ON FLUCTUATIONS OF COMPLEXITY MEASURES FOR THE FIND ALGORITHM

JASPER ISCHEBECK AND RALPH NEININGER

Abstract. The FIND algorithm (also called Quickselect) is a fundamental algorithm to select ranks or quantiles within a set of data. It was shown by Grübel and Rösler that the number of key comparisons required by Find as a process of the quantiles $\alpha \in [0,1]$ in a natural probabilistic model converges after normalization in distribution within the càdlàg space $D[0,1]$ endowed with the Skorokhod metric. We show that the process of the residuals in the latter convergence after normalization converges in distribution to a mixture of Gaussian processes in $D[0,1]$ and identify the limit's conditional covariance functions. A similar result holds for the related algorithm QuickVal. Our method extends to other cost measures such as the number of swaps (key exchanges) required by Find or cost measures which are based on key comparisons but take into account that the cost of a comparison between two keys may depend on their values, an example being the number of bit comparisons needed to compare keys given by their bit expansions.

1. Introduction

In 1961, Hoare [11] introduced the algorithm FIND, also called Quickselect, to select a key (an element) of a given rank from a linearly ordered finite set of data. We assume that the data are distinct real numbers. To be definite a simple version of the FIND algorithm is given as follows: FIND(S,k) has as input a set $S = \{s_1, \dots, s_n\}$ of distinct real numbers of size n and an integer $1 \le k \le n$. The algorithm FIND operates recursively as follows: If $n = 1$ we have $k = 1$ and FIND returns the single element of S. If $n \ge 2$ and $S = \{s_1, \dots, s_n\}$ the algorithm first chooses an elements from S, say s_j, called pivot, and generates the sets

$$S_< := \{s_i \,|\, s_i < s_j, i \in \{1, \dots, n\} \setminus \{j\}\}, \quad S_\ge := \{s_i \,|\, s_i \ge s_j, i \in \{1, \dots, n\} \setminus \{j\}\}.$$

If $k = |S_<| + 1$, the algorithm returns s_j. If $k \le |S_<|$, recursively FIND$(S_<, k)$ is applied. If $k \ge |S_<| + 2$, recursively FIND$(S_\ge, k - |S_<| - 1)$ is applied. Note that FIND(S,k) returns the element of rank k from S. There are various variants of the algorithm, in particular regarding how the pivot element is chosen and how S is partitioned into the subsets $S_<$ and $S_\ge$.

In a standard probabilistic model one assumes that the data are ordered, i.e. given as a vector $(s_1, \dots, s_n)$, and are randomly permuted, all permutations being equally likely. This can be achieved assuming that the data are given as $(U_1, \dots, U_n)$ where $(U_j)_{j \in \mathbb{N}}$ is a sequence of i.i.d. random variables with distribution unif$[0,1]$, the uniform distribution over the unit interval $[0,1]$. This is the probabilistic model considered below. Note that the randomness is within the data, while the algorithm is deterministic.

Various cost measures have been considered for FIND, mainly the number of key comparisons required which we analyze in detail below. At the end of this extended abstract we state related results for the number of swaps (key exchanges) required and for cost measures which

Date: March 12, 2024.

2020 *Mathematics Subject Classification.* 60F17, 68Q25, 68P10, 60C05.

are based on key comparisons, where the cost of a comparison may depend on the values of the keys s_i, s_j, the number of bit comparisons required to decide whether $s_i < s_j$ or not being a prominent example.

For analysis purposes a related process, called QuickVal, has been considered, see [18, 8]. Informally, QuickVal for an $\alpha \in [0,1]$ mimics FIND to select (or to try to select) the value α from the set of data, which, in our probabilistic model for large n, comes close to FIND selecting rank $\lfloor \alpha n \rfloor$. To be definite, QuickVal$((U_1, \dots, U_n), \alpha)$ compares the U_i with U_1 to generate sublists

$$S_< := (U_{j_1}, \dots, U_{j_{m-1}}), \quad S_\geq := (U_{j_{m+1}}, \dots, U_{j_n}),$$

with $U_{j_i} < U_1$ for $i = 1, \dots, m-1$ and $2 \leq j_1 < \cdots < j_{m-1}$ and $U_{j_i} \geq U_1$ for $i = m+1, \dots, n$ and $2 \leq j_{m+1} < \cdots < j_n$. Hence, $m - 1 \in \{0, \dots, n-1\}$ is the number of the U_i, $2 \leq i \leq n$, being smaller than U_1. The algorithm recursively calls QuickVal$(S_<, \alpha)$ if $\alpha < U_1$ and $|S_<| > 0$. If $\alpha \geq U_1$ and $|S_\geq| > 0$ recursively QuickVal$(S_\geq, \alpha - U_1)$ is called. The number of key comparisons required by QuickVal$((U_1, \dots, U_n), \alpha)$ is denoted by $S_{\alpha,n}$.

To describe the processes $(S_{\alpha,n})_{\alpha\in[0,1]}$ and their limit (after scaling) conveniently we also consider the binary search tree constructed from the the data $(U_i)_{i\in\mathbb{N}}$. Part of the following definitions are depicted in Figure 1. The data are inserted into the rooted infinite binary tree, where we denote its nodes by the elements of $\{0,1\}^* := \cup_{n=0}^\infty \{0,1\}^n$ as follows. Its root is denoted by the empty word ϵ and for each node $\phi \in \{0,1\}^*$ we denote by $\phi 0$ and $\phi 1$ (the word ϕ appended with a 0 resp. 1) its left and right child respectively. Moreover $|\phi|$ denotes the length of the word ϕ, which is the depth of the corresponding node in the tree. To construct the binary search tree for $(U_1, \dots, U_n)$ the first key U_1 is inserted into the root and occupies the root. Then, successively the following keys are inserted, where each key traverses the already occupied nodes starting at the root as follows: Whenever the key traversing is less than the occupying key at a node it moves on to the left child of that node, otherwise to its right child. The first empty node found is occupied by the key.

To describe the costs of the algorithms we organize, using notation of Fill and Nakama [8], the sub-intervals $([L_\phi, R_\phi))_{\phi\in\{0,1\}^*}$ implicitly generated starting with $[0,1) =: [L_\epsilon, R_\epsilon)$ and recursively setting

$$\begin{aligned} \tau_\phi &:= \inf\{i \in \mathbb{N} \,|\, L_\phi < U_i < R_\phi\}, \\ L_{\phi 0} := L_\phi, \quad R_{\phi 1} := R_\phi, \quad L_{\phi 1} &:= R_{\phi 0} := U_{\tau_\phi}, \quad I_\phi := R_\phi - L_\phi. \end{aligned} \tag{1}$$

Now, if a sublist starting with pivot U_{τ_ϕ} has to be split by QuickVal, the keys which are inserted in the subtree rooted at U_{τ_ϕ} need to be compared with U_{τ_ϕ}. Hence, we get a contribution of key comparisons of

$$S_{\phi,n} = \sum_{\tau_\phi < k \leq n} \mathbf{1}_{[L_\phi, R_\phi)}(U_k). \tag{2}$$

Now, for $\alpha \in [0,1]$, QuickVal$((U_1, \dots, U_n), \alpha)$ generates and splits sublists encoded by $\phi(\alpha, k)$ for $k = 0, 1, \dots$ for which we obtain by $\phi(\alpha, 0) = \epsilon$ and

$$\phi(\alpha, k+1) = \begin{cases} \phi(\alpha,k)0, & \text{if } \alpha < U_{\tau_{\phi(\alpha,k)}}, \\ \phi(\alpha,k)1, & \text{if } \alpha \geq U_{\tau_{\phi(\alpha,k)}}. \end{cases} \tag{3}$$

When using the variables defined in (1) or (2), we abbreviate the notation $\phi(\alpha, k)$ by α, k, such as writing $I_{\alpha,k} := I_{\phi(\alpha,k)}$ or $S_{\alpha,k,n} := S_{\phi(\alpha,k),n}$.

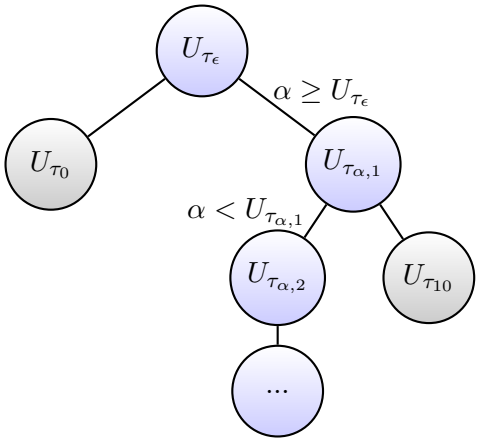

Figure 1. Part of the binary search tree. The pivots of sublists split by $\mathrm{QuickVal}((U_1, \ldots, U_n), \alpha)$ for some $\alpha \in [0,1]$ are on the path indicated. Note that we have $\tau_\epsilon = \tau_{\phi(\alpha,0)} = \tau_{\alpha,0} = 1$ and in this example $\alpha \geq U_1$ and $\alpha < U_{\tau_{\alpha,1}}$ so that $\phi(\alpha, 2) = 10 \in \{0,1\}^2$.

The number of key comparisons required by $\mathrm{QuickVal}((U_1, \ldots, U_n), \alpha)$ is then given by the (finite) sum

$$S_{\alpha,n} = \sum_{k=1}^{\infty} S_{\alpha,k,n}.$$

Fill and Nakama [8, Theorem 3.2] showed (considering more general complexity measures) that for each $\alpha \in [0,1]$ almost surely

$$\frac{1}{n} S_{\alpha,n} \to S_\alpha := \sum_{k=0}^{\infty} I_{\alpha,k}, \quad (n \to \infty). \tag{4}$$

The latter convergence also holds in L_p, see Fill and Matterer [7, Proposition 6.1].

We take the point of view that such an almost sure asymptotic result may be considered a strong law of large numbers (SLLN). The subject of the present extended abstract is to study the fluctuations in such SLLN, sometimes called a central limit analogue. We study these fluctuations as processes in the metric space $(D[0,1], d_{\mathrm{SK}})$ of càdlàg functions endowed with the Skorokhod metric; see Billingsley [2] for background on weak convergence of probability measures on metric spaces in general and on $(D[0,1], d_{\mathrm{SK}})$ in particular. Note that, by definition, $(S_{\alpha,n})_{\alpha\in[0,1]}$ and $(S_\alpha)_{\alpha\in[0,1]}$ have càdlàg paths almost surely. As the normalized process of fluctuations we define

$$G_n := (G_{\alpha,n})_{\alpha\in[0,1]} := \left(\frac{S_{\alpha,n} - nS_\alpha}{\sqrt{n}} \right)_{\alpha\in[0,1]}. \tag{5}$$

Then we have the following result:

Theorem 1. *Let $S_{\alpha,n}$ be the number of key comparisons required by* $\mathrm{QuickVal}((U_1, \ldots, U_n), \alpha)$ *and $(S_\alpha)_{\alpha\in[0,1]}$ as in (4). Then for the fluctuation process G_n defined in (5) we have*

$$G_n \xrightarrow{d} G_\infty \quad \text{in } (D[0,1], d_{\mathrm{SK}}) \qquad (n \to \infty),$$

where G_∞ is a mixture of centered Gaussian processes with random covariance function given by

$$\Sigma_{\infty,\alpha,\beta} := \sum_{k=0}^{J}\sum_{j=0}^{\infty} I_{\alpha,j\vee k} + \mathbf{1}_{\{\alpha\neq\beta\}}(J+1)\sum_{j=J+1}^{\infty}(I_{\beta,j}) - S_\alpha S_\beta, \quad \alpha,\beta\in[0,1], \tag{6}$$

where $J = J(\alpha,\beta) := \max\{k\in\mathbb{N}_0 \,|\, \tau_{\alpha,k} = \tau_{\beta,k}\} \in \mathbb{N}_0\cup\{\infty\}$.

Remark 2. In his PhD thesis, Matterer [14, Theorem 6.4] showed the convergence of the one-dimensional marginals for the functional limit law in Theorem 1.

Remark 3. An alternative representation of the random covariance function in (6) is as follows: With an independent random variable V uniformly distributed over $[0,1]$, we have

$$\Sigma_{\infty,\alpha,\beta} = \mathrm{Cov}\big(J(V,\alpha), J(V,\beta)\,\big|\,\mathcal{F}_\infty\big), \tag{7}$$

with the σ-algebra

$$\mathcal{F}_\infty := \sigma\big\{I_\phi \mid \phi\in\{0,1\}^*\big\}. \tag{8}$$

Remark 4. A related functional limit law for the complexity of Radix Selection, an algorithm to select ranks based on the bit expansions of the data, with a limiting Gaussian process with a covariance function related to (7) can be found in [12, Theorem 1.2]. See [17, Theorem 1.1] for another related functional limit law.

The analysis of QuickVal is usually considered an intermediate step to analyze the original FIND algorithm. Grübel and Rösler [9] already pointed out that a version of FIND such as stated above with $C_n^*(k)$ denoting the number of key comparisons for finding rank k within $(U_1,\dots,U_n)$ does not lead to convergence within $(D[0,1], d_{\mathrm{SK}})$ after the normalization $\alpha\mapsto \frac{1}{n}C_n^*(\lfloor\alpha n\rfloor+1)$, where here and below the convention $C_n^*(n+1) := C_n^*(n)$ is used. To overcome this problem they propose a version that does not stop in case a pivot turns out to be the rank to be selected by including the pivot in the list $S_<$ and proceeding until a list of size 1 is generated. Moreover, their pivots are chosen uniformly at random. The number of key comparisons $C_n'(k)$ for Grübel and Rösler's FIND-version has the property that

$$\left(\frac{1}{n}C_n'\big(\lfloor\alpha n\rfloor+1\big)\right)_{\alpha\in[0,1]} \xrightarrow{d} (S_\alpha)_{\alpha\in[0,1]}, \quad \text{in } (D[0,1], d_{\mathrm{SK}}), \tag{9}$$

see [9, Theorem 4]. Without using random pivots we may also obtain right-continuous limits by just recursively calling $\mathrm{FIND}(S_\geq, 0)$ in case the pivot turns out to be the rank sought. We denote the number of key comparisons for this version by $C_n(k)$, which is close to Grübel and Rösler's FIND-version and also satisfies (9).

The convergence in (9) could only be stated weakly (not almost surely) since Grübel und Rösler's FIND-version due to randomization within the algorithm does not have a natural embedding on a probability space. Note that the formulation of the QuickVal complexity does have such an embedding which, e.g., makes the almost sure convergence in (4) possible. However, it is easy to see that we have the distributional equality

$$\big(C_n\big(|\{U_i\le\alpha : 1\le i\le n\}|\big)\big)_{\alpha\in[0,1]} \stackrel{d}{=} (S_{\alpha,n})_{\alpha\in[0,1]}. \tag{10}$$

This allows to naturally couple the complexities on one probability space, which we call its *natural coupling*. See [7, page 807] for a related discussion of natural couplings.

To transfer Theorem 1 to FIND we need to align jumps to come up with a suitable fluctuation process. The conventions $C_n(0) := C_n(1)$ and $C_n(n+1) := C_n(n)$ are used.

Corollary 5. *Let $C_n(k)$ be the number of key comparisons required to select rank $1 \le k \le n$ within a set of n data by* FIND *with the natural coupling (10). Let $\Lambda_n : [0,1] \to [0,1]$, $n \in \mathbb{N}$, be any (random) monotone increasing bijective function such that $\Lambda_n(\frac{k}{n+1})$ is equal to the element of rank k within $\{U_1, \dots, U_n\}$. Then we have*

$$\left(\frac{C_n(\lfloor t(n+1) \rfloor) - nS_{\Lambda_n(t)}}{\sqrt{n}} \right)_{t \in [0,1]} \xrightarrow{d} G_\infty \quad \text{in } (D[0,1], d_{\mathrm{SK}}),$$

where G_∞ is the process defined in Theorem 1.

The extended abstract is organized as follows: In Section 2 we introduce a novel perturbation argument which is the basis of our analysis. Section 3 contains a criterion for weak convergence of probability measures on $(D[0,1], d_{\mathrm{SK}})$, which is applied in Section 4 to proof Theorem 1 and Corollary 5. In Section 5 further functional fluctuation results are stated for the number of swaps (key exchanges) required by QuickVal (depending on the specific algorithm used to partition S into the sublists $S_<$ and $S_\ge$) as well as functional fluctuation results for cost measures which are based on key comparisons, where the cost of a comparison may depend on the values of the keys.

2. Perturbation of the data

QuickVal splits an interval $[L_\phi, R_\phi)$ by the first value falling into $[L_\phi, R_\phi)$ denoted by U_{τ_ϕ}. Obviously, this implies dependencies between the data U_i and the lengths I_ϕ of the intervals $[L_\phi, R_\phi)$. In the present section we construct a perturbed sequence $(\widetilde{U}_i)_{i \in \mathbb{N}}$ to the data $(U_i)_{i \in \mathbb{N}}$ such that we gain independence of $(\widetilde{U}_i)_{i \in \mathbb{N}}$ from the σ-algebra $\mathcal{F}_\infty$ of the interval lengths defined in (8). In particular, we aim that conditional on $\mathcal{F}_\infty$ the number of data $(\widetilde{U}_1, \dots, \widetilde{U}_n)$ falling into an interval $[L_\phi, R_\phi)$ is binomial $B(n, I_\phi)$ distributed, see Lemma 7 below.

Every value U_i, $i \in \mathbb{N}$, falls successively into subintervals generated by QuickVal until becoming a pivot element. These subintervals correspond to the path between the root of the corresponding binary search tree and the node where U_i is inserted. Let $\phi_i \in \{0,1\}^*$ denote the node where U_i is inserted. Hence, we have $\tau_{\phi_i} = i$ and $U_i = L_{\phi_i} + I_{\phi_i 0}$.

Let $(V_i)_{i \in \mathbb{N}}$ be a sequence of i.i.d. unif$[0,1]$ random variables being independent of $(U_i)_{i \in \mathbb{N}}$. We define

$$\widetilde{U}_i := L_{\phi_i} + I_{\phi_i} V_i. \tag{11}$$

Lemma 6. *The sequence $(\widetilde{U}_i)_{i \in \mathbb{N}}$ defined in (11) consists of i.i.d.* unif$[0,1]$ *distributed random variables and is independent of $\mathcal{F}_\infty$.*

Proof. It suffices to show that $\widetilde{U}_i$ conditional on $\mathcal{F}_\infty$ and $\widetilde{U}_1, \dots, \widetilde{U}_{i-1}$ is uniformly distributed on $[0,1]$ for all $i \in \mathbb{N}$. We use infinitesimal notation to denote this claim by

$$\mathbb{P}\left(\widetilde{U}_i \in \mathrm{d}u \,\middle|\, \mathcal{F}_\infty, \widetilde{U}_1, \dots, \widetilde{U}_{i-1} \right) = \mathbf{1}_{[0,1]}(u)\mathrm{d}u, \quad i \in \mathbb{N}.$$

For each $i \in \mathbb{N}$ the random variables $\widetilde{U}_i$ and U_i fall into the same interval $[L_{\phi_i}, R_{\phi_i})$, hence $\phi_1, \dots, \phi_{i-1}$ are determined by $\widetilde{U}_1, \dots, \widetilde{U}_{i-1}$. Let us additionally condition on ϕ_i, then, by definition,

$$\mathbb{P}\left(\widetilde{U}_i \in \mathrm{d}u \,\middle|\, \mathcal{F}_\infty, \widetilde{U}_1, \dots, \widetilde{U}_{i-1}, \phi_i \right) = \frac{1}{I_{\phi_i}} \mathbf{1}_{[L_{\phi_i}, R_{\phi_i})}(u)\mathrm{d}u.$$

Note that ϕ_i denotes one of the i external nodes of the binary search tree with internal nodes denoted by $\phi_1, \dots, \phi_{i-1}$. We denote by Ext_{i-1} the set of the labels of these external nodes. Hence, conditional on $\mathcal{F}_\infty, \phi_1, \dots, \phi_{i-1}$ the label ϕ_i is chosen from Ext_{i-1} with probability given by the length of the corresponding interval, i.e., $\mathbb{P}(\phi_i = \phi \,|\, \mathcal{F}_\infty, \phi_1, \dots, \phi_{i-1}) = I_\phi$ for all $\phi \in \mathrm{Ext}_{i-1}$. Thus, by the law of total probability we obtain

$$\mathbb{P}\left(\widetilde{U}_i \in \mathrm{d}u \,\middle|\, \mathcal{F}_\infty, \widetilde{U}_1, \dots, \widetilde{U}_{i-1}\right) = \sum_{\phi \in \mathrm{Ext}_{i-1}} I_\phi \frac{1}{I_\phi} \mathbf{1}_{[L_\phi, R_\phi)}(u)\mathrm{d}u = \mathbf{1}_{[0,1]}(u)\mathrm{d}u.$$

This implies the assertion. □

The $\widetilde{U}_i$ are now coupled with the U_i but independent of the I_ϕ. To compare with the number of key comparisons required by QuickVal$((U_1, \dots, U_n), \alpha)$ we define

$$\widetilde{S}_{\alpha,k,n} := \sum_{i=1}^{n} \mathbf{1}_{[L_{\alpha,k}, R_{\alpha,k})}(\widetilde{U}_i).$$

Lemma 7. *Conditional on $I_{\alpha,k}$ we have that $\widetilde{S}_{\alpha,k,n}$ has the binomial $B(n, I_{\alpha,k})$ distribution. Moreover, for all $\alpha \in [0,1]$, $n \in \mathbb{N}$ and $0 \le k \le n$ we have*

$$S_{\alpha,k,n} \le (\widetilde{S}_{\alpha,k,n} - 1)^+ \le S_{\alpha,k,n} + k - 1. \tag{12}$$

Proof. The conditional distribution of $\widetilde{S}_{\alpha,k,n}$ follows from Lemma 6. Recall that $S_{\alpha,k,n}$ is defined as $\sum_{i=\tau_{\alpha,k}}^{n} \mathbf{1}\{L_{\alpha,k-1} \le U_i < R_{\alpha,k-1}\}$. By definition, U_i and $\widetilde{U}_i$ are in the interval $[L_{\phi_i}, R_{\phi_i})$ for all $i \in \mathbb{N}$. If $U_i \in (L_{\alpha,k}, R_{\alpha,k})$, then U_i appears as a pivot after the k-th pivot. Hence, its interval $[L_{\phi_i}, R_{\phi_i})$ and thus also $\widetilde{U}_i$ are contained in $(L_{\alpha,k}, R_{\alpha,k})$. The k-th pivot $U_{\tau_{\alpha,k}}$ itself does not contribute to $S_{\alpha,k,n}$, which implies the left inequality stated in the present lemma.

For the right inequality, assume for some $i \in \mathbb{N}$ that the perturbed value $\widetilde{U}_i$ is in $(L_{\alpha,k}, R_{\alpha,k})$, but U_i is not. Then the corresponding interval (L_{ϕ_i}, R_{ϕ_i}) must contain $(L_{\alpha,k}, R_{\alpha,k})$, thus making U_i a pivot that appears before the k-th pivot. Since there are only k such pivots, the right inequality follows. □

3. On Weak Convergence in $D[0,1]$

To prove the convergence in distribution in Theorem 1 within the space $(D[0,1], d_{\mathrm{SK}})$ we use the following Proposition 8. It can be proved by classical tools of weak convergence theory based on a study of the modulus of continuity and the Arzelà–Ascoli theorem in form of a general theorem of Billingsley [2, Theorem 13.2].

Proposition 8. *Let $X_1, X_2, \dots$ be a sequence of random variables in $(D[0,1], d_{SK})$. Suppose that for every $K \in \mathbb{N}$, there exist random càdlàg step functions $X_1^K, X_2^K, \dots$ with all jumps contained in $\{U_\phi \mid \phi \in \{0,1\}^*, |\phi| < K\}$. If*

(i) *for all $r \in \mathbb{N}$ and $\alpha_1, \dots, \alpha_r \in [0,1]$, the marginals $\mathcal{L}(X_n(\alpha_1), \dots, X_n(\alpha_r))$ converge weakly to some distribution $\mu_{\alpha_1,\dots,\alpha_r}$,*

(ii) *for all $\varepsilon > 0$,*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\left\|X_n - X_n^K\right\|_\infty > \varepsilon\right) \to 0, \tag{13}$$

then $(X_n)_{n\in\mathbb{N}}$ converges in distribution to a random variable X on $(D[0,1], d_{SK})$, and for all $r \in \mathbb{N}$ and $\alpha_1, \dots, \alpha_r \in [0,1]$ we have

$$\mathcal{L}(X(\alpha_1), \dots, X(\alpha_r)) = \mu_{\alpha_1,\dots,\alpha_r}. \tag{14}$$

4. Proof of Theorem 1

To split the contributions to the process G_n into costs resulting from above and below a level $K \in \mathbb{N}$ we define

$$G_{\alpha,k,n} := \frac{S_{\alpha,k,n} - nI_{\alpha,k}}{\sqrt{n}} \tag{15}$$

as the normalized fluctuations of the contribution at level k, and set

$$G^{\le K}_{\alpha,n} := \sum_{k=0}^{K} G_{\alpha,k,n}, \qquad G^{\le K}_{n} := \big(G^{\le K}_{\alpha,n}\big)_{\alpha\in[0,1]}, \qquad G^{>K}_{\alpha,n} := \sum_{k=K+1}^{\infty} G_{\alpha,k,n}. \tag{16}$$

Hence, $G_{\alpha,n} = G^{\le K}_{\alpha,n} + G^{>K}_{\alpha,n}$. Analogously, for the perturbed values $\widetilde{S}_{k,n}$ we define

$$W_{\alpha,k,n} := \frac{\widetilde{S}_{\alpha,k,n} - nI_{\alpha,k}}{\sqrt{n}}, \qquad W^{\le K}_{\alpha,n} := \sum_{k=0}^{K} W_{\alpha,k,n} \qquad W^{\le K}_{n} := \big(W^{\le K}_{\alpha,n}\big)_{\alpha\in[0,1]}. \tag{17}$$

Lemma 9. *For all $K \in \mathbb{N}$ we have convergence in distribution of $(G^{\le K}_n)_{n\in\mathbb{N}}$ towards a mixture $G^{\le K}_\infty = (G^{\le K}_{\alpha,\infty})_{\alpha\in[0,1]}$ of centered Gaussian processes within $\|\cdot\|_\infty$. Conditional on $\mathcal{F}_\infty$, the limit $G^{\le K}_\infty$ is a centered Gaussian process with covariance function given, for $\alpha, \beta \in [0,1]$ by*

$$\operatorname{Cov}\big(G^{\le K}_{\alpha,\infty}, G^{\le K}_{\beta,\infty} \mid \mathcal{F}_\infty\big) = \sum_{k=0}^{K\wedge J}\sum_{j=0}^{K} I_{\alpha,j\vee k} + \big(1 + (K\wedge J)\big)\sum_{j=J+1}^{K} I_{\beta,j} - S^{\le K}_{\alpha} S^{\le K}_{\beta}, \tag{18}$$

where $J = J(\alpha,\beta)$ is as in Theorem 1 and $S^{\le K}_\alpha := \sum_{k=0}^{K} I_{\alpha,k}$. The stated convergence in distribution also holds conditionally in $\mathcal{F}_\infty$, i.e., we have almost surely that $\mathcal{L}(G^{\le K}_n \mid \mathcal{F}_\infty)$ converges weakly towards $\mathcal{L}(G^{\le K}_\infty \mid \mathcal{F}_\infty)$.

Proof. First note that by Lemma 7 we have $\big\|G^{\le K}_n - W^{\le K}_n\big\|_\infty < K^2/\sqrt{n}$, so it suffices to show the lemma for $W^{\le K}_n$. Conditional on $\mathcal{F}_\infty$, the value of $W^{\le K}_{\alpha,n}$ is given by

$$W^{\le K}_{\alpha,n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=0}^{K} \mathbf{1}\{L_{\alpha,k} \le \widetilde{U}_i < R_{\alpha,k}\} - (R_{\alpha,k} - L_{\alpha,k}), \tag{19}$$

thus the 2^k different values of the process $W^{\le K}_n$ can be expressed as the sum of n centered, bounded i.i.d. random vectors, scaled by $1/\sqrt{n}$. By the multivariate central limit theorem, these converge towards a multivariate, centered normal variable. As the positions of the jumps, still conditional on $\mathcal{F}_\infty$, are fixed, we have convergence of $W^{\le K}_n$ and thus also of $G^{\le K}_n$ towards a Gaussian process. Define $X_{\alpha,k} := \mathbf{1}\{L_{\alpha,k-1} \le \widetilde{U}_1 < R_{\alpha,k-1}\}$. The covariance

function then is given by

$$\begin{aligned}\operatorname{Cov}\big(G_{\alpha,\infty}^{\leq K}, G_{\beta,\infty}^{\leq K} \mid \mathcal{F}_\infty\big) &= \operatorname{Cov}\Big(\sum_{k=0}^{K} X_{\alpha,k}, \sum_{j=0}^{K} X_{\beta,j} \;\Big|\; \mathcal{F}_\infty\Big) \\ &= \sum_{k=0}^{K}\sum_{j=0}^{K} \mathbb{E}\left[X_{\alpha,k}X_{\beta,j} \;\middle|\; \mathcal{F}_\infty\right] - S_\alpha^{\leq K} S_\beta^{\leq K} \\ &= \sum_{k=0}^{K\wedge J}\sum_{j=0}^{K} I_{\alpha,k\vee j} + \big(1+(K\wedge J)\big)\sum_{j=J+1}^{K} I_{\beta,j} - S_\alpha^{\leq K} S_\beta^{\leq K}. \qquad (20)\end{aligned}$$

The assertion follows. □

To see that the covariance functions in (20) converge towards the covariance function of G_∞ stated in Theorem 1 we restate a Lemma of Grübel and Rösler [9, Lemma 1] that the maximal length of the intervals at a level is decreasing geometrically with increasing levels. It is obtained observing that $\mathbb{E}\left[\sum_{|\phi|=k} I_\phi^2\right] = (2/3)^k$ and states:

Lemma 10. *There exists an a.s. finite random variable K_1 such that for all $k \geq K_1$:*

$$\max_{\alpha\in[0,1]} I_{\alpha,k} \leq k\left(\frac{2}{3}\right)^{k/2}. \qquad (21)$$

Lemma 10 implies that the covariance functions of $G_\infty^{\leq K}$ from (18) converge a.s. to the covariance function of G_∞ from (6).

For the costs from levels $k > K$ we find:

Proposition 11. *For all $\varepsilon, \eta > 0$ there are constants $K, N \in \mathbb{N}$ such that for all $n \geq N$*

$$\mathbb{P}\left(\|G_n^{>K}\|_\infty > \eta\right) < \varepsilon. \qquad (22)$$

We postpone the proof of the latter proposition and first use Proposition 11 and Lemma 9 to show convergence of the finite-dimensional distributions, denoted fdd-convergence.

Lemma 12. *We have fdd-convergence of G_n towards G_∞.*

Proof. For any K, we can split $G_n = G_n^{\leq K} + G_n^{>K}$. By Lemma 9, we have

$$G_n^{\leq K} \xrightarrow{\text{fdd}} G_\infty^{\leq K} \qquad (n\to\infty).$$

Furthermore, because the covariance functions of the $G_\infty^{\leq K}$ converge a.s., we obtain

$$G_\infty^{\leq K} \xrightarrow{\text{fdd}} G_\infty \qquad (K\to\infty)$$

by Lévy's continuity theorem. Hence, for all $\alpha_1, \ldots, \alpha_\ell \in [0,1]$ and all $t_1, \ldots, t_\ell \in \mathbb{R}$ we find a sequence $(K_n)_{n\in\mathbb{N}}$ in $\mathbb{N}$ such that

$$\mathbb{P}\left(G_{\alpha_1,n}^{\leq K_n} < t_1, \ldots, G_{\alpha_\ell,n}^{\leq K_n} < t_\ell\right) \longrightarrow \mathbb{P}\left(G_{\alpha_1,\infty} < t_1, \ldots, G_{\alpha_\ell,\infty} < t_\ell\right) \quad (n\to\infty).$$

Now, since $\|G_n^{>K_n}\|_\infty \to 0$ in probability by Proposition 11 the claim of Lemma 12 follows by Slutzky's theorem. □

To prepare for the proof of Proposition 11, we show that the fluctuations on each level are also at least geometrically decreasing. Recall K_1 from Lemma 10.

Lemma 13. *There exists a constant $a > 1$ such that for all $k, n \in \mathbb{N}$*

$$\mathbb{P}\left(\max_{\alpha\in[0,1]}|W_{\alpha,k,n}| > a^{-k},\, K_1 \le k\right) \le b(k) + c(k,n)$$

with $b(k), c(k,n) \ge 0$ such that

$$\sum_{k=1}^{\infty} b(k) < \infty \quad \text{and} \quad \sum_{k=1}^{\lfloor (9/2)\log n\rfloor} c(k,n) \to 0 \quad (n\to\infty). \tag{23}$$

For the proof of Lemma 13 we require the following Chernoff bound:

Lemma 14. *Let S_n be binomial $B(n,p)$ distributed for some $p \in [0,1]$ and $n \in \mathbb{N}$ and let $\mu := \mathbb{E}[S_n]$, $\varepsilon \ge 0$. Then*

$$\mathbb{P}\left(S_n \notin \left((1-\varepsilon)\mu, (1+\varepsilon)\mu\right)\right) \le 2\exp\left(-\frac{\varepsilon^2\mu}{2+\varepsilon}\right).$$

Proof. Combine upper and lower bound in McDiarmid [15, Theorem 2.3]. □

Proof of Corollary 13. Fix some $\alpha \in [0,1]$. Conditionally on $I_{k,\alpha}$, the costs $\widetilde{S}_{\alpha,k,n}$ are $B(n, I_{k,\alpha})$-distributed by Lemma 6. The Chernoff bound in Lemma 14 implies

$$\begin{aligned}
\mathbb{P}\left(|W_{\alpha,k,n}| > a^{-k} \,\middle|\, I_{\alpha,k}\right) &= \mathbb{P}\left(\left|\widetilde{S}_{\alpha,k,n} - nI_{\alpha,k}\right| > \sqrt{n}a^{-k} \,\middle|\, I_{\alpha,k}\right) \\
&\le 2\exp\left(-\frac{na^{-2k}}{nI_{\alpha,k}(2+\sqrt{n}a^{-k}/(nI_{\alpha,k}))}\right) \\
&= 2\exp\left(-\left(2a^{2k}I_{\alpha,k} + a^k/\sqrt{n}\right)^{-1}\right). \qquad (24)
\end{aligned}$$

For the two summands in the exponent in (24) we have the following behavior: Summand $2a^{2k}I_{k,\alpha}$ is falling geometrically with k for sufficiently small $a > 1$. Summand $a^k/\sqrt{n}$ is falling with n, but growing with k. To separate these two contributions, note that $\exp(x^{-1}) \ge \frac{1}{m!}x^{-m}$ and thus $\exp(-x^{-1}) \le m!x^m$ for all $m \in \mathbb{N}$ and $x \ge 0$. Choosing $m = 7$, we obtain

$$\begin{aligned}
\mathbb{P}\left(|W_{\alpha,k,n}| > a^{-k} \,\middle|\, I_{\alpha,k}\right) &\le 2\cdot 7!\left(2a^{2k}I_{\alpha,k} + a^k/\sqrt{n}\right)^7 \\
&\le 2^{14}7!a^{14k}I_{\alpha,k}^7 + 2^7 7! a^{7k} n^{-7/2}
\end{aligned}$$

by convexity of $x \mapsto x^7$. Note that the 2^k intervals at level k have lengths $I_{\alpha,k}$ summing to 1. Hence,

$$\mathbb{P}\left(\max_{\alpha\in[0,1]}|W_{\alpha,k,n}| > a^{-k} \,\middle|\, I_{\alpha,k}\right) \le 2^{14}7!a^{14k}\max_{\alpha\in[0,1]} I_{\alpha,k}^6 + 2^7 7!(2a^7)^k n^{-7/2}.$$

When furthermore $K_1 \le k$, by Lemma 10 the length $I_{\alpha,k}$ is bounded by $k\frac{2}{3}^{k/2}$, hence

$$\mathbb{P}\left(\max_{\alpha\in[0,1]}|W_{\alpha,k,n}| > a^{-k},\, K_1 \le k\right) \le 2^{14}7!ka^{14k}\left(\frac{2}{3}\right)^{3k} + 2^7 7!(2a^7)^k n^{-7/2}.$$

Define the first summand on the right hand side of the latter inequality by $b(k)$ and the second summand by $c(k,n)$. For all $1 < a < (3/2)^{3/14}$ the $b(k)$ form a convergent series. To

also show the second statement in (23) note that

$$\sum_{k=1}^{\lfloor (9/2)\log n\rfloor} c(k,n) = \mathrm{O}\Big(\big(2a^7\big)^{(9/2)\log n} n^{-7/2}\Big) = \mathrm{O}\Big(n^{(9/2)\log 2+(9/2)\cdot 7\log a-7/2}\Big).$$

The latter O-term converges to 0 for $(63/2)\log a < 7/2-(9/2)\log 2 \approx 0.381\ldots$, thus we may choose a as required. □

We are now prepared for the proof of Proposition 11.

Proof of Proposition 11. Let $\varepsilon, \eta > 0$. To K_1 from Lemma 10 and a and $b(k)$ from Lemma 13 we choose K sufficiently large such that

$$\mathbb{P}(K_1 > K) \le \frac{\varepsilon}{4}, \qquad \sum_{k=K}^{\infty} a^{-k} \le \frac{\eta}{4} \qquad \text{and} \qquad \sum_{k=K}^{\infty} b(k) \le \frac{\varepsilon}{4}. \tag{25}$$

Let H_n be the maximum amount of steps needed by $\mathrm{QuickVal}((U_1,\ldots,U_n),\alpha)$ for any α. Thus, H_n is also the height of the binary search tree built from $U_1,\ldots,U_n$. Devroye [4] showed that the height has expectation $\mathbb{E}[H_n] = \gamma\log n + o(\log n)$ with $\gamma = 4.311\ldots$ Reed [16] further showed that $\mathrm{Var}(H_n) = \mathrm{O}(1)$. Hence, we can choose N sufficiently large such that

$$\mathbb{P}\left(H_n > \lfloor (9/2)\log n\rfloor\right) < \frac{\varepsilon}{4}, \qquad \sum_{k=1}^{\lfloor (9/2)\log n\rfloor} c(k,n) \le \frac{\varepsilon}{4}, \qquad \text{and} \qquad \frac{((9/2)\log n)^2}{\sqrt{n}} \le \frac{\eta}{4} \tag{26}$$

for all $n \ge N$. Subsequently we use the decomposition

$$G_{\alpha,n}^{>K} = \sum_{k=K+1}^{\lfloor (9/2)\log n\rfloor} \frac{S_{\alpha,k,n} - nI_{\alpha,k}}{\sqrt{n}} + \sum_{\lfloor (9/2)\log n\rfloor+1}^{\infty} \frac{S_{\alpha,k,n} - nI_{\alpha,k}}{\sqrt{n}} =: \Gamma_n + G_{\alpha,n}^{>\lfloor (9/2)\log n\rfloor}$$

and consider the event

$$A_n := \{K_1 > K\} \cup \{H_n > \lfloor (9/2)\log n\rfloor\}.$$

We have $\mathbb{P}(A_n) < \varepsilon/2$ for all $n \ge N$. Note that on A_n^c (the complement of A_n) we have $S_{\alpha,k,n} = 0$ for all $k > \lfloor (9/2)\log n\rfloor$ and also the bound on $I_{\alpha,k}$ from Lemma 10 applies, hence

$$\begin{aligned}\left|G_{\alpha,n}^{>\lfloor (9/2)\log n\rfloor}\right| &\le \sum_{\lfloor (9/2)\log n\rfloor+1}^{\infty} \sqrt{n} I_{\alpha,k} \le \sum_{\lfloor (9/2)\log n\rfloor+1}^{\infty} \sqrt{n} k \left(\frac{2}{3}\right)^{k/2} \\ &= \mathrm{O}\Big(n^{1/2-(9/4)\log(3/2)}\log n\Big) = \mathrm{o}(1)\end{aligned}$$

since $(9/4)\log(3/2) = 0.912\ldots$ Hence, we can enlarge N so that on A_n^c we have $\left|G_{\alpha,n}^{>\lfloor (9/2)\log n\rfloor}\right| < \eta/2$ for all $n \ge N$. This implies the bound

$$\begin{aligned}\mathbb{P}\left(|G_{\alpha,n}^{>K}| > \eta\right) &\le \mathbb{P}(A_n) + \mathbb{P}\left(\left\{|G_{\alpha,n}^{>K}| > \eta\right\} \cap A_n^c\right) \\ &\le \frac{\varepsilon}{2} + \mathbb{P}\left(\left\{|\Gamma_n| > \frac{\eta}{2}\right\} \cap A_n^c\right) + \mathbb{P}\left(\left\{\left|G_{\alpha,n}^{>\lfloor (9/2)\log n\rfloor}\right| > \frac{\eta}{2}\right\} \cap A_n^c\right).\end{aligned} \tag{27}$$

Note that the third summand in (27) is 0. Hence, it remains to bound the second summand in (27). To this end note that

$$|\Gamma_n| \le \sup_{\alpha\in[0,1]} \sum_{k=K+1}^{\lfloor (9/2)\log n\rfloor} \left(\left| \frac{\widetilde{S}_{\alpha,k,n} - nI_{\alpha,k}}{\sqrt{n}} \right| + \left| \frac{S_{\alpha,k,n} - \widetilde{S}_{\alpha,k,n}}{\sqrt{n}} \right| \right)$$

$$\le \left(\sum_{k=K+1}^{\lfloor (9/2)\log n\rfloor} \max_{\alpha\in[0,1]} |W_{\alpha,k,n}| \right) + \frac{\lfloor (9/2)\log n\rfloor^2}{\sqrt{n}}, \tag{28}$$

where Lemma 7 is used. The third relation in (26) assures that the second term in (28) is smaller than $\eta/4$. In view of the second relation in (25) and (28), we have

$$\left\{ |\Gamma_n| > \frac{\eta}{2} \right\} \cap A_n^c \subset \bigcup_{k=K}^{\lfloor (9/2)\log n\rfloor} \left\{ \max_{\alpha\in[0,1]} |W_{\alpha,k,n}| > a^{-k},\, K_1 \le k \right\}.$$

Thus, Lemma 13 together with (25) and (26) imply that the second summand in (27) is bounded by $\varepsilon/2$. This implies the assertion. □

Proof of Theorem 1. We apply Lemma 8 to G_n and $G_n^{\le K}$. The first condition, fdd convergence, is Lemma 12, the second condition is Proposition 11. □

We now transfer the fluctuation result for QuickVal in Theorem 1 to the original Find process in Corollary 5.

Proof of Corollary 5. Let $\widetilde{F}_n$ be the inverse of Λ_n in the statement of Corollary 5. By definition of Λ_n, the value of the element $U_{(k)}$ of rank k within $U_1, \dots, U_n$ is given by $\frac{k}{n+1}$, so

$$\big\lfloor (n+1)\widetilde{F}_n(\alpha) \big\rfloor = \big| \{ U_i \le \alpha \mid 1 \le i \le n \} \big| \tag{29}$$

for all $\alpha \in [0,1)$. Thus, $C_n\big(\lfloor (n+1)\widetilde{F}_n(\alpha)\rfloor\big) = S_{\alpha,n}$ a.s. for all $\alpha \in [0,1]$, see (10). For $\alpha = 1$ note that $\widetilde{F}_n(\alpha) = 1$ and $C_n(n+1) = C_n(n)$ by definition. The Skorokhod distance d_{SK} is then bounded by

$$\begin{aligned} d_{\mathrm{SK}}\left(G_n, \left(\frac{C_n(\lfloor t(n+1)\rfloor) - nS_{\Lambda_n(t)}}{\sqrt{n}} \right)_{t\in[0,1]} \right) &= d_{\mathrm{SK}}\left(G_n, \left(\frac{S_{\Lambda_n(t),n} - nS_{\Lambda_n(t)}}{\sqrt{n}} \right)_{t\in[0,1]} \right) \\ &= d_{\mathrm{SK}}\big(G_n, (G_{\Lambda_n(t),n})_{t\in[0,1]} \big) \\ &\le \|\widetilde{F}_n - \mathrm{id}\|_\infty. \end{aligned}$$

By (29), $\widetilde{F}_n$ is close to the empirical distribution function and thus converges a.s. uniformly to the identity id by the Glivenko–Cantelli theorem. The statement of Corollary 5 then follows from Slutzky's theorem. □

5. Further cost measures

In this section we sketch results analogous to Theorem 1 for other cost measures than the number of key comparisons. We consider the number of swaps required by QuickVal which however depends on the implementation of the procedure to partition the input $(U_1, \dots, U_n)$ into the sublists $S_<$ and $S_\ge$. We consider two such procedures, the one originally proposed

by Hoare [10] and one that is attributed to Lomuto, see [1, 3, 13]. Our results are stated in Subsection 5.1.

As a further cost measure we consider the model where the costs to compare two keys may depend on their values, e.g., the number of bit comparisons required to compare them when they are given by their binary expansions. The total cost for all key comparisons required by QuickVal$((U_1,\dots,U_n),\alpha)$ or FIND$((U_1,\dots,U_n),k)$ is no longer determined by the fact that the ranks of $(U_1,\dots,U_n)$ form an uniformly random permutation. Here, the distribution of the U_i matters. We only consider the uniform distribution as in the previous sections and hope to report on other distributions in the full paper version of this extended abstract. Our results are stated in Subsection 5.2. A probabilistic analysis for the number of bit comparisons of the related Quicksort algorithms was given in [6, 5].

5.1. **Number of swaps.** Usually, QuickSelect is implemented in-place, meaning that it only requires the memory for the list S of values and a bounded amount of additional memory. This is achieved by swapping values within S so that the elements of $S_<$ and $S_\geq$ are contained in contiguous parts of the list. Such a procedure is called *partition*. There are various procedures of *partition*.

The original procedure by Hoare [10] searches the list S from both ends at once: It repeatedly finds the index $i = \min\{2 \leq i \leq n \mid U_i > U_1\}$ of the leftmost element bigger than the pivot and the index $j := \max\{2 \leq j \leq n \mid U_j < U_1\}$ of the rightmost element smaller than the pivot. If $i < j$, it swaps U_i and U_j. Else the algorithm terminates.

A simpler, but less efficient implementation is the so-called Lomuto partition scheme [1, 3, 13] that only searches from one end of S. It keeps track of the amount i of elements at the start of the list it has already swapped. In every step, it finds the index $j := \max\{2 \leq j \leq n \mid U_j < U_1\}$ of the rightmost element smaller than the pivot. If $i+1 < j$, it swaps U_{i+1} and U_j and increases i by one. Otherwise the algorithm terminates.

Both partition schemes only compare elements to the pivot, so the model of randomness is preserved within the sublists $S_<$ and $S_\geq$. However, their original order is not preserved, so QuickSelect run on $U_1,\dots,U_n$ will usually not select the same pivots as QuickSelect on $U_1,\dots,U_{n+1}$. For convenience, we assume that the pivot to split a sublist S' of S is the element of S' that came first in the original list S. We call this choice of the pivots a *suitable embedding*.

5.1.1. *Hoare's partition.* For Hoare's partition, via a hypergeometric distribution the expected number of swaps in step k given $\mathcal{F}_\infty$ is approximately $nI_{\alpha,k+1}(I_\alpha - I_{\alpha,k})/I_{\alpha,k}$, which leads to the limit process $L = (L_\alpha)_{\alpha\in[0,1]}$ given by

$$L_\alpha := \sum_{k=0}^{\infty} \frac{I_{\alpha,k+1}(I_\alpha - I_{\alpha,k})}{I_{\alpha,k}}, \quad \alpha \in [0,1].$$

It is now possible to study the fluctuations by their contributions from the individual levels and combine them as for the number of key comparisons above. Since we are still in the range of the central limit theorem we again obtain a mixture of centered Gaussian processes. To be explicit, first denote by Z_ϕ the limit of the $G_{\alpha,k,n}$ as $n \to \infty$ where $\phi = \phi(\alpha,k) \in \{0,1\}^k$. Further, denote by $\{Y_\phi \mid \phi \in \{0,1\}^*\}$ a set of i.i.d. $\mathcal{N}(0,1)$ random variables being independent

of $\{Z_\phi \,|\, \phi \in \{0,1\}^*\}$ and of $\mathcal{F}_\infty$. Then the limiting process $G^{\text{swap}} = (G^{\text{swap}}_\alpha)_{\alpha\in[0,1]}$ is given by

$$G^{\text{swap}}_\alpha := \sum_{\phi\in\{\phi(k,\alpha)\,|\,k\in\mathbb{N}_0\}} Y_\phi \frac{I_{\phi0}I_{\phi1}}{I_\phi^{3/2}} + Z_{\phi0}\frac{I_{\phi1}}{I_\phi} + Z_{\phi1}\frac{I_{\phi0}}{I_\phi} - Z_\phi\frac{I_{\phi0}I_{\phi1}}{I_\phi}, \quad \alpha\in[0,1]. \tag{30}$$

Then we have the following result for key exchanges corresponding to Theorem 1.

Theorem 15. *Let $K_{\alpha,n}$ be the number of key exchanges required by* QuickVal$((U_1,\dots,U_n),\alpha)$ *with Hoare's partition algorithm in a suitable embedding. Then, as $n\to\infty$, we have*

$$\left(\frac{K_{\alpha,n}-nL_\alpha}{\sqrt{n}}\right)_{\alpha\in[0,1]} \xrightarrow{d} G^{\text{swap}} \quad \text{in } (D[0,1],d_{\text{SK}}).$$

5.1.2. *Lomuto's partition.* The Lomuto partition is simpler to implement and much easier to analyze. The Lomuto partition swaps every element smaller than the pivot, so the amount of swaps at some path $\phi\in\{0,1\}^*$ is given by $S_{\phi0}+1$. With the Z_ϕ introduced in Subsection 5.1.1 we find that $(Z_\phi)_{\phi\in\{0,1\}^*}$ is a mixture of centered Gaussian processes with conditional covariance function given by

$$\operatorname{Cov}(Z_\phi, Z_\psi \mid \mathcal{F}_\infty) = I_{\phi\vee\psi} - I_\phi I_\psi, \quad \phi,\psi\in\{0,1\}^*,$$

where $I_{\phi\vee\psi}$ is the length of the interval $[L_\phi,R_\phi)\cap[L_\psi,R_\psi)$, thus $I_{\phi\vee\psi}$ is only nonzero if one of ψ and ϕ is a prefix of the other. Then the limiting process $G^{\text{Lo}} = (G^{\text{Lo}}_\alpha)_{\alpha\in[0,1]}$ is given by

$$G^{\text{Lo}}_\alpha = \sum_{k=0}^{\infty} Z_{\phi(\alpha,k)0}, \quad \alpha\in[0,1].$$

We can directly apply Lemma 13 and our proof for the number of key comparisons can be straightforwardly transferred.

Theorem 16. *Let $K^{\text{Lo}}_{\alpha,n}$ be the number of key exchanges required by* QuickVal$((U_1,\dots,U_n),\alpha)$ *with Lomuto's partition procedure in a suitable embedding. Then, as $n\to\infty$, we have*

$$\left(\frac{K^{\text{Lo}}_{\alpha,n}-n\sum_{k=0}^{\infty} I_{\phi(\alpha,k)0}}{\sqrt{n}}\right)_{\alpha\in[0,1]} \xrightarrow{d} G^{\text{Lo}} \quad \text{in } (D[0,1],d_{\text{SK}}).$$

5.2. **Number of bit comparisons.** We now consider the model where the cost to compare two keys depends on their values. These costs are described by a measurable *cost function* $\beta : [0,1]^2\to[0,\infty)$, and we require that they have a polynomial tail, that is: There are constants $c,\varepsilon>0$ such that for all $u\in[0,1], x\in\mathbb{N}$ and for $V\sim\text{unif}[0,1]$

$$\mathbb{P}\left(\beta(u,V)\geq x\right) \leq cx^{-1/\varepsilon}.$$

This condition is called (c,ε)-*tameness*, see Matterer [14], and β is called to be ε-*tame* if it is (c,ε)-tame for some $c>0$. Note that, e.g., β counting the number of bit comparisons is ε-tame for all $\varepsilon>0$. The costs of QuickVal$((U_1,\dots,U_n),\alpha)$ in this model are given by

$$S^\beta_{\alpha,n} := \sum_{k=0}^{\infty}\sum_{\tau_{\alpha,k}<i\leq n} \mathbf{1}_{[L_{\alpha,k},R_{\alpha,k})}(U_i)\beta(U_{\tau_{\alpha,k}},U_i)$$

and the limit is, with $V\sim\text{unif}[0,1]$ being independent of the $U_1,\dots,U_n$, given as

$$S^\beta_{\alpha,\infty} := \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{[L_{\alpha,k},R_{\alpha,k})}(V)\beta(U_{\tau_{\alpha,k}},V)\,\middle|\,\mathcal{F}_\infty\right]$$

Matterer [14, Theorem 6.4 and Theorem 6.14] shows for $\varepsilon < \frac{1}{2}$ that for fixed $\alpha \in [0,1]$ the resulting residual

$$G^{\beta}_{\alpha,n} := \frac{S^{\beta}_{\alpha,n} - nS^{\beta}_{\alpha,\infty}}{\sqrt{n}}$$

converges to a mixed centered Gaussian random variable $G^{\beta}_{\alpha,\infty}$ in distribution and with all moments. It is possible to combine them to a mixture of centered Gaussian processes

$$G^{\beta}_{\infty} = (G^{\beta}_{\alpha,\infty})_{\alpha\in[0,1]}, \tag{31}$$

defined by the conditional covariance functions given, with $X^{\beta}_{\alpha,k} := \mathbf{1}_{[L_{\alpha,k},R_{\alpha,k})}(V)\cdot\beta(U_{\tau_{\alpha,k,n}},V)$, by

$$\operatorname{Cov}\big(G^{\beta}_{\alpha,\infty}, G^{\beta}_{\gamma,\infty} \mid \mathcal{F}_{\infty}\big) = \operatorname{Cov}\left(\sum_{k=0}^{\infty} X^{\beta}_{\alpha,k}, \sum_{k=0}^{\infty} X^{\beta}_{\gamma,k} \,\middle|\, \mathcal{F}_{\infty}\right), \quad \alpha,\gamma \in [0,1]. \tag{32}$$

It can be shown that the latter expression is a.s. well-defined. We have the following result corresponding to Theorem 1.

Theorem 17. *Let β be an ε-tame cost function with $\varepsilon < \frac{1}{4}$. Then we have*

$$\left(\frac{S^{\beta}_{\alpha,n} - nS^{\beta}_{\alpha,\infty}}{\sqrt{n}}\right)_{\alpha\in[0,1]} \xrightarrow{d} G^{\beta}_{\infty} \quad \text{in } (D[0,1], d_{\mathrm{SK}}),$$

where G^{β}_{∞} is the mixture of centered Gaussian processes defined in (31).

References

[1] Jon Bentley, *Programming pearls*, Addison-Wesley, Reading, Mass., 2000.

[2] Patrick Billingsley, *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999, A Wiley-Interscience Publication. MR MR1700749 (2000e:60008)

[3] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to algorithms*, third ed., MIT Press, Cambridge, MA, 2009. MR 2572804

[4] Luc Devroye, *A note on the height of binary search trees*, J. ACM **33** (1986), no. 3, 489–498.

[5] James Allen Fill, *Distributional convergence for the number of symbol comparisons used by QuickSort*, Ann. Appl. Probab. **23** (2013), no. 3, 1129–1147. MR 3076680

[6] James Allen Fill and Svante Janson, *The number of bit comparisons used by Quicksort: an average-case analysis*, Electron. J. Probab. **17** (2012), no. 43, 22. MR 2928726

[7] James Allen Fill and Jason Matterer, *QuickSelect tree process convergence, with an application to distributional convergence for the number of symbol comparisons used by worst-case find*, Combin. Probab. Comput. **23** (2014), no. 5, 805–828. MR 3249225

[8] James Allen Fill and Takehiko Nakama, *Distributional convergence for the number of symbol comparisons used by quickselect*, Advances in Applied Probability **45** (2013), no. 2, 425–450.

[9] Rudolf Grübel and Uwe Rösler, *Asymptotic distribution theory for Hoare's selection algorithm*, Adv. in Appl. Probab. **28** (1996), no. 1, 252–269. MR 1372338

[10] C. A. R. Hoare, *Algorithm 63: Partition*, Commun. ACM **4** (1961), no. 7, 321.

[11] ———, *Algorithm 65: Find*, Commun. ACM **4** (1961), no. 7, 321–322.

[12] Kevin Leckey, Ralph Neininger, and Henning Sulzbach, *Process convergence for the complexity of radix selection on markov sources*, Stochastic Processes and their Applications **129** (2019), no. 2, 507–538.

[13] Hosam M. Mahmoud, *Distributional analysis of swaps in quick select*, Theoretical Computer Science **411** (2010), no. 16, 1763–1769.

[14] Jason Matterer, *Quickselect process and quickval residual convergence*, Ph.D. thesis, The Johns Hopkins University, Baltimore, Maryland, 2015.

[15] Colin McDiarmid, *Concentration*, Probabilistic Methods for Algorithmic Discrete Mathematics (Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed, eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 195–248.
[16] Bruce Reed, *The height of a random binary search tree*, J. ACM **50** (2003), no. 3, 306–332.
[17] Henning Sulzbach, Ralph Neininger, and Michael Drmota, *A Gaussian limit process for optimal FIND algorithms*, Electronic Journal of Probability **19** (2014), no. none, 1 – 28.
[18] Brigitte Vallée, Julien Clément, James Allen Fill, and Philippe Flajolet, *The number of symbol comparisons in QuickSort and QuickSelect*, Automata, languages and programming. Part I, Lecture Notes in Comput. Sci., vol. 5555, Springer, Berlin, 2009, pp. 750–763. MR 2544890

Institute of Mathematics, Goethe University Frankfurt, Frankfurt a.M., Germany
Email address: ischebec@math.uni-frankfurt.de

Institute of Mathematics, Goethe University Frankfurt, Frankfurt a.M., Germany
Email address: neininger@math.uni-frankfurt.de