Analysis of radix selection on Markov sources

Kevin Leckey and Ralph Neininger*
Institute for Mathematics
J.W. Goethe University Frankfurt
60054 Frankfurt am Main
Germany

Henning Sulzbach†
Equipe-projet RAP
INRIA Paris - Rocquencourt
78153 le Chesnay Cedex
France

April 14, 2014

Abstract

The complexity of the algorithm Radix Selection is considered for independent data generated from a Markov source. The complexity is measured by the number of bucket operations required and studied as a stochastic process indexed by the ranks; also the case of a uniformly chosen rank is considered. The orders of mean and variance of the complexity and limit theorems are derived. We find weak convergence of the appropriately normalized complexity towards a Gaussian process with explicit mean and covariance functions (in the space $\mathrm{D}[0,1]$ of càdlàg functions on $[0,1]$ with the Skorokhod metric) for uniform data and the asymmetric Bernoulli model. For uniformly chosen ranks and uniformly distributed data the normalized complexity was known to be asymptotically normal. For a general Markov source (excluding the uniform case) we find that this complexity is less concentrated and admits a limit law with non-normal limit distribution.

AMS 2010 subject classifications. Primary 68P10, 60F17; secondary 60G15, 60C05, 68Q25.
Key words. Radix Selection, Markov source model, complexity, weak convergence, Gaussian process.

*Supported by DFG grant NE 828/2-1
†Supported by the FSMP, reference: ANR-10-LABX-0098

1 Introduction

Radix Selection is an algorithm to select an order statistic from a set of data in $[0, 1]$ as follows. An integer $b \geq 2$ is fixed. In the first step the unit interval is decomposed into the intervals, also called *buckets*, $[0, 1/b), [1/b, 2/b), \ldots, [(b-2)/b, (b-1)/b)$ and $[(b-1)/b, 1]$ and the data are assigned to these buckets according to their value. If the bucket containing the datum with rank to be selected contains further data the algorithm is recursively applied by again decomposing this bucket equidistantly and recursing. The algorithm stops once the bucket containing the rank to be selected contains no other data. Assigning a datum to a bucket is called a *bucket operation* and the algorithm's complexity is measured by the total number of bucket operations required.

Radix Selection is especially suitable when data are stored as expansions in base (radix) b, the case $b = 2$ being the most common on the level of machine data. For such expansions a bucket operation breaks down to access a digit (or bit).

In this extended abstract we study the complexity of Radix Selection in a probabilistic model. We assume that n data are modeled independently with b-ary expansions generated from a Markov chain on the alphabet $\{0, \ldots, b-1\}$. For the ranks to be selected we use two models. First, we consider the complexity of a random rank uniformly distributed over $\{1, \ldots, n\}$ and independent from the data. This is the model proposed and studied (for independent, uniformly over $[0, 1]$ distributed data) in Mahmoud, Flajolet, Jacquet and Régnier [15]. The complexities of all ranks are averaged in this model and, in accordance with the literature, we call it the model of *grand averages*. Second, all possible ranks are considered simultaneously. Hence, we study the stochastic process of the complexities indexed by the ranks $1, \ldots, n$. We choose a scaling in time and space which asymptotically gives access to the complexity to select quantiles from the data, i.e., ranks of the size tn with $t \in [0, 1]$. We call this model for the ranks the *quantile-model*.

The main results of this extended abstract are on the asymptotic orders of mean and variance and limit laws for the complexity of Radix Selection for our Markov source model both for grand averages and for the quantile-model. For the quantile-model we find Gaussian limit processes for the uniform model (defined below) for the data as well as for the asymmetric Bernoulli model (defined below). For the general Markov source model with $b = 2$ we identify the first asymptotic term of the mean complexity. For grand averages and uniform data it was shown in Mahmoud et al. [15] that the normalized complexity is asymptotically normal. We find that for Markov sources (with $b = 2$) other than uniform the limit distribution is no longer normal and the complexity is less concentrated. An explanation of this behavior is given at the end of section 3.2.

We present our analysis separately for uniform data in section 2 and for Markov sources different from the uniform model in section 3, where the quantile-model is discussed in section 3.1, the grand averages in section 3.2.

A general reference on bucket algorithms is Devroye [5]. A large body of probabilistic analysis of digital structures is based on methods from analytic combinatorics, see Flajolet and Sedgewick [7], Knuth [13] and Szpankowski [18]. For an approach based on renewal theory see Janson [10] and the references given there. Our Markov source model is a special case of the model of dynamical sources, see Clément, Flajolet and Vallée [4].

We close this introduction defining the Markov source model explicitly, fixing some standard notation and stating corresponding results for the related Radix Sorting algorithm.

The Markov source model: We model data strings over the alphabet $\Sigma = \{0, \dots, b-1\}$ with a fixed integer $b \geq 2$ generated by a homogeneous Markov chain. The data strings $s = (s_i)_{i \geq 1}$ are also interpreted as b-ary expansions of a real number $s \in [0,1]$ via the identification

$$s = \sum_{i=1}^{\infty} s_i b^{-i}.$$

Conversely, if to $s \in [0,1)$ a b-ary expansion $s = (s_i)_{i \geq 1}$ is associated, to avoid ambiguity, we chose the expansion such that we have $s_i < b-1$ for infinitely many $i \in \mathbb{N}$. (For $s = 1$ we use the expansion where $s_i = b-1$ for all $i \in \mathbb{N}$.) The most important case is $b = 2$ where the data are binary strings.

In general, a homogeneous Markov chain on Σ is given by its initial distribution $\mu = \sum_{\ell=0}^{b-1} \mu_\ell \delta_\ell$ on Σ and the transition matrix $(p_{ij})_{i,j \in \Sigma}$. Here, δ_x denotes the Dirac measure in $x \in \mathbb{R}$. Hence, the initial state is ℓ with probability μ_ℓ for $\ell = 0, \dots, b-1$. We have $\mu_\ell \in [0,1]$ and $\sum_{\ell=0}^{b-1} \mu_\ell = 1$. A transition from state i to j happens with probability p_{ij}, $i, j \in \Sigma$. Now, a data string is generated as the sequence of states taken by the Markov chain. In our Markov source model assumed subsequently all data strings are independent and identically distributed according to the given Markov chain.

We always assume that $p_{ij} < 1$ for all $i, j \in \Sigma$. Note that we do not necessarily assume the Markov chain to converge to a stationary distribution nor that it starts in a stationary distribution.

The case $p_{ij} = \mu_i = 1/b$ for all $i, j \in \Sigma$ is the case where all symbols within all data are independent and uniformly distributed over Σ. Then the associated numbers are independent and uniformly distributed over $[0,1]$. We call this the *uniform model.* For $b = 2$ the uniform model is also called *symmetric Bernoulli model.* The *asymmetric Bernoulli model* for $b = 2$ is the case where $p_{i1} = \mu_1 = p$ for $i = 0, 1$ and a $p \in (0,1)$ with $p \neq \frac{1}{2}$.

Notation. We write $\xrightarrow{d}$ for convergence in distribution and $\stackrel{d}{=}$ for equality in distribution. By $\mathrm{B}(n,p)$ with $n \in \mathbb{N}$ and $p \in [0,1]$ the binomial distribution is denoted, by $\mathrm{B}(p)$ the Bernoulli distribution with success probability p, by $\mathcal{N}(\mu, \sigma^2)$ the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The Bachmann–Landau symbols are used.

Radix Sorting. The Radix Sorting algorithm consists of assigning all data to the buckets as for Radix Selection. Then the algorithm recurses on all buckets containing more than one datum. Clearly, this leads to a sorting algorithm. The complexity of Radix Sorting is also measured by the number of bucket operations. It has thoroughly been analyzed in the uniform model with refined expansions for mean and variance involving periodic functions and a central limit law for the normalized complexity, see Knuth [13], Jacquet and Régnier [9], Kirschenhofer, Prodinger and Szpankowski [11] and Mahmoud et al. [15].

For the Markov source model (with $b = 2$ and $0 < p_{ij} < 1$ for all $i, j = 1, 2$) the orders of mean and variance and a central limit theorem for the complexity of Radix Sorting were derived in Leckey, Neininger and Szpankowski [14].

Acknowledgements. We thank the referees for their careful reading and constructive remarks.

2 The uniform model — selection of quantiles

In this section our model consists of independent data, identically and uniformly distributed over $[0, 1]$. We fix $b \geq 2$ and consider bucket selection using b buckets in each step. The number $Y_n(\ell)$ of bucket operations needed by bucket selection to select rank $\ell \in \{1, \dots, n\}$ in a set of n such data is studied as a process in $1 \leq \ell \leq n$. We write $Y_n := (Y_n(\ell))_{1\leq \ell \leq n}$. For a refined asymptotic analysis we normalize the process in space and time and consider $X_n = (X_n(t))_{0\leq t\leq 1}$ defined for $n \geq 1$ and $t \in [0, 1]$ by

$$X_n(t) := \frac{Y_n(\lfloor tn\rfloor + 1) - \frac{b}{b-1}n}{\sqrt{n}}, \tag{1}$$

where we set $Y_n(n + 1) := Y_n(n)$. The process X_n has càdlàg paths and is considered as a random variable in $\mathrm{D}[0, 1]$ endowed with the Skorokhod metric d_{sk}, see Billingsley [2, Chapter 3].

Subsequently, we use prefixes of b-ary expansions. For $s, t \in [0, 1]$ based on their b-ary expansions $s = \sum_{i=1}^{\infty} s_i \cdot b^{-i}$, $t = \sum_{i=1}^{\infty} t_i \cdot b^{-i}$ with $s_i, t_i \in \{0, \dots, b - 1\}$ with the conventions stated in the introduction we denote the length of the longest common prefix by

$$j(s, t) := \max\{i \in \mathbb{N} \,|\, (s_1, \dots, s_i) = (t_1, \dots, t_i)\} \tag{2}$$

with the conventions $\max \emptyset := 0$ and $\max \mathbb{N} := \infty$.

Theorem 2.1. *Let $b \in \mathbb{N}$ with $b \geq 2$. Consider bucket selection using b buckets on a set of independent data uniformly distributed on $[0, 1]$. For the process $X_n = (X_n(t))_{0\leq t\leq 1}$ of the normalized number of bucket operations $Y_n(\ell)$ as defined in (1) we have weak convergence, as $n \to \infty$, in $(\mathrm{D}[0, 1], d_{sk})$:*

$$X_n \xrightarrow{d} G.$$

Here, $G = (G(t))_{t\in[0,1]}$ is a centered Gaussian process (depending on b) with covariance function

$$\mathbb{E}[G(s)G(t)] = \frac{b}{(b-1)^2} - \frac{b+1}{(b-1)^2} b^{-j(s,t)}, \quad s,t \in [0,1],$$

where $j(s,t)$ is the length of the longest common prefix defined in (2) and $b^{-\infty} := 0$.

Theorem 2.1 implies the asymptotic behavior of the worst case complexity $\max_{\ell=1,\dots,n} Y_n(\ell)$ of Radix Selection:

Corollary 2.2. *For the worst case complexity of Radix Selection in the model and notation of Theorem 2.1 we have, as $n \to \infty$, that*

$$\frac{1}{\sqrt{n}} \left(\sup_{1\le \ell \le n} Y_n(\ell) - \frac{b}{b-1} n \right) \xrightarrow{d} \sup_{t\in[0,1]} G(t).$$

For the Gaussian process G in Theorem 2.1 we have the following results on the tails of its supremum and regarding the continuity of its paths.

Theorem 2.3. *For the supremum $S = \sup_{t\in[0,1]} G(t)$ of the Gaussian process G in Theorem 2.1 we have for any $t > 0$ that*

$$\mathbb{P}(|S - \mathbb{E}[S]| \ge t) \le 2\exp\left(-\frac{(b-1)^2}{2b} t^2 \right).$$

In the Euclidean topology on $[0,1]$ (induced by absolute value) the Gaussian process G in Theorem 2.1 does not have continuous paths. Typically, in the study of Gaussian processes a metric on the index set is derived from the covariance function. We consider

$$d(s,t) := \sqrt{\mathbb{E}[(G(t) - G(s))^2]} = \frac{\sqrt{2(b+1)}}{b-1} \cdot b^{-j(s,t)/2}, \quad s,t \in [0,1].$$

The subsequent results in this section are stated with respect to the (topologically) equivalent metric

$$d_b(s,t) := b^{-j(s,t)}, \quad s,t \in [0,1].$$

Theorem 2.4 (Modulus of continuity)**.** *For the Gaussian process $G = (G(t))_{t\in[0,1]}$ in Theorem 2.1 we have, almost surely,*

$$2\frac{\sqrt{\log b}}{\sqrt{b-1}} \le \limsup_{n\to\infty} \sup_{\substack{s,t\in[0,1],\\ d_b(s,t)=b^{-n}}} \frac{|G(t) - G(s)|}{\sqrt{nb^{-n}}} \le 2\frac{\sqrt{2\log b}}{\sqrt{b-1}(1-b^{-1/2})}.$$

Theorem 2.5 (Hölder continuity)**.** *For any $\beta < 1/2$, almost surely, the paths of the Gaussian process $G = (G(t))_{t\in[0,1]}$ in Theorem 2.1 are Hölder continuous with exponent β with respect to d_b. For any $\beta > 1/2$, almost surely, the paths of G are nowhere pointwise Hölder continuous with exponent β with respect to d_b.*

Outline of the analysis: We outline the analysis leading to Theorems 2.1–2.5. To set up a recurrence for the process $Y_n := (Y_n(\ell))_{1\le \ell\le n}$ we denote by $I^n = (I^n_1, \dots, I^n_b)$ the numbers of elements in the b buckets after distribution of all n elements in the first partitioning stage. We abbreviate $F^n_0 := 0$ and

$$F^n_r := \sum_{j=1}^r I^n_j, \qquad 1 \le r \le b.$$

Note that we have $F^n_b = n$. Then, after the first partitioning phase, the element of rank ℓ is in bucket r if and only if $F^n_{r-1} < \ell \le F^n_r$. This implies the recurrence

$$Y_n \stackrel{d}{=} \left(\sum_{r=1}^b \mathbf{1}_{\left\{F^n_{r-1}<\ell\le F^n_r\right\}} Y^r_{I^n_r}\left(\ell - F^n_{r-1}\right) + n \right)_{1\le \ell\le n}, \tag{3}$$

where $(Y^1_j), \dots, (Y^b_j), I^n$ are independent and the Y^r_j have the same distribution as Y_j for all $j \ge 0$ and $r = 1, \dots, b$.

By the model of independent and uniformly distributed data we have that the vector I^n has the multinomial $M(n; \frac{1}{b}, \dots, \frac{1}{b})$ distribution. Hence, we have $\frac{1}{n} I^n \to (\frac{1}{b}, \dots, \frac{1}{b})$ almost surely as $n \to \infty$ and

$$\frac{I^n - \frac{1}{b}(n, \dots, n)}{\sqrt{n}} \to (N_1, \dots, N_b),$$

where $(N_1, \dots, N_b)$ is a multivariate normal distribution $\mathcal{N}(0, \Omega)$ with mean zero and covariance matrix Ω given by $\Omega_{ij} = \frac{b-1}{b^2}$ if $i = j$ and $\Omega_{ij} = -\frac{1}{b^2}$ if $i \ne j$. Note that for $b = 2$ we have $N_2 = -N_1$. Below, we denote by

$$\mathcal{N} = (\mathcal{N}_1, \dots, \mathcal{N}_b)$$

a vector with distribution $\frac{b}{b-1}(N_1, \dots, N_b)$. Hence $(\mathcal{N}_1, \dots, \mathcal{N}_b)$ has a multivariate normal distribution with mean zero and covariance matrix $\Upsilon = (\Upsilon_{ij})_{i,j\in\Sigma}$ given by

$$\Upsilon_{ij} = \begin{cases} \frac{1}{b-1}, & \text{if } i = j, \\ -\frac{1}{(b-1)^2}, & \text{if } i \ne j. \end{cases} \tag{4}$$

For the normalized processes X_n in (1) we thus obtain

$$\begin{aligned} X_n \stackrel{d}{=} \Bigg(&\sum_{r=1}^b \mathbf{1}_{\left\{F^n_{r-1}<\lfloor tn\rfloor+1\le F^n_r\right\}} \sqrt{\frac{I^n_r}{n}} X^r_{I^n_r}\left(\frac{nt - F^n_{r-1}}{I^n_r}\right) \\ &+ \sum_{r=1}^b \mathbf{1}_{\left\{F^n_{r-1}<\lfloor tn\rfloor+1\le F^n_r\right\}} \frac{b}{b-1} \frac{I^n_r - \frac{1}{b}n}{\sqrt{n}} \Bigg)_{0\le t\le 1}, \end{aligned} \tag{5}$$

with conditions on independence and identical distributions as in (3).

To associate to recurrence (5) a limit equation in the spirit of the contraction method we introduce the indicator functions

$$\mathbf{I}_r(x) := \mathbf{1}_{[r-1,r)}(x) \text{ for } r = 1, \ldots, b-1, \qquad \mathbf{I}_b(x) := \mathbf{1}_{[b-1,b]}(x)$$

and the sawtooth function $s_b : [0,1] \to [0,1]$

$$s_b(t) := \begin{cases} bt - \lfloor bt \rfloor, & 0 \leq t < 1, \\ 1, & t = 1. \end{cases}$$

Moreover, we use the transformations $\mathfrak{A}_r : \mathrm{D}[0,1] \to \mathrm{D}[0,1]$ for $r = 1, \ldots, b$ with

$$f \mapsto \mathfrak{A}_r(f), \quad \mathfrak{A}_r(f)(t) = \mathbf{I}_r(tb) f(s_b(t)) \text{ for } t \in [0,1],$$

and $\mathfrak{B} : \mathbb{R}^b \to \mathrm{D}[0,1]$ with (for $v = (v_1, \ldots, v_b)$)

$$v \mapsto \mathfrak{B}(v), \quad \mathfrak{B}(v)(t) = \sum_{r=1}^{b} \mathbf{I}_r(tb) v_r \text{ for } t \in [0,1].$$

Then we associate the limit equation

$$X \stackrel{d}{=} \sum_{r=1}^{b} \frac{1}{\sqrt{b}} \mathfrak{A}_r(X^r) + \mathfrak{B}(\mathcal{N}), \tag{6}$$

where $X^1, \ldots, X^b, \mathcal{N}$ are independent, the X^r are identically distributed random variables with values in $(\mathrm{D}[0,1], d_{sk})$ and distribution of X, and $\mathcal{N}$ has the centered multivariate normal distribution with covariance matrix Υ given in (4).

A distributional fixed-point equation related to (6) appeared in Sulzbach et al. [17], see the map T in equation (2.5) of [17]. The proof of our Theorem 2.1 can be carried out analogously to the proof in sections 2.1, 2.2 and 3.3 of [17]. Note that in analogy to Lemma 2.3 in [17] our fixed-point equation (6) characterizes the Gaussian limit process G in Theorem 2.1 as the unique fixed-point of (6) subject to the constraint $\mathbb{E}[\|X\|_\infty^{2+\varepsilon}] < \infty$ for any $\varepsilon > 0$ and, hence, in the analysis one has to adapt exponents appropriately. The proofs of our Theorems 2.3–2.5 can be carried out as the corresponding results in section 4 of [17] which are related to and partly based on fundamental work on Gaussian processes, see Dudley [6], Talagrand [19], Adler [1] and Boucheron, Lugosi and Massart [3].

3 The Markov source model

Now the Markov source model is considered for the data. For the rank to be selected the quantile-model is studied in section 3.1, the model of grand averages in section 3.2. We restrict ourselves to the study of Radix Selection using $b = 2$ buckets.

3.1 Selection of quantiles

We consider the complexity of Radix Selection with $b = 2$ buckets assuming the Markov source model for the data and the quantile-model for the rank to be selected. We first define functions $m_\mu : [0,1] \to (0,\infty)$ which appear in the average complexity. For $n \geq 1$ and $i = 0,1$ we recursively define sets $\mathcal{D}^i_n = \{s^i_{n,k} \,|\, k = 0,\dots,2^n\}$ as follows: For $n = 1$ we set $(s^i_{1,0}, s^i_{1,1}, s^i_{1,2}) := (0, p_{i0}, 1)$ for $i = 0,1$. Further, for all $n \geq 1$, $i = 0,1$ and $0 \leq k \leq 2^n$ we set

$$s^i_{n+1,k} := \begin{cases} s^i_{n,k/2}, & \text{if } k \bmod 4 \in \{0,2\}, \\ p_{00}s^i_{n,(k+1)/2} + p_{01}s^i_{n,(k-1)/2}, & \text{if } k \bmod 4 = 1. \\ p_{10}s^i_{n,(k+1)/2} + p_{11}s^i_{n,(k-1)/2}, & \text{if } k \bmod 4 = 3, \end{cases}$$

We further define $\mathcal{D}^i_\infty := \cup_{n=1}^\infty \mathcal{D}^i_n$. Note that for each $n \geq 1$ the set $\mathcal{D}^i_n$ decomposes the unit interval into 2^n sub-intervals. For $t \in [0,1] \setminus \mathcal{D}^i_\infty$ we denote by $\lambda^i_n(t)$ the length of the (unique) sub-interval of this decomposition that contains t. Then, for $i = 0,1$ and $t \in [0,1] \setminus \mathcal{D}^i_\infty$ we set

$$m_i(t) := 1 + \sum_{n=1}^\infty \lambda^i_n(t).$$

Further, for an initial distribution $\mu = \mu_0\delta_0 + \mu_1\delta_1$ with $\mu_0 \in [0,1]$ we denote

$$\mathcal{D}^\mu_\infty := \mu_0\mathcal{D}^0_\infty \cup \left(\mu_0 + \mu_1\mathcal{D}^1_\infty\right)$$

and, for $t \in [0,1] \setminus \mathcal{D}^\mu_\infty$,

$$m_\mu(t) := \begin{cases} \mu_0 m_0\left(\frac{t}{\mu_0}\right) + 1, & \text{if } t < \mu_0, \\ (1-\mu_0)m_1\left(\frac{t-\mu_0}{1-\mu_0}\right) + 1, & \text{if } t > \mu_0. \end{cases}$$

We have the following asymptotic behavior of the average complexity:

Theorem 3.1. *Let $Y^\mu_n(\ell)$ denote the number of bucket operations of Radix Selection with $b = 2$ selecting a rank $1 \leq \ell \leq n$ among n independent data generated from the Markov source model with initial distribution $\mu = \mu_0\delta_0 + \mu_1\delta_1$ where $\mu_0 \in [0,1]$ and transition matrix $(p_{ij})_{i,j\in\{0,1\}}$ with $p_{ij} < 1$ for all $i,j = 0,1$. Then, for all $t \in [0,1] \setminus D^\mu_\infty$ as $n \to \infty$, we have*

$$\mathbb{E}[Y^\mu_n(\lfloor tn \rfloor + 1)] = m_\mu(t)n + o(n). \tag{7}$$

Outline of the analysis: We denote by $Y^0_n = (Y^0_n(\ell))_{1\leq\ell\leq n}$ and $Y^1_n = (Y^1_n(\ell))_{1\leq\ell\leq n}$ the number of bucket operations for a Markov source model as in Theorem 3.1 for initial distributions $p_{00}\delta_0 + p_{01}\delta_1$ and $p_{10}\delta_0 + p_{11}\delta_1$ respectively. Then we have the system of recursive distributional equations, for $n \geq 2$,

$$Y^i_n \stackrel{d}{=} \left(\mathbf{1}_{\{\ell \leq J^i_n\}} Y^0_{J^i_n}(\ell) + \mathbf{1}_{\{\ell > J^i_n\}} Y^1_{n-J^i_n}(\ell - J^i_n) + n\right)_{1\leq\ell\leq n}, \quad i = 0,1, \tag{8}$$

where $Y_0^0, \dots, Y_n^0, Y_0^1, \dots, Y_n^1, J_n^0, J_n^1$ are independent (the independence between J_n^0 and J_n^1 is not required) and we have that J_n^i is $\mathrm{B}(n, p_{i0})$ distributed for $i = 0, 1$. Moreover, for general initial distribution μ we further have

$$Y_n^\mu \overset{d}{=} \left(\mathbf{1}_{\{\ell \le K_n\}} Y_{K_n}^0(\ell) + \mathbf{1}_{\{\ell > K_n\}} Y_{n-K_n}^1(\ell - K_n) + n \right)_{1 \le \ell \le n}, \tag{9}$$

where $Y_0^0, \dots, Y_n^0, Y_0^1, \dots, Y_n^1, K_n$ are independent and K_n has the binomial $\mathrm{B}(n, \mu_0)$ distribution.

The proof of Theorem 3.1 is based on k times iterating the system (8) with $k = k(n) = \Theta(\log n)$ chosen appropriately. The contributions of the toll functions within these k iterations yield the main contribution, the other terms are asymptotically negligible.

A distributional analysis of the quantile-model is left for the full paper version of this extended abstract as well as the behavior at the $t \in \mathcal{D}_\infty^\mu$. For these t the expansion (7) still holds when defining $m_\mu(t)$ as the average of the left-hand and right-hand limit of m_μ at t.

A special case where the analysis is simplified considerably is the asymmetric Bernoulli model discussed next where we obtain a functional limit law as for the uniform model in Theorem 2.1.

The asymmetric Bernoulli model: The data model called *asymmetric Bernoulli model* consists of all data being independent and having independent bits all identically distributed over $\Sigma = \{0, 1\}$ with Bernoulli $\mathrm{B}(p)$ distribution for a fixed $p \in (0, 1)$ with $p \neq \frac{1}{2}$. Note that this can also be considered as a special case of the Markov source model by choosing $\mu_0 = p_{00} = p_{10} = 1 - p$ and $\mu_1 = p_{01} = p_{11} = p$. Here, the analysis simplifies considerably compared to the general Markov source model due to the fact that the mean function m corresponding to m_0, m_1, m_μ in Theorem 3.1 becomes an affine function. We have the following results:

Theorem 3.2. *Consider bucket selection using $b = 2$ buckets on a set of n independent data generated from the asymmetric Bernoulli model with success probability $p \in (0, 1) \setminus \{\frac{1}{2}\}$. For the process $X_n^{\mathrm{asyB}} = (X_n^{\mathrm{asyB}}(t))_{0 \le t \le 1}$ of the normalized number of bucket operations $Y_n^{\mathrm{asyB}}(\ell)$ defined by*

$$X_n^{\mathrm{asyB}}(t) := \frac{Y_n^{\mathrm{asyB}}(\lfloor tn \rfloor + 1) - m(t)n}{\sqrt{n}}, \quad t \in [0, 1],$$

with $Y_n^{\mathrm{asyB}}(n+1) := Y_n^{\mathrm{asyB}}(n)$ and

$$m(t) = \frac{2p - 1}{p(1 - p)} t + \frac{1}{p}, \quad t \in [0, 1],$$

we have weak convergence, as $n \to \infty$, in $(\mathrm{D}[0, 1], d_{sk})$:

$$X_n^{\mathrm{asyB}} \xrightarrow{d} G^{\mathrm{asyB}}.$$

Here, $G^{\mathrm{asyB}} = (G^{\mathrm{asyB}}(t))_{t\in[0,1]}$ is a centered Gaussian process (depending on p) with covariance function given, for $s, t \in [0,1]$, by

$$\mathbb{E}[G^{\mathrm{asyB}}(s)G^{\mathrm{asyB}}(t)] = -\prod_{k=1}^{r(s,t)} p[g(t,k)] + \sum_{k=1}^{r(s,t)} \frac{\prod_{j=1}^{k} p[g(t,j)]}{p[1-g(t,k)]},$$

where $p[0] := 1-p$, $p[1] := p$ and the functions $r : [0,1]^2 \to \mathbb{N}_0 \cup \{\infty\}$ and $g : [0,1] \times \mathbb{N}_0 \to \{0,1\}$ are defined as follows:

$$r(s,t) = \max\{n \in \mathbb{N}_0 | g(s,\ell) = g(t,\ell),\ 1 \le \ell \le n\}$$

and g and $h : [0,1] \times \mathbb{N}_0 \to [0,1]$ are recursively defined by $g(t,0) = 0$, $h(t,0) = t$ for $t \in [0,1]$ and for $k \ge 1$ by

$$g(t,k) = \begin{cases} 0, & \text{if } h(t,k-1) < 1-p, \\ 1, & \text{if } h(t,k-1) \ge 1-p, \end{cases}$$

$$h(t,k) = \begin{cases} \frac{h(t,k-1)}{1-p}, & \text{if } h(t,k-1) < 1-p, \\ \frac{h(t,k-1)-(1-p)}{p}, & \text{if } h(t,k-1) \ge 1-p. \end{cases}$$

For the maximum of the complexities we obtain the following corollary:

Corollary 3.3. *In the model and notation of Theorem 3.2 we have, as $n \to \infty$, that*

$$\frac{1}{\sqrt{n}} \left(\sup_{1 \le \ell \le n} \left(Y_n^{\mathrm{asyB}}(\ell) - m\left(\frac{\ell}{n}\right) n \right) \right) \xrightarrow{d} \sup_{t\in[0,1]} G^{\mathrm{asyB}}(t).$$

Theorem 3.4. *For the supremum $S' = \sup_{t\in[0,1]} G^{\mathrm{asyB}}(t)$ of the Gaussian process G^{asyB} in Theorem 3.2 we have for any $t > 0$ with $p_\vee := \max\{p, 1-p\}$ that*

$$\mathbb{P}(|S' - \mathbb{E}[S']| \ge t) \le 2\exp\left(-\frac{(1-p_\vee)^2}{2p_\vee} t^2\right).$$

3.2 Selection of a uniform rank

We now consider the complexity of Radix Selection with $b = 2$ buckets assuming the Markov source model for the data and the model of grand averages for the rank. We have the following asymptotic behavior:

Theorem 3.5. *Let W_n denote the number of bucket operations of Radix Selection with $b = 2$ selecting a uniformly distributed rank independent from n independent data generated from the Markov source model with initial distribution $\mu = \mu_0\delta_0 + \mu_1\delta_1$ where $\mu_0 \in [0,1]$ and transition matrix $(p_{ij})_{i,j\in\{0,1\}}$ with $p_{ij} < 1$ for all $i, j = 0, 1$. Then, as $n \to \infty$, we have*

$$\mathbb{E}[W_n] = \kappa_\mu n + o(n)$$

with $\kappa_\mu > 0$ given in (15) and

$$\frac{W_n}{n} \xrightarrow{d} Z_\mu,$$

where the convergence also holds with all moments. The distribution of Z_μ is given by

$$Z_\mu \stackrel{d}{=} B_{\mu_0}\mu_0 Z^0 + (1 - B_{\mu_0})(1-\mu_0)Z^1 + 1, \tag{10}$$

where B_{μ_0}, Z^0, Z^1 are independent and B_{μ_0} has the Bernoulli distribution $\mathrm{B}(\mu_0)$. The distributions of Z^0 and Z^1 are the unique integrable solutions of the system (12).

Outline of the analysis: We denote by $W_n^\mu := W_n$ the complexity as stated in Theorem 3.5 and, for initial distributions $p_{i0}\delta_0 + p_{i1}\delta_1$, write $W_n^i := W_n^{p_{i0}\delta_0+p_{i1}\delta_1}$ for $i = 0, 1$. We have the system of distributional recurrences, for $n \geq 2$,

$$W_n^i \stackrel{d}{=} B_{ii} W^0_{J_n^i} + (1 - B_{ii}) W^1_{n-J_n^i} + n, \quad i = 0, 1, \tag{11}$$

where $W_1^0, \ldots, W_n^0, W_1^1, \ldots, W_n^1$ and $(J_n^0, J_n^1, B_{00}, B_{11})$ are independent and we have that J_n^i is Binomial $\mathrm{B}(n, p_{i0})$ distributed and B_{ii} is mixed Bernoulli distributed with distribution $\mathrm{B}(J_n^i/n)$ for $i = 0, 1$. We normalize

$$Z_n^i := \frac{W^i}{n}, \quad n \geq 1, \ i = 0, 1$$

and obtain, for all $n \geq 2$ that

$$Z_n^i \stackrel{d}{=} B_{ii} \frac{J_n^i}{n} Z^0_{J_n^i} + (1 - B_{ii}) \frac{n - J_n^i}{n} Z^1_{n-J_n^i} + 1, \quad i = 0, 1,$$

with independence relations as in (11). This leads to the limit system

$$Z^i \stackrel{d}{=} B_{p_{i0}} p_{i0} Z^0 + (1 - B_{p_{i0}})(1 - p_{i0}) Z^1 + 1, \quad i = 0, 1, \tag{12}$$

where Z^0, Z^1 and $B_{p_{i0}}$ are independent and $B_{p_{i0}}$ has the Bernoulli $\mathrm{B}(p_{i0})$ distribution for $i = 0, 1$. It is easy to show that subject to $\mathbb{E}[|Z^i|] < \infty$ the limit system (12) has a unique solution; cf. Knape and Neininger [12, section 5]. Also, the convergences $Z_n^i \to Z^i$ can be shown by a contraction argument in any Wasserstein ℓ_p metric with $p \geq 1$. From the limit system (12) we obtain for the expectations $\kappa_i := \mathbb{E}[Z^i]$ for $i = 0, 1$ that

$$\kappa_0 = \frac{1 + p_{01}^2 - p_{11}^2}{2(p_{00} + p_{11})(1 + p_{00}p_{11}) - 2(p_{00} + p_{11})^2} > 0, \tag{13}$$

$$\kappa_1 = \frac{1 + p_{10}^2 - p_{00}^2}{2(p_{00} + p_{11})(1 + p_{00}p_{11}) - 2(p_{00} + p_{11})^2} > 0. \tag{14}$$

Now, for a general initial distribution μ we have

$$W_n^\mu \stackrel{d}{=} B_{\mu\mu} W_{K_n}^0 + (1 - B_{\mu\mu}) W_{n-K_n}^1 + n,$$

where $W_1^0, \ldots, W_n^0, W_1^1, \ldots, W_n^1, (K_n, B_{\mu\mu})$ are independent, K_n has the binomial $\mathrm{B}(n, \mu_0)$ distribution and $B_{\mu\mu}$ has the mixed Bernoulli $\mathrm{B}(K_n/n)$ distribution. This implies for the limit Z_μ of W_n^μ/n the representation

$$Z_\mu \stackrel{d}{=} B_{\mu_0} \mu_0 Z^0 + (1 - B_{\mu_0})(1 - \mu_0) Z^1 + 1,$$

where B_{μ_0}, Z^0, Z^1 are independent and B_{μ_0} has the Bernoulli distribution $\mathrm{B}(\mu_0)$. Hence, we obtain for $\kappa_\mu := \mathbb{E}[Z_\mu]$ the representation

$$\kappa_\mu = \mu_0^2 \kappa_0 + (1 - \mu_0)^2 \kappa_1 + 1, \tag{15}$$

with κ_0, κ_1 given in (13) and (14). The claims of Theorem 3.5 follow from this outline by an application of the contraction method within the Wasserstein metrics.

A remark on concentration for grand averages: Note that for the special case $p_{ij} = \mu_i = \frac{1}{2}$ for $i = 0, 1$ the Markov source model reduces to the uniform model. In this case it was shown (together with more refined results) in Mahmoud et al. [15] that the complexity W_n in Theorem 3.5 as $n \to \infty$ satisfies

$$\frac{W_n - 2n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{16}$$

Our Theorem 3.5 also applies: We find that for the uniform model system (12) is solved deterministically by $Z^i = 2$ almost surely for $i = 0, 1$, hence plugging in into (10) we also obtain $Z_\mu = 2$ and thus $W_n/n \to 2$, which is, although only a law of large numbers, consistent with (16). (The full limit law in (16) is a corollary to our Theorem 2.1.)

However, for $(p_{00}, p_{01}, p_{10}, p_{11}) \neq (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ the system (12) does no longer solve deterministically, so that W_n/n then has a nondeterministic limit and is less concentrated, typical fluctuations are of linear order compared to $\sqrt{n}$ for the uniform model. This behavior becomes transparent when looking at the quantile-model: In the uniform model we have the same leading linear term $2n$ in the expansion of the means of all quantiles, which implies that a uniformly chosen rank conditional on its size will always lead to a complexity of the same linear order $2n$. This does no longer hold for a non-uniform Markov model. The constant $m_\mu(t)$ in the linear growth $m_\mu(t)n$ of the complexity depends on the quantiles $t \in [0, 1]$, see Theorem 3.1. This implies that the complexity can no longer remain concentrated: The fluctuations are now forced to be at least of linear order since different choices of the ranks lead to different linear orders. This is consistent with the fact that in Theorem 3.5 we then find a non-deterministic limit for W_n/n and a variance of the order $\Theta(n^2)$. Further note, that this also implies a simple representation of the limit distribution of Z_μ in Theorem 3.5 as

$$Z_\mu \overset{d}{=} m_\mu(U),$$

with m_μ as in Theorem 3.1 und U uniformly distributed on $[0,1]$.

References

[1] Adler, R.J. (1990) An introduction to continuity, extrema, and related topics for general Gaussian processes. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 12. *Institute of Mathematical Statistics, Hayward, CA. x+160 pp.*

[2] Billingsley, P. (1999) Convergence of probability measures. Second edition. Wiley Series in Probability and Statistics: Probability and Statistics. A Wiley-Interscience Publication. *John Wiley & Sons, Inc., New York.*

[3] Boucheron, S., Lugosi, G. and Massart, P. (2013), Concentration Inequalities: A Nonasymptotic Theory of Independence. *Oxford University Press.*

[4] Clément, J., Flajolet, P. and Vallée, B. (2001) Dynamical sources in information theory: a general analysis of trie structures. Average-case analysis of algorithms (Princeton, NJ, 1998). *Algorithmica* **29**, 307–369.

[5] Devroye, L. (1986) Lecture notes on bucket algorithms. Progress in Computer Science, 6. *Birkhäuser Boston, Inc., Boston, MA.*

[6] Dudley, R.M. (1973) Sample functions of the Gaussian process. *Ann. Probab.* **1**, 66–103.

[7] Flajolet, P. and Sedgewick, R. (2009) Analytic combinatorics. *Cambridge University Press, Cambridge.*

[8] Fuchs, M., Hwang, H.-K. and Zacharovas, V. (2014+) An analytic approach to the asymptotic variance of trie statistics and related structures. *Theoret. Comput. Sci.*, to appear. Electronic preprint available via arXiv:1303.4244

[9] Jacquet, P. and Régnier, M. (1988) Normal limit distribution for the size and the external path length of tries. INRIA Research Report 827.

[10] Janson, S. (2012) Renewal theory in the analysis of tries and strings. *Theoret. Comput. Sci.* **416**, 33–54.

[11] Kirschenhofer, P., Prodinger, H. and Szpankowski, W. (1989) On the variance of the external path length in a symmetric digital trie. Combinatorics and complexity (Chicago, IL, 1987). *Discrete Appl. Math.* **25**, 129–143.

[12] Knape, M. and R. Neininger (2014+) Pólya urns via the contraction method. *Comb. Probab. & Comput.*, to appear. Electronic preprint vailable via arXiv:1301.3404.

[13] Knuth, D.E. (1998) The art of computer programming. Vol. 3. Sorting and searching. Second edition. *Addison-Wesley, Reading, MA.*

[14] Leckey, K., Neininger, R. and Szpankowski, W. (2013) Towards More Realistic Probabilistic Models for Data Structures: The External Path Length in Tries under the Markov Model. *Proceedings ACM-SIAM Symp. Disc. Algo. (SODA)*, 877–886.

[15] Mahmoud, H.M., Flajolet, P., Jacquet, P. and Régnier, M. (2000) Analytic variations on bucket selection and sorting. *Acta Inform.* **36**, 735–760.

[16] Ragab, M. and Rösler, U. (2014) The Quicksort process. *Stochastic Process. Appl.* **124**, 1036–1054.

[17] Sulzbach, H., Neininger, R. and Drmota, M. (2014) A Gaussian limit process for optimal FIND algorithms. *Elect. J. Probab.* **19**, no. 3, 1–28.

[18] Szpankowski, W. (2001) Average case analysis of algorithms on sequences. With a foreword by Philippe Flajolet. Wiley-Interscience Series in Discrete Mathematics and Optimization. *Wiley-Interscience, New York.*

[19] Talagrand, M. (1987) Regularity of Gaussian processes. *Acta Math.* **159**, 99–149.