of type $D1$ are Λ–strongly tamed. The estimates of Theorem 1 and 2 are then applicable to all these sources, and, in the case of Λ–strongly tamed sources, the error term of* `QuickSort` *in* (1) *can be improved to $O(n^{1-\delta})$, for some $\delta > 0$.*

It is also possible to study "intermittent sources" in the style of what is done for the subtractive Euclidean algorithm [20, §2,6]: a higher order pole of $\Lambda(s)$ then arises, leading to complexities of the form $n \log^3 n$ for `QuickSort`, whereas `QuickQuant` remains of order n.

Acknowledgements. The authors are thankful to Také Nakama for helpful discussions. Research for Clément, Flajolet, and Vallée was supported in part by the SADA Project of the ANR (French National Research Agency). Research for Fill was supported by The Johns Hopkins University's Acheson J. Duncan Fund for the Advancement of Research in Statistics. Thanks finally to Jérémie Lumbroso for a careful reading of the manuscript.

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. *Proc. AMS 133*, 3 (2005), 865–874.
3. 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.
4. Dolgopyat, D. On decay of correlations in Anosov flows, *Annals of Mathematics* 147 (1998) 357-390.
5. Dolgopyat, D. Prevalence of rapid mixing (I) *Ergodic Theory and Dynamical Systems* 18 (1998) 1097-1114.
6. Delange, H. Généralisation du théorème de Ikehara. *Annales scientifiques de l'École Normale Supérieure Sér. 3 71*, 3 (1954), 213–142.
7. 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.
8. Fill, J. A., and Nakama, T. Analysis of the expected number of bit comparisons required by Quickselect. `ArXiv:0706.2437`. To appear in *Algorithmica*. Extended abstract in *Proceedings ANALCO* (2008), SIAM Press, 249–256.
9. Flajolet, P., and Sedgewick, R. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Comp. Sc. 144* (1995), 101–124.
10. Flajolet, P., and Sedgewick, R. *Analytic Combinatorics*. Cambridge University Press, 2009. Available electronically from the authors' home pages.
11. Grabner, P., and Prodinger, H. On a constant arising in the analysis of bit comparisons in Quickselect. *Quaestiones Mathematicae 31* (2008), 303–306.
12. Kirschenhofer, P., Prodinger, H., and Martínez, C. Analysis of Hoare's FIND Algorithm. *Random Structures & Algorithms. 10* (1997), 143–156.
13. Knuth, D. E. *The Art of Computer Programming*, 2nd ed., vol. 3: Sorting and Searching. Addison-Wesley, 1998.
14. Nörlund, N. E. Leçons sur les équations linéaires aux différences finies. Gauthier-Villars, Paris, 1929.
15. Nörlund, N. E. *Vorlesungen über Differenzenrechnung*. Chelsea Publishing Company, New York, 1954.
16. Sedgewick, R. *Quicksort*. Garland Pub. Co., New York, 1980. Reprint of Ph.D. thesis, Stanford University, 1975.
17. Sedgewick, R. *Algorithms in C, Parts 1–4*, third ed. Addison–Wesley, Reading, Mass., 1998.
18. Szpankowski, W. *Average-Case Analysis of Algorithms on Sequences*. John Wiley, 2001.
19. Vallée, B. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica 29*, 1/2 (2001), 262–306.
20. Vallée, B. Euclidean dynamics. *Discrete and Continuous Dynamical Systems 15*, 1 (2006), 281–352.