INDIVIDUAL DISPLACEMENTS FOR LINEAR PROBING HASHING WITH DIFFERENT INSERTION POLICIES

SVANTE JANSON

Abstract. We study the distribution of the individual displacements in hashing with linear probing for three different versions: First Come, Last Come and Robin Hood. Asymptotic distributions and their moments are found when the the size of the hash table tends to infinity with the proportion of occupied cells converging to some α, $0 < \alpha < 1$. (In the case of Last Come, the results are more complicated and less complete than in the other cases.)

We also show, using the diagonal Poisson transform studied by Poblete, Viola and Munro, that exact expressions for finite m and n can be obtained from the limits as $m, n \to \infty$.

We end with some results, conjectures and questions about the shape of the limit distributions. These have some relevance for computer applications.

1. Introduction

The standard version of hashing with linear probing can be described as follows, where n and m are integers with $0 \le n \le m$. (For a thorough discussion, see Knuth [15, Section 6.4, in particular Algorithm 6.4.L].)

> n items $x_1, \dots, x_n$ are placed sequentially into a table with m cells $1, \dots, m$, using n integers $h_i \in \{1, \dots, m\}$, by inserting x_i into cell h_i if it is empty, and otherwise trying cells $h_i + 1$, $h_i + 2$, ..., until an empty cell is found; all positions being interpreted modulo m.

For our probabilistic treatment, we assume that the hash addresses h_i are independent random numbers, uniformly distributed on $\{1, \dots, m\}$. In other words, each of the m^n possible hash sequences $(h_i)_1^n$ has the same probability m^{-n}.

If item x_i is inserted into cell q_i, then its *displacement* is $d_i := (q_i - h_i) \bmod m$. This is the number of unsuccessful probes when this item is inserted, as well as each time we later search for the item in the table. (The number of probes to find the item in the table is thus $d_i + 1$. This should be noted

Date: June 15, 2003; revised November 21, 2003; June 1, 2005.

1991 *Mathematics Subject Classification.* Primary: 68W40, 68P10; Secondary: 60C05, 60F05.

This work was first presented at the Knuthfest in honour of Donald Knuth's 64th birthday in January 2002.

when comparing the results below with other papers.) The displacement is a measure of the time (or cost) to find the item in the table; for simplicity we say that the search time is the displacement.

We began our study of hashing with linear probing in [10], where we studied the total displacement $\sum_i d_i$. In the present paper, we will study the individual displacements.

It turns out that the version of hashing described above leads to large variations among the displacements, especially for full or almost full tables. Several people have therefore suggested variations of the basic algorithm. We will study three versions of hashing with linear probing, differing in their insertion policies when there is a conflict.

- FC *First-Come(-First-Served)*. The usual version described above where the first item that probes a cell is inserted there and remains there.
- LC *Last-Come(-First-Served)*, see Poblete and Munro [22]. Each new item is inserted where it arrives. If the cell is already occupied, the old inhabitant is moved to the next cell. If that too is occupied, its old inhabitant is moved, etc.
- RH *Robin Hood*, see Celis, Larson and Munro [4] and [15, Answer 6.4-67]. When an item wants a cell that is already occupied by another item, the item (of the two) with the largest current displacement is put in the cell and the other is moved to the next cell, where the same rule applies recursively. (Ties are resolved in either way.) Robin Hood hashing minimizes the variance of the displacements for all linear probing algorithms [3], [23].

Note that the insertion of a sequence of items results in the same set of occupied cells in all three versions, and thus the same total displacement, while the individual displacements may differ. As has been shown before, and is seen by our results below, the Last Come and Robin Hood versions tend to give more evenly distributed displacements, thus reducing extreme values that may be annoying or dangerous.

Remark 1.1. It has been suggested [22, 23] that the displacements in the LC and RH versions may be so concentrated around their mean that searches would be quicker using centered probing, first probing cells at the mean displacement from the hash address. This seems to be true for double hashing and random probing, but we will see in Section 11 that for hashing with linear probing, this is not the case.

The situation we consider in this paper is a computer program where a large hash table is constructed once, and then used many times for finding the items. We mainly consider successful searches, although we give results for unsuccessful searches too, and we always assume that each item in the table is equally likely to be requested. We therefore have two levels of randomness: given a hash table $\mathcal{T}$, and its displacements (d_i), the time to find a random element in the table is a random variable $d(\mathcal{T}) = d_I$, where $I \in \{1, \dots, n\}$ is a uniformly distributed random index. As the program runs with many searches in the hash table, the search times are independent observations of this random variable. It is thus interesting to study the distribution of this random variable and its properties such as its mean and variance, and perhaps the probability of extremely large values. Note that this distribution depends on $\mathcal{T}$.

On the other hand, the hash table $\mathcal{T}$ is itself random; another run of the program yields another $\mathcal{T}$ and another set of displacements. Hence the distribution of the displacement $d(\mathcal{T})$ is a random distribution and its mean $\mathbb{E}(d(\mathcal{T})|\mathcal{T})$ and variance $\mathrm{Var}(d(\mathcal{T})|\mathcal{T}) = \mathbb{E}(d(\mathcal{T})^2|\mathcal{T}) - \mathbb{E}(d(\mathcal{T})|\mathcal{T})^2$ are random variables.

As has been noted earlier, see e.g. [10], the asymptotic behaviour of hashing with linear probing when n and m and tend to ∞ depend on the relative size of m and n. We will in the present paper, for simplicity as well as for lack of time and space, only consider the case $n/m \to \alpha$ with $0 < \alpha < 1$. (This is also the range of most interest for computer applications.) The case $n/m \to 0$ is more degenerate, with most displacements 0, and will be ignored. (It can be treated by similar methods, cf. the discussion of the total displacement in [10].) It will be seen below, that in this range, the dependency on $\mathcal{T}$ is negligible. For example, the mean and variance above, which are functions of $\mathcal{T}$, converge after suitable scaling in probability to some constants. In other words, we observe essentially the same distribution of search times for every run of the program.

The case $n/m \to 1$ is mathematically very interesting, and we plan to treat it in a later paper (jointly with Philippe Chassaing). There are two subcases, again cf. [10]: If $m - n \gg \sqrt{n}$, the results are similar to those in the present paper (but in some respects simpler); in particular, the random variation between different hash tables $\mathcal{T}$ is insignificant. On the other hand, in the almost full case when $(m-n)/\sqrt{n} \to c$ for some $c \geq 0$, the dependency on $\mathcal{T}$ is important and e.g. the mean $\mathbb{E}(d(\mathcal{T})|\mathcal{T})$ has a non-degenerate limit in distribution. Moreover, this is the Brownian phase, where the limits can be described using Brownian motion and derivatives of it such as the Brownian excursion.

The paper begins with some definitions and other preliminaries in Sections 2 and 3. A general limit theorem is given in Section 4 together with some variations.

In Section 5 we review these limit results in the context of the *diagonal Poisson transform* introduced by Poblete, Viola and Munro [23, 28]. This will show that the limit as $m, n \to \infty$ with $n/m \to \alpha$ of, for example, a certain moment of the displacements in random hash tables, equals a certain generating function (the *Poisson transform*) of the same moment. By inverting the Poisson transform, we are thus able to derive exact expressions for finite m and n from the limits as $m, n \to \infty$, a rather unusual situation!

The limit distributions for the different versions are found explicitly (more or less) in Sections 6–9. The reader will observe that our results for Last Come are much less satisfactory than for the other versions, and it is possible that others will succeed to find simpler forms of the result.

To illustrate the limit distributions, some numerical probabilities are given in Section 10.

Finally, the mode of the limit distributions are studied in Section 11, with a mixture of theorems, numerically based conjectures and open questions.

$n^{\underline{k}} := n(n-1)\cdots(n-k+1)$ denotes the falling factorial.

Acknowledgements. This paper is part of a joint project with Philippe Chassaing. I am also grateful to Alfredo Viola and Patricio Poblete for helpful discussions. Many related results, including some of the results below, have independently been found by Viola [29] by related but differently formulated methods. The reader is invited to compare (and combine) the two approaches.

2. Preliminaries

By a hash table $\mathcal{T}$ we mean not only the final table, but also its construction history; moreover, we consider the three possible results and construction histories under the three different policies together. Formally, a hash table can be regarded as encoded by the numbers m and n and the sequence $(h_1, \dots, h_n)$ of hash addresses. We always let $m = m(\mathcal{T})$ be the number of cells and $n = n(\mathcal{T})$ the number of items in the table. Thus there are $m-n$ empty cells. We always have $0 \le n \le m$. We sometimes exclude the cases $n = 0$ or $n = m$ to avoid trivial complications.

Our prime object of study is the random hash table $\mathcal{T}_{m,n}$ with m cells and n items ($0 \le n \le m$) and the hash addresses $h_1, \dots, h_n$ i.i.d. random variables, uniformly distributed on $\{1, \dots, m\}$. (This random hash table thus has a fixed size. In the analysis below we will also meet random hash tables where both the size and the number of items are random, see Section 3.)

We say that a hash table is *confined* if it leaves the last cell empty. We let $\mathcal{T}'_{m,n}$ ($n < m$) denote a random confined hash table, defined as $\mathcal{T}_{m,n}$ conditioned on the last cell being empty. By symmetry, the sequence of displacements has the same distribution for $\mathcal{T}'_{m,n}$ as for $\mathcal{T}_{m,n}$.

We denote the three insertion policies defined in the introduction by FC, LC and RH, and use Ξ to denote any of these.

Given a hash table $\mathcal{T}$, random or not, and a policy $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$, we let $d_i^{\Xi}(\mathcal{T})$ be the (final) displacement of the i:th item, $1 \le i \le n$, and

$$n_k^{\Xi}(\mathcal{T}) := \#\{i : d_i^{\Xi}(\mathcal{T}) = k\}, \qquad k = 0, 1, \dots,$$

the number of items with displacement k. Note that

$$\sum_k n_k^{\Xi}(\mathcal{T}) = n. \tag{2.1}$$

The total displacement is

$$d^*(\mathcal{T}) := \sum_{i=1}^{n} d_i^{\Xi}(\mathcal{T}) = \sum_{k=0}^{\infty} k n_k^{\Xi}(\mathcal{T})$$

and the average displacement is, when $n > 0$,

$$\bar{d}(\mathcal{T}) := \frac{1}{n} d^*(\mathcal{T}) = \frac{1}{n} \sum_{i=1}^{n} d_i^{\Xi}(\mathcal{T}).$$

As remarked above, $d^*(\mathcal{T})$ and $\bar{d}(\mathcal{T})$ do not depend on the policy Ξ.

If $n > 0$, we let $d^{\Xi}(\mathcal{T})$ denote a randomly chosen displacement in a given hash table $\mathcal{T}$ using policy Ξ, i.e. the random variable $d_I^{\Xi}(\mathcal{T})$ where $I \in \{1, \dots, n\}$ is a random index with a uniform distribution. Thus, given $\mathcal{T}$, $d^{\Xi}(\mathcal{T})$ has the distribution

$$\mathbb{P}\big(d^{\Xi}(\mathcal{T}) = k\big) = \frac{1}{n} n_k^{\Xi}(\mathcal{T}) \tag{2.2}$$

and the expectation

$$\mathbb{E}\, d^{\Xi}(\mathcal{T}) = \bar{d}(\mathcal{T}).$$

Similarly, we let $d_j^{\mathsf{U}}(\mathcal{T})$ denote the number of occupied cells encountered in an *unsuccessful* search starting at hash address j, $1 \le j \le m$, and let $d^{\mathsf{U}}(\mathcal{T})$ denote the number of occupied cells encountered in a random unsuccessful search, i.e. $d^{\mathsf{U}}(\mathcal{T}) := d_J^{\mathsf{U}}(\mathcal{T})$, where $J \in \{1, \dots, m\}$ is a uniformly distributed random index. (As for successful searches, the number of probes thus is one more, $d_j^{\mathsf{U}}(\mathcal{T}) + 1$ and $d^{\mathsf{U}}(\mathcal{T}) + 1$.) We further let

$$n_k^{\mathsf{U}}(\mathcal{T}) := \#\{j : d_j^{\mathsf{U}}(\mathcal{T}) = k\}, \qquad k = 0, 1, \dots,$$

and note that now, in contrast to (2.1),

$$\sum_k n_k^{\mathsf{U}}(\mathcal{T}) = m. \tag{2.3}$$

Thus, given $\mathcal{T}$, $d^{\mathsf{U}}(\mathcal{T})$ has the distribution

$$\mathbb{P}\big(d^{\mathsf{U}}(\mathcal{T}) = k\big) = \frac{1}{m} n_k^{\mathsf{U}}(\mathcal{T}).$$

2.1. **Blocks.** A *block* in a hash table (with $n < m$) is a sequence of cells $\{i+1, \dots, j\}$ (modulo m) where i and j are two consecutive empty cells; thus the last cell in a block is empty but all others are occupied. Clearly, $\mathcal{T}$ contains $m-n$ blocks which form a partition of $\{1, \dots, m\}$. We denote the block lengths in $\mathcal{T}$ by $\ell_1(\mathcal{T})$, ..., $\ell_{m-n}(\mathcal{T})$. (In the confined case, we take the blocks in the natural order, but in the unconfined case we start at a randomly chosen block. Otherwise, if for example in the unconfined case we would let the first block be the block containing cell 1, we would introduce an unwanted bias.)

Note that by our convention, each block includes the final, empty cell. Hence each block length is ≥ 1, and a block of length 1 is just a single empty cell (preceded by another empty cell). Further,

$$\sum_{i=1}^{m-n} \ell_i(\mathcal{T}) = m.$$

Each block may be regarded as an almost full confined hash table, with ℓ cells and $\ell - 1$ items, where ℓ is the length of the block.

The block lengths ℓ_i do not depend on the insertion policy.

If B is a block of length ℓ in $\mathcal{T}$, then the ℓ values of $d_j^{\mathsf{U}}(\mathcal{T})$ for $j \in B$ are $\ell - 1$, $\ell - 2$, ..., 0. Hence, for $k = 0, 1, \dots$,

$$n_k^{\mathsf{U}}(\mathcal{T}) = \#\{i : \ell_i(\mathcal{T}) > k\}. \tag{2.4}$$

We further see that $d^{\mathsf{U}}_j(\mathcal{T})$ and $d^{\mathsf{U}}(\mathcal{T})$ do not depend on the insertion policy.

2.2. **Profile.** Let, for $j = 1, \dots, m$, $X_j := \{i : h_i = j\}$ be the number of items with hash address j. Further, let H_j be the number of items that make an attempt to be inserted in cell j, whether they succeed or not. We extend these definitions to arbitrary integer indices by interpreting the index j modulo m. We call $(H_j)_{j=1}^m$ the *profile* of the hashing. If $H_j \geq 1$, then exactly one of the H_j items that try cell j ends up there, while the remaining $H_j - 1$ items either are rejected immediately or are admitted first but later thrown out; thus these continue to cell $j + 1$, and

$$H_{j+1} = X_{j+1} + (H_j - 1)_+.$$

When $n < m$, this set of equations has a unique solution [15, Exercise 6.4-32], [10, Lemma 2.1]

$$H_j = \max_{-\infty<i\leq j}\left(\sum_{k=i}^{j} X_k - (j-i)\right).$$

In particular, the profile does not depend on the insertion policy.

The profile has a simple relation to Robin Hood hashing.

Lemma 2.1. *For Robin Hood,*

$$n^{\mathsf{RH}}_k(\mathcal{T}) = \#\{i : d^{\mathsf{RH}}_i = k\} = \#\{j : H_j = k+1\}.$$

In other words, the displacements $\left(d^{\mathsf{RH}}_i\right)_{i=1}^n$ *are a permutation of the numbers* $(H_j - 1)_{j=1}^m$, *ignoring the* $m - n$ *negative values when* $H_i = 0$.

Proof. Consider a cell j and the X_j items that arrive at j. Under the RH rule, the final order of the items in a block is the same as the order of their places of arrivals. Hence, the $(H_{j-1} - 1)_+$ items that try cell $j - 1$ but are rejected (immediately or later) and therefore continue and try cell j, will end up in cells $j, \dots, j+(H_{j-1} - 1)_+ - 1$, and the X_j items that arrive at place j will end up in $j+(H_{j-1} - 1)_+, \dots, j+(H_{j-1} - 1)_+ + X_j - 1$, Consequently, the displacements of these X_j items are $(H_{j-1} - 1)_+, \dots, (H_{j-1} - 1)_+ + X_j - 1 = H_j - 1$. In particular, these displacements are all different, and, if $k \geq 0$, one of them equals k if and only if $H_{j-1} - 1 \leq k \leq H_j - 1$. Hence,

$$\begin{aligned} n^{\mathsf{RH}}_k &= \#\{j : H_{j-1} - 1 \leq k \leq H_j - 1\} \\ &= \#\{j : H_{j-1} \leq k+1 \leq H_j\}. \end{aligned} \tag{2.5}$$

Consider the sequence $\{H_j\}$, where the index runs through $\{1, \dots, m\}$ regarded as a cycle, with 1 following m. The number of times H_j increases across $[k, k+1]$ has to equal the number of times it decreases across the same interval, and since $H_j - H_{j-1} \geq -1$ always, such a decrease can only be from $k+1$ to k.

Hence, (2.5) yields

$$\begin{aligned} n_k^{\mathsf{RH}} &= \#\{j : H_{j-1} \le k+1 \le H_j\} \\ &= \#\{j : H_{j-1} = k+1 \le H_j\} + \#\{j : H_{j-1} \le k < k+1 \le H_j\} \\ &= \#\{j : H_{j-1} = k+1 \le H_j\} + \#\{j : H_{j-1} = k+1 > k = H_j\} \\ &= \#\{j : H_{j-1} = k+1\}. \end{aligned}$$

□

3. Random blocks and infinite hashing

Let $T(z)$ be the tree function

$$T(z) := \sum_{l=1}^{\infty} \frac{l^{l-1}z^l}{l!}, \qquad |z| \le e^{-1}. \tag{3.1}$$

Recall the well-known formulas $T(z)e^{-T(z)} = z$ ($|z| \le e^{-1}$), $T(\alpha e^{-\alpha}) = \alpha$ ($0 \le \alpha \le 1$),

$$\sum_{l=1}^{\infty} \frac{l^{l-2}z^l}{l!} = T(z) - \tfrac{1}{2}T(z)^2 \tag{3.2}$$

and

$$T'(z) = \frac{T(z)}{z\bigl(1-T(z)\bigr)}. \tag{3.3}$$

We say, for $0 \le \alpha \le 1$, that a random variable B_α has the *Borel distribution* $\mathrm{Bo}(\alpha)$ if

$$\mathbb{P}(B_\alpha = l) = \frac{l^{l-1}}{l!}\alpha^{l-1}e^{-l\alpha} = \frac{1}{T(\alpha e^{-\alpha})}\frac{l^{l-1}}{l!}(\alpha e^{-\alpha})^l, \qquad l = 1,2,\dots \tag{3.4}$$

(B_α always denotes such a variable.) The probability generating function of the Borel distribution is

$$\mathbb{E}\, z^{B_\alpha} = \sum_{l=1}^{\infty} \mathbb{P}(B_\alpha = l)z^l = \frac{T(\alpha e^{-\alpha}z)}{T(\alpha e^{-\alpha})} = \frac{T(\alpha e^{-\alpha}z)}{\alpha}. \tag{3.5}$$

Moments are easily computed from this. In particular, see e.g. [10, §4],

$$\mathbb{E}\, B_\alpha = \frac{1}{1-T(\alpha e^{-\alpha})} = \frac{1}{1-\alpha}, \tag{3.6}$$

$$\mathbb{E}\, B_\alpha^2 = \frac{1}{(1-\alpha)^3}. \tag{3.7}$$

As shown in [10, Lemma 4.1], for any $\alpha > 0$, the sequence of block lengths $\{\ell_i(\mathcal{T}_{m,n})\}_{i=1}^{m-n}$ of the random hash table $\mathcal{T}_{m,n}$ or $\mathcal{T}'_{m,n}$ has the same distribution as a sequence $\{X_i\}_{i=1}^{m-n}$ of independent random variables X_i with the common distribution $\mathrm{Bo}(\alpha)$, conditioned on $\sum_{i=1}^{m-n} X_i = m$. Moreover, conditioned on the block lengths ℓ_i, the internal structures of the blocks are independent, and the same as for a sequence of independent random almost full confined hash tables $\mathcal{T}'_{\ell_i,\ell_i-1}$, $i = 1,\dots,m-n$.

We let $\mathcal{T}_\alpha$, where $0 \le \alpha \le 1$, denote the random hash table of random size constructed by first selecting ℓ at random with the Borel distribution $\mathrm{Bo}(\alpha)$, and then taking a random confined hash table $\mathcal{T}'_{\ell,\ell-1}$. (Thus $\mathcal{T}_\alpha$ is a confined hash table.) The result just quoted then implies the following.

Lemma 3.1. *The random confined hash table $\mathcal{T}'_{m,n}$ can be obtained by juxtaposing $m-n$ independent copies of $\mathcal{T}_\alpha$, conditioning on the total size being m. The random hash table $\mathcal{T}_{m,n}$ can be obtained by first constructing $\mathcal{T}'_{m,n}$ in this way, followed by a random cyclic shift. Here α is any number with $0<\alpha\le 1$.* $\square$

Remark 3.2. There is a one-to-one correspondence between hash tables and rooted forests, see e.g. [15, Exercise 6.4-31] and [5], and the lemma is essentially equivalent to a result for random rooted forests by Pavlov [17, 19, 20]. Furthermore, Lemma 3.1 is closely related to results for generating functions for the total displacement in [7, 16].

Next, let us observe that the Borel distribution arises in connection with random walks. More precisely, let $\xi_1, \xi_2, \dots$ be i.i.d. random variables with the Poisson distribution $\mathrm{Po}(\alpha)$, and let $S_k := \sum_{i=1}^k (\xi_i - 1)$ be a random walk starting at $S_0 = 0$ with increments $\xi_i - 1$. Then $\tau := \min\{k : S_k = -1\}$ has the Borel distribution $\mathrm{Bo}(\alpha)$ [11, 12]; see also [6, 21].

Remark 3.3. It is easily seen that this result has equivalent reformulations in the theories of queues and branching processes: $\mathrm{Bo}(\alpha)$ is the distribution of the number of customers in a busy period in a queue with arrivals according to a Poisson process and constant service time; $\mathrm{Bo}(\alpha)$ is also the total progeny of a Galton–Watson process where each individual has $\mathrm{Po}(\alpha)$ children. For these results and generalizations, see e.g. [2, 18, 13, 27, 25, 26, 6, 21].

Furthermore, let us use this random walk to construct a hash table $\widetilde{\mathcal{T}}$ on $\{1, \dots, \tau\}$ by taking ξ_i items with hash address i, $1 \le i \le \tau$, inserting them in the table in random order. Then, for $1 \le k \le \tau$, the number of probes at cell k, i.e. H_k, is $S_k + 1$; thus the number of unsuccessful probes at k is S_k for $k < \tau$, and τ is the first empty cell. It is easily seen that, conditioned on $\tau = l$, the resulting table has a uniform distribution over all almost full confined hash tables of length l, i.e. it is $\mathcal{T}'_{l,l-1}$ Consequently, the random hash table $\widetilde{\mathcal{T}}$ equals (in distribution) $\mathcal{T}_\alpha$. This yields the following description of the profile.

Lemma 3.4. *Let $0<\alpha<1$. The profile $(H_j)_{j=1}^{m(\mathcal{T}_\alpha)}$ of $\mathcal{T}_\alpha$ has the same distribution as $(S_j+1)_{j=1}^{\tau}$, where $\xi \in \mathrm{Po}(\alpha)$, S_k and τ are as above.* $\square$

This may also be expressed using the following random *infinite* hashing.

Consider a hash table with infinitely many cells $1, 2, 3, \dots$ and suppose that items arrive to the cells according to independent Poisson processes with rate 1. When an item arrives, it is placed in the cell if the cell is empty, otherwise either the new or the old item (according to the chosen policy) is moved to the next cell, and so on. All movements are instantaneous. Denote the resulting table at time $t \ge 0$ by $\mathcal{T}_\infty(t)$. We define blocks in $\mathcal{T}_\infty(t)$ as for finite tables, starting at cell 1; we also consider an infinite string of occupied cells as a block. There are either an infinite number of finite blocks, or a finite (possibly zero) number of finite blocks followed by a single infinite block.

Lemma 3.5. *If $0 \le \alpha \le 1$, then the first block in $\mathcal{T}_\infty(\alpha)$ equals $\mathcal{T}_\alpha$ in distribution.*

Proof. Let ξ_i be the number of items arriving to cell i up to time α. Then $\xi_1, \xi_2, \dots$ are i.i.d. $\mathrm{Po}(\alpha)$ as above, and we define S_k and τ as before. Thus, for $k \le \tau$, the number of items probing cell k up to time α is $S_k + 1$, and τ is the first empty cell at time α, i.e. the length of the first block in $\mathcal{T}_\infty(\alpha)$. Conditioned on τ and $\xi_1, \dots, \xi_\tau$, the $\tau - 1$ items that have arrived to $\{1, \dots, \tau\}$ have come in random order, and we are in the situation of $\widetilde{\mathcal{T}}$ discussed above. $\square$

Remark 3.6. As a digression, let us study $\mathcal{T}_\infty(\alpha)$ further.

For $\alpha \le 1$, the random walk S_k defined above has negative or zero drift and thus a.s. hits -1, i.e. $\tau < \infty$ (as is implicit in the discussion above). After cell τ, the same process starts again, and thus $\mathcal{T}_\infty(\alpha)$ consists of an infinite string of independent random blocks, each a copy of $\mathcal{T}_\alpha$. In particular, the description above of the blocks in $\mathcal{T}_{m,n}$ is equivalent to the fact that if we condition $\mathcal{T}_\infty(\alpha)$ on the $(m-n)$:th empty cell being cell m, then the m first cells form a random confined hash table $\mathcal{T}'_{m,n}$, which easily is seen directly.

For $\alpha > 1$, the random walk $\{S_k\}$ has positive drift and thus has positive probability of never hitting -1. This means that $\mathcal{T}_\infty(\alpha)$ may have one or several finite blocks in the beginning, but, a.s., eventually there is an infinite block covering the rest of the table. As a consequence, the infinite hashing works for time $\alpha \le 1$, but for $\alpha > 1$ it fails because items are moved away to infinity (in zero time) and lost. In other words, there is a phase transition at time 1.

For $\alpha > 1$, the probability of the first block being finite is α'/α, where $\alpha' := T(\alpha e^{-\alpha}) < 1$, and conditioned on being finite, the block has the distribution of $T_{\alpha'}$. In particular, its length then has the Borel distribution $\mathrm{Bo}(\alpha')$ with mean $1/(1-\alpha')$.

Hence, the number of finite blocks has the geometric distribution $\mathrm{Ge}(1 - \alpha'/\alpha)$ with mean $\alpha'/(\alpha - \alpha')$. As a consequence, the expected number of cells in the finite blocks equals $\frac{\alpha'}{\alpha-\alpha'}\frac{1}{1-\alpha'}$.

Finally, we note that $\mathcal{T}_\infty(\alpha)$ is a discrete version of the queueing process studied by Borel [2].

Remark 3.7. We can similarly define two-way infinite hashing with the cells indexed by $\mathbb{Z}$, all integers. This can be regarded as infinite unconfined hashing, while the one-way infinite hashing on $\mathbb{Z}_+$ is infinite confined hashing.

In this case, we have a similar structure as for the one-way infinite case. If the time $\alpha \le 1$, the table consists a.s. of an infinite number of finite blocks. On the other hand, for $\alpha > 1$, the whole table is filled a.s.

4. A general limit theorem

We begin with a general limit theorem for the distribution of the individual displacements (in the case $n/m \to \alpha$, $0 < \alpha < 1$).

Recall that we first take a random hash table $\mathcal{T}_{m,n}$ and then consider the distribution of a random displacement for that hash table, i.e. we consider the conditional distribution of a random displacement given the hash table. We are thus really studying a random probability distribution; the reader who finds this too mind-boggling can instead think of the proportions $n_k^\Xi(\mathcal{T}_{m,n})/n$ of items with given displacements in the table, which is the same thing by (2.2).

Theorem 4.1. *Suppose that $m, n \to \infty$ with $n/m \to \alpha$, where $0 < \alpha < 1$, and let $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$.*
(i) *For every $k = 0, 1, \dots,$*

$$\mathbb{P}\big(d^\Xi(\mathcal{T}_{m,n}) = k \bigm| \mathcal{T}_{m,n}\big) = \frac{1}{n} n_k^\Xi(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{p}} p_\alpha^\Xi(k) := \frac{1-\alpha}{\alpha} \operatorname{\mathbb{E}} n_k^\Xi(\mathcal{T}_\alpha), \tag{4.1}$$

and $p_\alpha^\Xi = \{p_\alpha^\Xi(k)\}_{k=0}^\infty$ is a probability distribution on $\mathbb{N}$.
(ii) *More precisely, for every $k = 0, 1, \dots$ and jointly for all k,*

$$\sqrt{n}\big(\mathbb{P}(d^\Xi(\mathcal{T}_{m,n}) = k \mid \mathcal{T}_{m,n}) - p_{n/m}^\Xi(k)\big) = n^{-1/2}\big(n_k^\Xi(\mathcal{T}_{m,n}) - np_{n/m}^\Xi(k)\big) \xrightarrow{\mathrm{d}} Z_k^\Xi, \tag{4.2}$$

where Z_k^Ξ are some Gaussian random variables with means $\operatorname{\mathbb{E}} Z_k^\Xi = 0$ and a nondegenerate covariance matrix given by

$$\begin{split}\operatorname{Cov}(Z_k^\Xi, Z_l^\Xi) = \frac{1-\alpha}{\alpha}\Big(&\operatorname{Cov}\big(n_k^\Xi(\mathcal{T}_\alpha), n_l^\Xi(\mathcal{T}_\alpha)\big) \\ &- (1-\alpha)^3 \operatorname{Cov}\big(n_k^\Xi(\mathcal{T}_\alpha), m(\mathcal{T}_\alpha)\big) \operatorname{Cov}\big(n_l^\Xi(\mathcal{T}_\alpha), m(\mathcal{T}_\alpha)\big)\Big);\end{split} \tag{4.3}$$

furthermore, all moments converge too. In particular, with $N_k := n_k^\Xi(\mathcal{T}_{m,n})$,

$$\operatorname{\mathbb{E}} N_k = np_{n/m}^\Xi(k) + o(n^{1/2}) = np_\alpha^\Xi(k) + o(n),$$

$$(N_k - \operatorname{\mathbb{E}} N_k)/(\operatorname{Var} N_k)^{1/2} \xrightarrow{\mathrm{d}} N(0,1).$$

Remark 4.2. A more fancy formulation of part (i) of the theorem is that the distribution of $d^\Xi(\mathcal{T}_{m,n})$ converges to p_α^Ξ in probability, in the space of all probability measures on $\mathbb{N}$, equipped with the weak topology (which coincides with the ℓ^1 topology on this space); see [1] for definitions.

We will find (more or less explicit) formulas for the limit probabilities p_α^Ξ in Sections 7–9. It seems possible that the variances and covariances can be found by the same methods, but we have not attempted to find them.

The theorem says that a typical instance of the random hash table $\mathcal{T}_{m,n}$ has its displacements distributed approximately according to p_α^Ξ, with some normal fluctuations. In particular, different realizations of $\mathcal{T}_{m,n}$ have (with large probability) almost the same distribution. Taking the expectation over the possible choices of $\mathcal{T}_{m,n}$, this yields the following, conceptually simpler, corollary on the distribution of the displacement of a random item in a random hash table.

Corollary 4.3. *If $n/m \to \alpha < 1$ and $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$, then*

$$d^\Xi(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{d}} D_\alpha^\Xi,$$

where D^{Ξ}_{α} is a random variable with distribution

$$\mathbb{P}(D^{\Xi}_{\alpha} = k) = p^{\Xi}_{\alpha}(k). \qquad \square$$

Proof of Theorem 4.1. Let $\alpha_n := n/m$ and let $\mathcal{T}_1, \mathcal{T}_2, \dots$ be independent copies of $\mathcal{T}_{\alpha_n}$. By Lemma 3.1, $n^{\Xi}_k(\mathcal{T}_{m,n})$ has the same distribution as $\sum_{i=1}^{m-n} n^{\Xi}_k(\mathcal{T}_i)$ conditioned on $\sum_{i=1}^{m-n} m(\mathcal{T}_i) = m$.

We define $X_i := m(\mathcal{T}_i)$ and $Y_i := n^{\Xi}_k(\mathcal{T}_i)$ and are thus led to study $\sum_{i=1}^{m-n} Y_i$ conditioned on $\sum_{i=1}^{m-n} X_i = m$. We use, as in [10], the following conditional limit theorem proved in [9].

Lemma 4.4. *Suppose that, for each ν, $(X,Y) = (X(\nu), Y(\nu))$ is a pair of random variables such that X is integer valued, and that $N = N(\nu)$ and $m = m(\nu)$ are integers. Suppose further that for some γ and c (independent of ν), with $0 < \gamma \le 2$ and $c > 0$, the following hold, where $\sigma_X^2 := \operatorname{Var} X$, $\sigma_Y^2 := \operatorname{Var} Y$ and all limits are taken as $\nu \to \infty$:*

(i) $\mathbb{E}\, X = m/N$.
(ii) $0 < \sigma_X^2 < \infty$.
(iii) *For every integer $r \ge 3$,* $\mathbb{E}\,|X - \mathbb{E}\, X|^r = o(N^{r/2-1}\sigma_X^r)$.
(iv) $\sigma_X^2 = O(N^{2/\gamma - 1})$.
(v) $\varphi_X(s) := \mathbb{E}\, e^{isX}$ *satisfies* $1 - |\varphi_X(s)| \ge c \min(|s|^{\gamma}, s^2\sigma_X^2)$ *for* $|s| \le \pi$.
(vi) $0 < \sigma_Y^2 < \infty$.
(vii) *For every integer $r \ge 3$,* $\mathbb{E}\,|Y - \mathbb{E}\, Y|^r = o(N^{r/2-1}\sigma_Y^r)$.
(viii) *The correlation* $\rho := \operatorname{Cov}(X,Y)/\sigma_X\sigma_Y$ *satisfies* $\limsup |\rho| < 1$.

Let, for each ν, (X_i, Y_i) be i.i.d. copies of (X,Y), and let $S_N := \sum_1^N X_i$, $T_N := \sum_1^N Y_i$ and $\tau^2 := \sigma_Y^2(1-\rho^2) = \sigma_Y^2 - \operatorname{Cov}(X,Y)^2/\sigma_X^2$. Then, as $\nu \to \infty$, the conditional distribution of $(T_N - N\,\mathbb{E}\, Y)/N^{1/2}\tau$ given $S_N = m$ converges to a standard normal distribution. In other words, if $U = U_\nu$ is a random variable whose distribution equals the conditional distribution of T_N given $S_N = m$, then

$$\frac{U - N\,\mathbb{E}\, Y}{N^{1/2}\tau} \xrightarrow{\mathrm{d}} N(0,1). \tag{4.4}$$

Moreover, $\mathbb{E}\, U = N\,\mathbb{E}\, Y + o(N^{1/2}\tau)$ and $\operatorname{Var} U \sim N\tau^2$, and thus also

$$\frac{U - \mathbb{E}\, U}{(\operatorname{Var} U)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1). \tag{4.5}$$

The limits (4.4) and (4.5) hold with convergence of all moments. $\square$

To apply Lemma 4.4, we consider a sequence $(m,n) = (m(\nu), n(\nu))$ tending to infinity and let $N := m - n$. First, fix $k \ge 1$ and let $X := m(\mathcal{T}_{n/m})$ and $Y := n^{\Xi}_k(\mathcal{T}_{n/m})$. By the remarks before the lemma, $n^{\Xi}_k(\mathcal{T}_{m,n})$ has the same distribution as T_N conditioned on $S_N = m$, i.e. the same distribution as U in the lemma.

We have to verify all conditions of the lemma. First, $X \in \operatorname{Bo}(n/m)$ and thus, by (3.6), $\mathbb{E}\, X = 1/(1-\alpha_n) = m/N$, which verifies (i). Next, $n/m \to \alpha \in (0,1)$, and thus $\mathcal{T}_{n/m} \xrightarrow{\mathrm{d}} \mathcal{T}_\alpha$ and $(X,Y) \xrightarrow{\mathrm{d}} (\bar{X}, \bar{Y}) := \bigl(m(\mathcal{T}_\alpha), n^{\Xi}_k(\mathcal{T}_\alpha)\bigr)$, with convergence of all moments. Further, $N \to \infty$, the distribution of $\bar{X}$ has span

1 (because $\mathbb{P}(\bar X = 1) > 0$) and $\bar Y$ is not a.s. equal to a linear function of $\bar X$ because both $\mathbb{P}(\bar X = k+2, \bar Y = 0) > 0$ and $\mathbb{P}(\bar X = k+2, \bar Y = 1) > 0$. It follows easily, see [9, Corollary 2.1] for a general statement, that all other conditions of the lemma hold for $\gamma = 2$, with $\tau^2 = \operatorname{Var}(\bar Y) - \operatorname{Cov}(\bar Y, \bar X)^2 / \operatorname{Var}(\bar X)$.

We may thus apply Lemma 4.4. We have

$$N \,\mathbb{E}\, Y = n \frac{1-n/m}{n/m} \,\mathbb{E}\, n_k^{\Xi}(\mathcal{T}_{n/m}) = n p^{\Xi}_{n/m}(k),$$

and (4.2) follows for a single k, with $\operatorname{Var}(Z_k^{\Xi}) = \lim (N/n)\tau^2 = \bigl((1-\alpha)/\alpha\bigr)\tau^2$. (Note that we divide by $n^{1/2}$ in (4.2) but $N^{1/2}$ in (4.4).) Furthermore, the same argument applies with the Y above replaced by a linear combination $Y := \sum_0^K a_k n_k^{\Xi}(\mathcal{T}_{n/m})$, which by the Cramér–Wold device yields joint convergence and, using (3.7), (4.3), see [9, Corollary 2.2] for details.

Again, no such Y is equal to a linear function of X, and thus every nontrivial finite linear combination $\sum_0^K a_k Z_k^{\Xi}$ has nonzero variance.

Part (ii) of Theorem 4.1 thus follows from Lemma 4.4. Part (i) follows from part (ii) and (2.1) and (3.6), which yields

$$\sum_{k=0}^{\infty} p_\alpha^{\Xi}(k) = \frac{1-\alpha}{\alpha} \,\mathbb{E} \sum_{k=0}^{\infty} n_k^{\Xi}(\mathcal{T}_\alpha) = \frac{1-\alpha}{\alpha} \,\mathbb{E}\, n(\mathcal{T}_\alpha) = \frac{1-\alpha}{\alpha} \,\mathbb{E}\bigl(m(\mathcal{T}_\alpha) - 1\bigr) = 1.$$

□

We have similar results, with minor differences, for the unsuccessful searches.

Theorem 4.5. *Suppose that $m, n \to \infty$ with $n/m \to \alpha$, where $0 < \alpha < 1$.*
(i) *For every $k = 0, 1, \dots,$*

$$\mathbb{P}\bigl(d^{\mathsf{U}}(\mathcal{T}_{m,n}) = k \bigm| \mathcal{T}_{m,n}\bigr) = \frac{1}{m} n_k^{\mathsf{U}}(\mathcal{T}_{m,n}) \xrightarrow{\text{p}} p_\alpha^{\mathsf{U}}(k) := (1-\alpha)\,\mathbb{P}(B_\alpha > k),$$

where $B_\alpha \in \operatorname{Bo}(\alpha)$, and $p_\alpha^{\mathsf{U}} = \{p_\alpha^{\mathsf{U}}(k)\}_{k=0}^{\infty}$ is a probability distribution on $\mathbb{N}$.
(ii) *More precisely, for every $k = 0, 1, \dots$ and jointly for all k,*

$$\sqrt{n}\bigl(\mathbb{P}(d^{\mathsf{U}}(\mathcal{T}_{m,n}) = k \mid \mathcal{T}_{m,n}) - p^{\mathsf{U}}_{n/m}(k)\bigr) = \frac{n^{1/2}}{m}\bigl(n_k^{\mathsf{U}}(\mathcal{T}_{m,n}) - m p^{\mathsf{U}}_{n/m}(k)\bigr) \xrightarrow{\text{d}} Z_k^{\mathsf{U}},$$

where Z_k^{U} are some Gaussian random variables with means $\mathbb{E}\, Z_k^{\mathsf{U}} = 0$ and a covariance matrix given by

$$\begin{aligned}\operatorname{Cov}(Z_k^{\mathsf{U}}, Z_l^{\mathsf{U}}) = \alpha(1-\alpha)\Bigl(&\operatorname{Cov}\bigl(\mathbf{1}[B_\alpha \le k], \mathbf{1}[B_\alpha \le l]\bigr) \\ &- (1-\alpha)^3 \operatorname{Cov}\bigl(\mathbf{1}[B_\alpha \le k], B_\alpha\bigr) \operatorname{Cov}\bigl(\mathbf{1}[B_\alpha \le l], B_\alpha\bigr)\Bigr);\end{aligned}$$

furthermore, all moments converge too. In particular, with $N_k := n_k^{\mathsf{U}}(\mathcal{T}_{m,n})$,

$$\mathbb{E}\, N_k = m p^{\mathsf{U}}_{n/m}(k) + o(m^{1/2}) = m p^{\mathsf{U}}_\alpha(k) + o(m),$$

$$(N_k - \mathbb{E}\, N_k)/(\operatorname{Var} N_k)^{1/2} \xrightarrow{\text{d}} N(0,1).$$

The case $k = 0$ is trivial, $n_0^{\mathsf{U}}(\mathcal{T}_{m,n}) = m - n = m p^{\mathsf{U}}_{n/m}(0)$ and thus $Z_0^{\mathsf{U}} = 0$, but the variables Z_k^{U}, $k \ge 1$, are linearly independent.

Proof. We argue as in the proof of Theorem 4.1, now taking $Y := n_k^{\mathsf{U}}(\mathcal{T}_{n/m})$, or a linear combination $\sum_0^K a_k n_k^{\mathsf{U}}(\mathcal{T}_{n/m})$ of such variables. By (2.4), recalling that $\mathcal{T}_\alpha$ has a single block, $n_k^{\mathsf{U}}(\mathcal{T}_{n/m}) = \mathbf{1}[m(\mathcal{T}_{n/m}) > k]$ and thus

$$N\,\mathbb{E}\, n_k^{\mathsf{U}}(\mathcal{T}_{n/m}) = m(1-n/m)\,\mathbb{P}\big(m(\mathcal{T}_{n/m}) > k\big) = m p_{n/m}^{\mathsf{U}}(k),$$

and the result follows from Lemma 4.4 as above, if we recall that $m(\mathcal{T}_\alpha) \in \mathrm{Bo}(\alpha)$ so $m(\mathcal{T}_\alpha) \overset{\mathrm{d}}{=} B_\alpha$, and use $\mathbf{1}[B_\alpha > k] = 1 - \mathbf{1}[B_\alpha \le k]$. ☐

Corollary 4.6. *If $n/m \to \alpha < 1$, then*

$$d^{\mathsf{U}}(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{d}} D_\alpha^{\mathsf{U}},$$

where D_α^{U} is a random variable with distribution

$$\mathbb{P}(D_\alpha^{\mathsf{U}} = k) = p_\alpha^{\mathsf{U}}(k).$$ ☐

In the proofs above we used finite linear combinations $\sum_0^K a_k n_k^{\Xi}(\mathcal{T})$, but we can just as well take infinite sums $\sum_0^\infty f(k) n_k^{\Xi}(\mathcal{T})$, provided the function f grows subexponentially, i.e. $f(k) = \exp\big(o(k)\big)$, which implies that $\sum_0^\infty f(k) n_k^{\Xi}(\mathcal{T}_\alpha)$ has finite moments of all orders. In fact, this was done in [10] to study the total displacement $d^*(\mathcal{T}_{m,n})$, which is obtained by the choice $f(k) = k$.

We obtain the following result, leaving the formulas for the asymptotic variances to the reader.

Theorem 4.7. *Let f be a function of subexponential growth on $\mathbb{N}$, for example a polynomial. Suppose that $m, n \to \infty$ with $n/m \to \alpha$, where $0 < \alpha < 1$. For every $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(d_i^{\Xi}(\mathcal{T}_{m,n})) = \frac{1}{n}\sum_{k=0}^{\infty} f(k) n_k^{\Xi}(\mathcal{T}_{m,n})$$
$$\xrightarrow{\mathrm{p}} \sum_{k=0}^{\infty} f(k) p_\alpha^{\Xi}(k) = \frac{1-\alpha}{\alpha}\,\mathbb{E}\sum_{k=0}^{\infty} f(k) n_k^{\Xi}(\mathcal{T}_\alpha) = \frac{1-\alpha}{\alpha}\,\mathbb{E}\sum_{i=1}^{n(\mathcal{T}_\alpha)} f(d_i^{\Xi}(\mathcal{T}_\alpha)).$$

and, similarly, with $B_\alpha \in \mathrm{Bo}(\alpha)$,

$$\frac{1}{m}\sum_{j=1}^{m} f(d_j^{\mathsf{U}}(\mathcal{T}_{m,n})) = \frac{1}{m}\sum_{k=0}^{\infty} f(k) n_k^{\mathsf{U}}(\mathcal{T}_{m,n})$$
$$\xrightarrow{\mathrm{p}} \sum_{k=0}^{\infty} f(k) p_\alpha^{\mathsf{U}}(k) = (1-\alpha)\,\mathbb{E}\sum_{k=0}^{B_\alpha - 1} f(k).$$

The convergences hold in L^p too, for any $p < \infty$; in particular, the expectations of the left hand sides converge to the right hand sides. Moreover, the random variables on the left hand side are asymptotically normal, with variances $O(n^{-1})$. ☐

Corollary 4.8. *Suppose that $m, n \to \infty$ with $n/m \to \alpha$, where $0 < \alpha < 1$, and let $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}, \mathsf{U}\}$. Then*

$$\mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n}) \mid \mathcal{T}_{m,n}) \xrightarrow{\mathrm{p}} \mathbb{E}\, D_{\alpha}^{\Xi}, \tag{4.6}$$

$$\operatorname{Var}(d^{\Xi}(\mathcal{T}_{m,n}) \mid \mathcal{T}_{m,n}) \xrightarrow{\mathrm{p}} \operatorname{Var} D_{\alpha}^{\Xi}. \tag{4.7}$$

The expectations of the left hand sides converge to the same limits.

Proof. Apply Theorem 4.7 with $f(k) = k$ and $f(k) = k^2$. □

Note that if $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$, then $\mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n}) \mid \mathcal{T}_{m,n}) = \bar{d}(\mathcal{T}_{m,n})$, the average displacement. We thus have the orthogonal decomposition $d^{\Xi}(\mathcal{T}_{m,n}) = \big(d^{\Xi}(\mathcal{T}_{m,n}) - \bar{d}(\mathcal{T}_{m,n})\big) + \bar{d}(\mathcal{T}_{m,n})$ and the decomposition of the variance

$$\begin{aligned}
\operatorname{Var}\big(d^{\Xi}(\mathcal{T}_{m,n})\big) &= \operatorname{Var}\big(d^{\Xi}(\mathcal{T}_{m,n}) - \bar{d}(\mathcal{T}_{m,n})\big) + \operatorname{Var}\big(\bar{d}(\mathcal{T}_{m,n})\big) \\
&= \mathbb{E}\big(\operatorname{Var}\big(d^{\Xi}(\mathcal{T}_{m,n}) \mid \mathcal{T}_{m,n}\big)\big) + \operatorname{Var}\big(\bar{d}(\mathcal{T}_{m,n})\big) \\
&\to \operatorname{Var}(D_{\alpha}^{\Xi}).
\end{aligned}$$

The second term, i.e. the part of the variance that comes from the variation between different hash tables is of order $O(n^{-1})$ only by Theorem 4.7, and thus much smaller than the first term which is the part of the variance coming from the variation between different items in the table. This, again, shows that the variation between tables is insignificant in the sparse range.

A similar result holds for d^{U}. Furthermore, we obtain similar results for higher moments by taking $f(k) = k^r$, $r > 0$. In particular, we can sharpen Corollaries 4.3 and 4.6.

Corollary 4.9. *If $n/m \to \alpha < 1$ and $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}, \mathsf{U}\}$, then $d^{\Xi}(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{d}} D_{\alpha}^{\Xi}$ with all moments, i.e.*

$$\mathbb{E}\big(d^{\Xi}(\mathcal{T}_{m,n})\big)^r \to \mathbb{E}\big(D_{\alpha}^{\Xi}\big)^r, \qquad r \geq 0. \qquad \square$$

We have so far treated all three insertion policies together. In Sections 6–9, we will study them one by one (beginning with unsucessful searches) and identify the limit distributions p_{α}^{Ξ}, i.e. the distributions of the limit random variables D_{α}^{Ξ}.

5. Exact distributions

Although we are mainly interested in asymptotic result, we make in this section a digression and consider exact formulas for the distributions of $d^{\Xi}(\mathcal{T}_{m,n})$ and, in particular, their moments. For simplicity, we consider only $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$ in this section, and leave the case of unsuccessful searches to the next section. Using results by Poblete, Viola and Munro [23, 28], we will see that the moments of D_{α}^{Ξ} not only are the limits of the moments of $d^{\Xi}(\mathcal{T}_{m,n})$ as $m, n \to \infty$ with $n/m \to \alpha$, they can also be regarded as *Poisson transforms* of the moments of $d^{\Xi}(\mathcal{T}_{m,n})$. The same is true for the probabilities $\mathbb{P}(d^{\Xi}(\mathcal{T}_{m,n}) = k)$ and for the probability generating function. This provides an interesting relation between the values of these moments (probabilities) and their limits. Moreover, it is possible to invert the Poisson transform and thus derive exact formulas from the limits. This yields interesting connections with earlier results by various authors.

We denote the probability generating function of D^{Ξ}_{α} by ψ^{Ξ}_{α} and have by (4.1), for $0 < \alpha < 1$, at least for $|z| \le 1$,

$$\psi^{\Xi}_{\alpha}(z) := \mathbb{E}\, z^{D^{\Xi}_{\alpha}} = \sum_{k=0}^{\infty} z^k p^{\Xi}_{\alpha}(k) = \frac{1-\alpha}{\alpha}\sum_{k=0}^{\infty} z^k\, \mathbb{E}\, n^{\Xi}_k(\mathcal{T}_{\alpha}). \tag{5.1}$$

Recall that the probabilities $p^{\Xi}_{\alpha}(k)$ can be obtained by differentiating $\psi^{\Xi}_{\alpha}(z)$ at $z = 0$, and that the (factorial) moments are obtained by differentiation at $z = 1$.

Remark 5.1. Since no displacement is larger than the size of the table, we have the bound $\mathbb{E}\, n^{\Xi}_k(\mathcal{T}_{\alpha}) \le \mathbb{E}\big(m(\mathcal{T}_{\alpha})\mathbf{1}[m(\mathcal{T}_{\alpha}) \ge k]\big)$, which together with (3.4) easily implies that the sums in (5.1) converge at least for $|z| < \big(\alpha e^{1-\alpha}\big)^{-1}$. Hence, for each $\alpha < 1$, $\psi^{\Xi}_{\alpha}(z)$ is analytic in a disc with radius greater than 1, and (5.1) is valid there. The same applies to various formulas for generating functions below; they are always valid for $|z| \le 1$, and actually in some larger open domain (possibly depending on parameters such as α and Ξ), but we will usually ignore mentioning this restriction on z.

Furthermore, for $1 \le n \le m$, let

$$\varphi^{\Xi}_{m,n}(z) := \mathbb{E}\, z^{d^{\Xi}(\mathcal{T}_{m,n})}, \tag{5.2}$$

the probability generating function of $d^{\Xi}(\mathcal{T}_{m,n})$. When $n < m$, $d^{\Xi}(\mathcal{T}'_{m,n})$ has the same distribution as $d^{\Xi}(\mathcal{T}_{m,n})$, see Section 2, and thus the same probability generating function $\varphi^{\Xi}_{m,n}(z)$.

Let, for $\ell \ge 1$, $\Phi^{\Xi}_{\ell}(z)$ be the sum of $z^{d^{\Xi}_i(\mathcal{T})}$ over all $\ell^{\ell-2}$ confined almost full hash tables of length ℓ, and all $i \in \{1, \dots, l-1\}$. Thus $\Phi^{\Xi}_1(z) = 0$, and for $\ell \ge 2$, $\Phi^{\Xi}_{\ell}(z) = (\ell-1)\ell^{\ell-2}\varphi^{\Xi}_{\ell,\ell-1}(z)$. Further,

$$\Phi^{\Xi}_{\ell}(z) = \ell^{\ell-2}\, \mathbb{E}\sum_{i=1}^{\ell-1} z^{d^{\Xi}_i(\mathcal{T}'_{\ell,\ell-1})} = \ell^{\ell-2}\, \mathbb{E}\sum_{k=0}^{\infty} n^{\Xi}_k(\mathcal{T}'_{\ell,\ell-1}) z^k. \tag{5.3}$$

Hence, by (5.1), the definition of $\mathcal{T}_{\alpha}$ and (3.4),

$$\begin{aligned}
\psi^{\Xi}_{\alpha}(z) &= \frac{1-\alpha}{\alpha}\sum_{k=0}^{\infty} z^k\, \mathbb{E}\, n^{\Xi}_k(\mathcal{T}_{\alpha}) = \frac{1-\alpha}{\alpha}\sum_{\ell=1}^{\infty} \mathbb{P}(B_{\alpha} = \ell)\ell^{-(\ell-2)}\Phi^{\Xi}_{\ell}(z) \\
&= (1-\alpha)\sum_{\ell=2}^{\infty} \frac{\alpha^{\ell-2}e^{-\ell\alpha}}{(\ell-1)!}\Phi^{\Xi}_{\ell}(z) && \text{(5.4a)} \\
&= (1-\alpha)\sum_{\ell=2}^{\infty} \frac{(\ell\alpha)^{\ell-2}e^{-\ell\alpha}}{(\ell-2)!}\varphi^{\Xi}_{\ell,\ell-1}(z) \\
&= (1-\alpha)\sum_{i=0}^{\infty} \frac{((i+2)\alpha)^{i}e^{-(i+2)\alpha}}{i!}\varphi^{\Xi}_{i+2,i+1}(z). && \text{(5.4b)}
\end{aligned}$$

In terms of the transforms defined in Poblete, Viola and Munro [23], (5.4b) shows that $\psi_\alpha^\Xi(z)$ is the *diagonal Poisson transform* $\mathcal{D}_2[\varphi^\Xi_{n+2,n+1};\alpha]$.

Before proceeding, we note that (5.4b) also shows that $\psi_\alpha^\Xi(z)$ is an analytic function of α too (in a suitable domain).

Fix z and let $f_{m,n-1} := \varphi^\Xi_{m,n}(z)$, $0 \le n < m$. (Thus $f_{m,n}$ refers to a hash table of size m with $n+1$ items; we follow here the notation in [23].) This is the expectation of a random variable that depends on a randomly chosen item and the block it belongs to. Poblete, Viola and Munro [23, §4.1] show that for such quantities $f_{m,n}$, it is easy to express $f_{m,n}$ ($n < m-1$) in terms of the values for almost full tables $f_{i+2,i}$, see [23, (28)].

Using this formula [23, (28)] as a definition of $f_{m,n}$ for $n \ge m-1$ too, it is further shown in [23, (29)] that the diagonal Poisson transform $\mathcal{D}_2[f_{n+2,n};x]$ equals the *Poisson transform* $\mathcal{P}_m[f_{m,n};x]$, which is defined as $\mathbb{E}\, f_{m,N}$ where $N \in \mathrm{Po}(mx)$. (In particular, the latter transform is independent of m.) This leads to the following result.

Theorem 5.2. *For every $m \ge 1$ and $\Xi \in \{\mathsf{FC},\mathsf{LC},\mathsf{RH}\}$, $\varphi^\Xi_{m,n}$ can be defined for $n > m$ too such that for $0 \le \alpha < 1$*

$$\psi_\alpha^\Xi(z) = \mathcal{P}_m[\varphi^\Xi_{m,n+1}(z);\alpha] = e^{-m\alpha}\sum_{n=0}^\infty \frac{(m\alpha)^n}{n!}\varphi^\Xi_{m,n+1}(z) \tag{5.5}$$

and

$$\varphi^\Xi_{m,n}(z) = \frac{(n-1)!}{m^{n-1}}[\alpha^{n-1}]\bigl(e^{m\alpha}\psi_\alpha^\Xi(z)\bigr) = \sum_{k=0}^{n-1}\frac{(n-1)^{\underline{k}}}{m^k}\,[\alpha^k]\psi_\alpha^\Xi(z). \tag{5.6}$$

Note that the functions $\varphi^\Xi_{m,n}$ that appear here only have a formal meaning when $n > m$. At least, we do not know any probabilistic interpretation of them in this case, and they are not always probability generating functions. (For example, it follows easily from Theorem 5.2 and Theorem 8.1 below that $\varphi^{\mathsf{RH}}_{1,3}(z) = \frac13 z^2 + \frac43 z - \frac23$.) It would be interesting to find such an interpretation, and a probabilistic proof of Theorem 5.2.

Proof. We have already shown (5.5), with $\varphi^\Xi_{m,n}(z) = f_{m,n-1}$ given by [23, (28)] for $n \ge m$.

To invert the Poisson transform, we multiply (5.5) by $e^{m\alpha}$ and extract the coefficient of α^n, which yields

$$\begin{aligned}\varphi^\Xi_{m,n+1}(z) &= \frac{n!}{m^n}\,[\alpha^n]\bigl(e^{m\alpha}\psi_\alpha^\Xi(z)\bigr) = \frac{n!}{m^n}\sum_{k=0}^n\frac{m^{n-k}}{(n-k)!}\,[\alpha^k]\psi_\alpha^\Xi(z)\\ &= \sum_{k=0}^n \frac{n^{\underline{k}}}{m^k}\,[\alpha^k]\psi_\alpha^\Xi(z).\end{aligned}$$

One subtle problem remains: The argument in [23] leading to [23, (28)] uses confined hash tables, and is thus restricted to $f_{m,n}$ for $n+1 < m$. Hence, the $\varphi^\Xi_{m,n}(z)$ that appears in (5.5) and (5.6) is indeed given by (5.2) for $n < m$, but it remains to show that this is true for $n = m$ too. In order to see this, fix z

and n, and note that the right hand side of (5.6) is a polynomial in $1/m$. We claim that the same is true for $\varphi^{\Xi}_{m,n}$ for all $m \geq n$; since the equality is verified for $m > n$, it then holds for $m = n$ too, and the theorem is completely proved.

To verify the claim, define a graph on the set $\{1, \dots, n\}$ where i and j are joined by an edge if items i and j conflict over the same cell sometime during the construction of the hash table (using the policy Ξ), and call the components of this graph *strict blocks*. It is easily seen that any partition of $\{1, \dots, n\}$ into strict blocks together with some given internal structure in the strict blocks has a probability that is a polynomial in $1/m$ for $m \geq n$, and the result follows. $\square$

We can take derivatives at $z = 1$ in Theorem 5.2 to obtain the corresponding result for fractional moments, and then take suitable linear combinations to obtain the moments. Similarly, taking derivatives at $z = 0$ we obtain results for the point probabilities. Alternatively, we can argue as above using moments or probabilities for $f_{m,n}$. We thus obtain the following result.

Corollary 5.3. *Let $m \geq 1$ and $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$. Then, for every $r \geq 0$,*

$$\mathbb{E}\big(D^{\Xi}_{\alpha}\big)^r = \mathcal{P}_m[\mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n+1}))^r; \alpha] = e^{-m\alpha} \sum_{n=0}^{\infty} \frac{(m\alpha)^n}{n!} \mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n+1}))^r$$

and, for $k \geq 0$,

$$\mathbb{P}\big(D^{\Xi}_{\alpha} = k\big) = \mathcal{P}_m[\mathbb{P}(d^{\Xi}(\mathcal{T}_{m,n+1}) = k); \alpha] = e^{-m\alpha} \sum_{n=0}^{\infty} \frac{(m\alpha)^n}{n!} \mathbb{P}(d^{\Xi}(\mathcal{T}_{m,n+1}) = k)$$

where $\mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n+1}))^r$ and $\mathbb{P}(d^{\Xi}(\mathcal{T}_{m,n+1}) = k)$ only have a formal meaning for $n + 1 > m$. Conversely, for $1 \leq n \leq m$,

$$\mathbb{E}\big(d^{\Xi}(\mathcal{T}_{m,n})\big)^r = \frac{(n-1)!}{m^{n-1}}[\alpha^{n-1}]\big(e^{m\alpha}\,\mathbb{E}\big(D^{\Xi}_{\alpha}\big)^r\big) = \sum_{i=0}^{n-1} \frac{(n-1)^{\underline{i}}}{m^i}\,[\alpha^i]\,\mathbb{E}\big(D^{\Xi}_{\alpha}\big)^r,$$

$$\begin{aligned}\mathbb{P}\big(d^{\Xi}(\mathcal{T}_{m,n}) = k\big) &= \frac{(n-1)!}{m^{n-1}}[\alpha^{n-1}]\big(e^{m\alpha}\,\mathbb{P}(D^{\Xi}_{\alpha} = k)\big) \\ &= \sum_{i=0}^{n-1} \frac{(n-1)^{\underline{i}}}{m^i}\,[\alpha^i]\,\mathbb{P}\big(D^{\Xi}_{\alpha} = k\big).\end{aligned}$$ $\square$

An important example is provided by the transform $(1-\alpha)^{-r-1}$, which has the inverse Poisson transform

$$\frac{n!}{m^n}\,[\alpha^n]\Big(e^{m\alpha}(1-\alpha)^{-r-1}\Big) = \frac{n!}{m^n} \sum_{k=0}^{n} \frac{m^{n-k}}{(n-k)!}\binom{r+k}{k} = Q_r(m,n),$$

the Q function defined in [15, Theorem 6.4.K]. Here r may be any real number. Note, in particular, that $Q_{-1}(m,n) = 1$ and $Q_{-2}(m,n) = 1 - n/m$.

Consequently, we have an explicit formula, covering several cases below. (We use the simple, well-known, identity [16, (5.8)] for $Q_r(m, n-1)$.)

Corollary 5.4. *Let $r \geq 1$ and $\Xi \in \{\mathsf{FC}, \mathsf{LC}, \mathsf{RH}\}$, and suppose that*

$$\mathbb{E}\big(D_\alpha^\Xi\big)^r = \sum_{j\in J} c_j (1-\alpha)^{-j}, \qquad 0<\alpha<1,$$

for some finite set $J \subset \mathbb{Z}$ and numbers c_j. Then, for $1 \leq n \leq m$,

$$\begin{aligned}\mathbb{E}\big(d^\Xi(\mathcal{T}_{m,n})\big)^r &= \sum_{j\in J} c_j Q_{j-1}(m,n-1)\\ &= \frac{m}{n}\sum_{j\in J} c_j\big(Q_{j-1}(m,n)-Q_{j-2}(m,n)\big).\end{aligned}$$ □

Of course, the converse of this corollary is immediate.

6. The limit distribution for unsuccessful searches

By Theorem 4.5 and (3.4),

$$p_\alpha^{\mathsf{U}}(k) = (1-\alpha)\,\mathbb{P}(B_\alpha > k) = (1-\alpha)\Big(1-\sum_{l=1}^{k}\frac{l^{l-1}}{l!}\alpha^{l-1}e^{-l\alpha}\Big). \tag{6.1}$$

In particular, $p_\alpha^{\mathsf{U}}(0) = 1-\alpha$, $p_\alpha^{\mathsf{U}}(1) = (1-\alpha)(1-e^{-\alpha})$ and $p_\alpha^{\mathsf{U}}(2) = (1-\alpha)(1-e^{-\alpha}-\alpha e^{-2\alpha})$.

We have the following further results.

Theorem 6.1. *The probability generating function of D_α^{U} is*

$$\psi_\alpha^{\mathsf{U}}(z) := \mathbb{E}\, z^{D_\alpha^{\mathsf{U}}} = \sum_{k=0}^\infty z^k p_\alpha^{\mathsf{U}}(k) = \frac{1-\alpha}{\alpha}\cdot\frac{T(\alpha e^{-\alpha})-T(\alpha e^{-\alpha}z)}{1-z}.$$

The first moments are given by

$$\begin{aligned}\mathbb{E}\, D_\alpha^{\mathsf{U}} &= \frac{2\alpha-\alpha^2}{2(1-\alpha)^2} = \tfrac12(1-\alpha)^{-2}-\tfrac12,\\ \mathbb{E}(D_\alpha^{\mathsf{U}})^2 &= \frac{6\alpha+3\alpha^2-4\alpha^3+\alpha^4}{6(1-\alpha)^4} = (1-\alpha)^{-4}-\tfrac23(1-\alpha)^{-3}-\tfrac12(1-\alpha)^{-2}+\tfrac16,\\ \operatorname{Var}(D_\alpha^{\mathsf{U}}) &= \frac{12\alpha-6\alpha^2+4\alpha^3-\alpha^4}{12(1-\alpha)^4} = \tfrac34(1-\alpha)^{-4}-\tfrac23(1-\alpha)^{-3}-\tfrac1{12}.\end{aligned}$$

Proof. By Theorem 4.5 and (3.4), or Theorem 4.7 with $f(k)=z^k$ (for $|z|\leq 1$), together with (3.5),

$$\begin{aligned}\psi_\alpha^{\mathsf{U}}(z) &= \sum_{k=0}^\infty z^k p_\alpha^{\mathsf{U}}(k) = \sum_{k=0}^\infty z^k(1-\alpha)\,\mathbb{P}(B_\alpha>k) = (1-\alpha)\,\mathbb{E}\sum_{k=0}^{B_\alpha-1} z^k\\ &= (1-\alpha)\,\mathbb{E}\,\frac{1-z^{B_\alpha}}{1-z} = (1-\alpha)\frac{1-T(\alpha e^{-\alpha}z)/\alpha}{1-z}.\end{aligned}$$

The moments are computed by calculating the derivatives of $\psi_\alpha^{\mathsf{U}}(z)$ at $z=1$. (A computer algebra program is helpful.) □

To obtain exact formulas for moments of $d^{\mathsf{U}}(\mathcal{T}_{m,n})$, we cannot directly apply the results of Section 5, since d^{U} is defined by taking a random new hash address rather than a random item. We circumvent this by defining $d^{\mathsf{u}}(\mathcal{T})$ to be $d^{\mathsf{U}}(\mathcal{T})$ conditioned on being nonzero; this equals the remaining length of the current block at a random item, and the arguments above apply. The limit random variable is D^{u}_α, defined as D^{U}_α conditioned on being nonzero. We have, using Theorem 6.1 and $\mathbb{P}(D^{\mathsf{U}}_\alpha > 0) = 1 - p^{\mathsf{U}}_\alpha(0) = \alpha$,

$$\mathbb{E}\, D^{\mathsf{u}}_\alpha = \alpha^{-1}\,\mathbb{E}\, D^{\mathsf{U}}_\alpha = \frac{2-\alpha}{2(1-\alpha)^2} = \tfrac12(1-\alpha)^{-2} + \tfrac12(1-\alpha)^{-1} \tag{6.2}$$

and Corollary 5.4 yields, since $\mathbb{P}\big(d^{\mathsf{U}}(\mathcal{T}_{m,n}) > 0\big) = n/m$,

$$\begin{aligned}\mathbb{E}\, d^{\mathsf{U}}(\mathcal{T}_{m,n}) &= \frac{n}{m}\,\mathbb{E}\, d^{\mathsf{u}}(\mathcal{T}_{m,n})\\ &= \tfrac12\big(Q_1(m,n) - Q_0(m,n)\big) + \tfrac12\big(Q_0(m,n) - Q_{-1}(m,n)\big)\\ &= \tfrac12 Q_1(m,n) - \tfrac12,\end{aligned}$$

in accordance with [14], [15, Theorem 6.4.K] (where $d^{\mathsf{U}} + 1$ is studied). Similarly,

$$\begin{aligned}\mathbb{E}(D^{\mathsf{u}}_\alpha)^2 = \alpha^{-1}\,\mathbb{E}(D^{\mathsf{U}}_\alpha)^2 &= \frac{6+3\alpha-4\alpha^2+\alpha^3}{6(1-\alpha)^4}\\ &= (1-\alpha)^{-4} + \tfrac13(1-\alpha)^{-3} - \tfrac16(1-\alpha)^{-2} - \tfrac16(1-\alpha)^{-1}\end{aligned} \tag{6.3}$$

and Corollary 5.4 yields

$$\begin{aligned}\mathbb{E}\big(d^{\mathsf{U}}(\mathcal{T}_{m,n})\big)^2 &= \frac{n}{m}\,\mathbb{E}\big(d^{\mathsf{u}}(\mathcal{T}_{m,n})\big)^2\\ &= Q_3(m,n) - \tfrac23 Q_2(m,n) - \tfrac12 Q_1(m,n) + \tfrac16,\end{aligned}$$

in accordance with the formula for $\mathbb{E}\big(d^{\mathsf{U}}(\mathcal{T}_{m,n}) + 1\big)^2$ in [15, Answer 6.4-28].

We finally observe from Theorem 6.1 that the radius of convergence $r^{\mathsf{U}}(\alpha)$ of $\psi^{\mathsf{U}}_\alpha(z)$ equals $\big(\alpha e^{1-\alpha}\big)^{-1}$, which is the radius of convergence for $\mathbb{E}\, z^{B_\alpha}$ too.

7. The limit distribution for FC

Theorem 7.1. *The distribution of D^{FC}_α, $0 < \alpha < 1$, is given by*

$$p^{\mathsf{FC}}_\alpha(k) = \mathbb{P}(D^{\mathsf{FC}}_\alpha = k) = \frac{1}{\alpha}\int_0^\alpha p^{\mathsf{U}}_x(k)\,dx = 1 - \frac{\alpha}{2} - \sum_{l=1}^{k} \frac{l^{l-2}}{l!}\alpha^{l-1}e^{-l\alpha}. \tag{7.1}$$

The probability generating function is

$$\psi^{\mathsf{FC}}_\alpha(z) := \mathbb{E}\, z^{D^{\mathsf{FC}}_\alpha} = \sum_{k=0}^{\infty} z^k p^{\mathsf{FC}}_\alpha(k) = \frac{\big(1 - T(z\alpha e^{-\alpha})\big)^2 - \big(1-\alpha\big)^2}{2\alpha(1-z)}. \tag{7.2}$$

The first moments are

$$\mathbb{E}\, D_\alpha^{\mathsf{FC}} = \frac{\alpha}{2(1-\alpha)} = \tfrac12(1-\alpha)^{-1} - \tfrac12, \tag{7.3}$$

$$\mathbb{E}(D_\alpha^{\mathsf{FC}})^2 = \frac{3\alpha-\alpha^3}{6(1-\alpha)^3} = \tfrac13(1-\alpha)^{-3} - \tfrac12(1-\alpha)^{-1} + \tfrac16, \tag{7.4}$$

$$\operatorname{Var}(D_\alpha^{\mathsf{FC}}) = \frac{6\alpha-3\alpha^2+\alpha^3}{12(1-\alpha)^3} = \tfrac13(1-\alpha)^{-3} - \tfrac14(1-\alpha)^{-2} - \tfrac1{12}. \tag{7.5}$$

Proof. In a random hash table $\mathcal{T}_{m,n}$ using the FC rule, the insertion of i:th item can be regarded as an unsuccessful search in the table $\mathcal{T}_{m,i-1}$ constructed so far; hence

$$d_i^{\mathsf{FC}}(\mathcal{T}_{m,n}) \overset{\mathrm{d}}{=} d^{\mathsf{U}}(\mathcal{T}_{m,i-1}), \qquad 1 \le i \le n, \tag{7.6}$$

and

$$\mathbb{E}\, n_k^{\mathsf{FC}}(\mathcal{T}_{m,n}) = \sum_{i=1}^{n} \mathbb{P}\big(d_i^{\mathsf{FC}}(\mathcal{T}_{m,n}) = k\big) = \sum_{i=1}^{n} \mathbb{P}\big(d^{\mathsf{U}}(\mathcal{T}_{m,i-1}) = k\big).$$

Consequently, using Corollary 4.6 and dominated convergence,

$$\begin{aligned}\frac1n \mathbb{E}\, n_k^{\mathsf{FC}}(\mathcal{T}_{m,n}) &= \frac1n \sum_{i=0}^{n-1} \mathbb{P}(d^{\mathsf{U}}(\mathcal{T}_{m,i}) = k) = \frac{m}{n}\int_0^{n/m} \mathbb{P}(d^{\mathsf{U}}(\mathcal{T}_{m,\lfloor mx\rfloor}) = k)\,dx \\ &\to \frac1\alpha \int_0^\alpha \mathbb{P}(D_x^{\mathsf{U}} = k)\,dx = \frac1\alpha\int_0^\alpha p_x^{\mathsf{U}}(k)\,dx,\end{aligned}$$

which by Theorem 4.1 yields $p_\alpha^{\mathsf{FC}}(k) = \frac1\alpha \int_0^\alpha p_x^{\mathsf{U}}(k)\,dx$, which is the first part of (7.1). The final equality in (7.1) is verified, using (6.1), by multiplying the last expression by α and differentiating.

As a consequence of (7.1) and (3.2),

$$\begin{aligned}\psi_\alpha^{\mathsf{FC}}(z) &= \sum_{k=0}^\infty z^k p_\alpha^{\mathsf{FC}}(k) = \sum_{k=0}^\infty z^k\Big(1-\frac\alpha2\Big) - \sum_{1\le l\le k<\infty} z^k \frac{l^{l-2}}{l!}\alpha^{l-1}e^{-l\alpha} \\ &= \frac{1-\alpha/2}{1-z} - \frac{1}{1-z}\sum_{l=1}^\infty z^l \frac{l^{l-2}}{l!}\alpha^{l-1}e^{-l\alpha} \\ &= \frac{1-\alpha/2}{1-z} - \frac{T(z\alpha e^{-\alpha}) - \frac12 T(z\alpha e^{-\alpha})^2}{(1-z)\alpha},\end{aligned}$$

which can be rewritten as (7.2). The moments are obtained by differentiation. ∎

For example, $p_\alpha^{\mathsf{FC}}(0) = 1-\frac12\alpha$, $p_\alpha^{\mathsf{FC}}(1) = 1-\frac12\alpha - e^{-\alpha}$, $p_\alpha^{\mathsf{FC}}(2) = 1-\frac12\alpha - e^{-\alpha} - \frac12\alpha e^{-2\alpha}$.

From (7.3) and Corollary 5.4 we obtain the well-known formula [14], [15, Theorem 6.4.K]

$$\mathbb{E}\, d^{\mathsf{FC}}(\mathcal{T}_{m,n}) = \tfrac12 Q_0(m,n-1) - \tfrac12.$$

Similarly, (7.4) and Corollary 5.4 yields

$$\mathbb{E}\big(d^{\mathsf{FC}}(\mathcal{T}_{m,n})\big)^2 = \frac{m}{n}\frac{2Q_2(m,n)-2Q_1(m,n)-3Q_0(m,n)+3}{6}+\frac{1}{6},$$

in accordance with [15, Answer 6.4-67].

It can be seen from the form of $\psi^{\mathsf{FC}}_\alpha$ in (7.2) and (3.3) that every integer moment $\mathbb{E}(D^{\mathsf{FC}}_\alpha)^r$ is a polynomial in $1/(1-\alpha)$, and thus Corollary 5.4 shows that every moment of $d^{\mathsf{FC}}(\mathcal{T}_{m,n})$ can be expressed in Q functions. We leave it to the reader as an exercise to find explicit formulas for, say, the third and fourth moments.

Similarly, we can obtain an exact formula for the distribution of $d^{\mathsf{FC}}(\mathcal{T}_{m,n})$.

Theorem 7.2.

$$\mathbb{P}\big(d^{\mathsf{FC}}(\mathcal{T}_{m,n})=k\big) = 1-\frac{n-1}{2m}-\sum_{l=1}^{k}\frac{l^{l-2}(n-1)^{\underline{l-1}}(m-l)^{n-l}}{l!\,m^{n-1}}.$$

Proof. By Corollary 5.3 and (7.1), the probability equals

$$\frac{(n-1)!}{m^{n-1}}[\alpha^{n-1}]\Big(e^{m\alpha}-\frac{\alpha}{2}e^{m\alpha}-\sum_{l=1}^{k}\frac{l^{l-2}}{l!}\alpha^{l-1}e^{(m-l)\alpha}\Big)$$

and the result follows. □

We finally observe from Theorem 7.1 that the radius of convergence $r^{\mathsf{FC}}(\alpha)$ of $\psi^{\mathsf{FC}}_\alpha(z)$ equals $r^{\mathsf{U}}(\alpha) = \big(\alpha e^{1-\alpha}\big)^{-1}$. Hence, $p^{\mathsf{FC}}_\alpha(k)$ decrease geometrically roughly as $r^{\mathsf{FC}}(\alpha)^{-k}$ as $k\to\infty$. More precisely, we have the following asymptotics.

Theorem 7.3. *Let $0<\alpha<1$ be fixed. Then, as $k\to\infty$,*

$$p^{\mathsf{FC}}_\alpha(k)\sim\frac{1}{\sqrt{2\pi}(e^{\alpha-1}-\alpha)}k^{-5/2}\big(\alpha e^{1-\alpha}\big)^k.$$

Proof. A simple consequence of (7.1) and Stirling's formula. We omit the details. □

8. The limit distribution for RH

For Robin Hood hashing, we have the following explicit formula for the generating function of the limit distribution in the sparse case.

Theorem 8.1. *The probability generating function of D^{RH}_α, $0<\alpha<1$, is*

$$\psi^{\mathsf{RH}}_\alpha(z) := \mathbb{E}\,z^{D^{\mathsf{RH}}_\alpha} = \sum_{k=0}^{\infty}z^k p^{\mathsf{RH}}_\alpha(k) = \frac{1-\alpha}{\alpha}\cdot\frac{e^{\alpha(1-z)}-1}{1-ze^{\alpha(1-z)}} = \frac{1-\alpha}{\alpha}\cdot\frac{e^{\alpha z}-e^{\alpha}}{ze^{\alpha}-e^{\alpha z}}.$$

The first moments are

$$\mathbb{E}\, D^{\mathsf{RH}}_{\alpha} = \frac{\alpha}{2(1-\alpha)} = \tfrac12(1-\alpha)^{-1}-\tfrac12, \tag{8.1}$$

$$\mathbb{E}(D^{\mathsf{RH}}_{\alpha})^2 = \frac{3\alpha-\alpha^2+\alpha^3}{6(1-\alpha)^2} = \tfrac12(1-\alpha)^{-2}-\tfrac23(1-\alpha)^{-1}+\tfrac16+\tfrac16\alpha, \tag{8.2}$$

$$\operatorname{Var}(D^{\mathsf{RH}}_{\alpha}) = \frac{6\alpha-5\alpha^2+2\alpha^3}{12(1-\alpha)^2} = \tfrac14(1-\alpha)^{-2}-\tfrac16(1-\alpha)^{-1}-\tfrac1{12}+\tfrac16\alpha. \tag{8.3}$$

Remark 8.2. The probabilities p^{RH}_{α} can be obtained from the generating function $\psi^{\mathsf{RH}}_{\alpha}$. For example,

$$p^{\mathsf{RH}}_{\alpha}(0) = (1-\alpha)\frac{e^{\alpha}-1}{\alpha}, \tag{8.4}$$

$$p^{\mathsf{RH}}_{\alpha}(1) = (1-\alpha)e^{\alpha}\frac{e^{\alpha}-1-\alpha}{\alpha}. \tag{8.5}$$

Moreover, as pointed out by the referee, there is a connection with Eulerian numbers [15, Section 5.1.3], [8, Section 6.2]. Indeed, from the formula above for $\psi^{\mathsf{RH}}_{\alpha}(z)$ and [8, (7.60)] or [15, 5.1.3-(20)] follows easily

$$p^{\mathsf{RH}}_{\alpha}(k) = \sum_{l=1}^{\infty}(1-\alpha)\alpha^{l-1}\left\langle {l \atop k} \right\rangle \Big/ l!.$$

This shows the curious fact that D^{RH}_{α} has the same distribution as the number of descents in a random permutation of random length L, where L has the geometric distribution $\mathbb{P}(L=l)=(1-\alpha)\alpha^{l-1}$, $l\ge 1$.

Proof. Let $F(z)$ be the generating function $\sum_{k=0}^{\infty}\mathbb{E}\, n^{\mathsf{RH}}_k(\mathcal{T}_\alpha)z^k$; thus $\psi^{\mathsf{RH}}_{\alpha}(z) = \big((1-\alpha)/\alpha\big)F(z)$ by Theorem 4.1. By Lemmas 2.1 and 3.4,

$$F(z) = \mathbb{E}\sum_{i=1}^{\tau-1} z^{d^{\mathsf{RH}}_i} = \mathbb{E}\sum_{j=1}^{\tau-1} z^{H_j-1} = \mathbb{E}\sum_{j=1}^{\tau-1} z^{S_j},$$

where S_j and τ are as in Section 3.

Recall that the random walk S_j starts with $S_0=0$ and ends with $S_\tau=-1$. We also consider the shifted random walk S_j+r, starting with $r\ge 0$, which we run until it hits 0, i.e. until $\tau_r := \min\{j\ge 0 : S_j=-r\}$. We define

$$G_r(z) := \mathbb{E}\sum_{j=1}^{\tau_r} z^{S_j+r},$$

and note that $G_0(z)=0$ and $G_1(z)=zF(z)+1$.

If $r\ge 1$, then $\tau=\tau_1$ is the time the shifted random walk first hits $r-1$, and $S_\tau+r, S_{\tau+1}+r,\dots,S_{\tau_r}+r$ is a random walk starting at $r-1$ and stopped when it hits 0, i.e. a copy of $S_0+(r-1),\dots,S_{\tau_{r-1}}+(r-1)$. Hence,

$$G_r(z) = \mathbb{E}\sum_{j=1}^{\tau} z^{S_j+r} + \mathbb{E}\sum_{j=\tau+1}^{\tau_r} z^{S_j+r} = z^{r-1}G_1(z)+G_{r-1}(z),$$

and thus

$$G_r(z) = (z^{r-1} + \cdots + 1)G_1(z) = \frac{1-z^r}{1-z}G_1(z), \qquad r \geq 0.$$

Moreover, taking $r = 1$, we start at $S_0 + 1 = 1$ and the next element is $S_1 + 1 = \xi_1 \in \mathrm{Po}(\alpha)$, so the remainder of the random walk is a random walk started at ξ_1. Conditioning on ξ_1, we thus find

$$\begin{aligned} G_1(z) &= \sum_{r=0}^{\infty} \mathbb{P}(\xi_1 = r)\big(z^r + G_r(z)\big) = \sum_{r=0}^{\infty} \mathbb{P}(\xi_1 = r)\Big(z^r + \frac{1-z^r}{1-z}G_1(z)\Big) \\ &= \sum_{r=0}^{\infty} \mathbb{P}(\xi_1 = r)z^r\Big(1 - \frac{1}{1-z}G_1(z)\Big) + \frac{1}{1-z}G_1(z) \\ &= e^{\alpha z-\alpha}\Big(1 - \frac{1}{1-z}G_1(z)\Big) + \frac{1}{1-z}G_1(z) \end{aligned}$$

with the solution

$$G_1(z) = \frac{1-z}{1-ze^{\alpha(1-z)}}.$$

Hence,

$$F(z) = \frac{G_1(z)-1}{z} = \frac{e^{\alpha(1-z)}-1}{1-ze^{\alpha(1-z)}}.$$

The formula for $\psi_\alpha^{\mathsf{RH}}$ follows, and the moments are obtained by differentiations at $z = 1$. $\square$

Note that $\mathbb{E}\, D_\alpha^{\mathsf{RH}} = \mathbb{E}\, D_\alpha^{\mathsf{FC}}$, as we already know since the average displacement is the same for any insertion policy. Similarly, $\mathbb{E}\, d^{\mathsf{RH}}(\mathcal{T}_{m,n}) = \mathbb{E}\, d^{\mathsf{FC}}(\mathcal{T}_{m,n}) = \frac{1}{2}Q_0(m, n-1) - \frac{1}{2}$. For the second moment, (8.2) can be written

$$\mathbb{E}(D_\alpha^{\mathsf{RH}})^2 = \tfrac{1}{2}(1-\alpha)^{-2} - \tfrac{2}{3}(1-\alpha)^{-1} + \tfrac{1}{3} - \tfrac{1}{6}(1-\alpha),$$

and Corollary 5.4 yields, using $Q_{-2}(m, n) = 1 - n/m$,

$$\begin{aligned} \mathbb{E}\big(d^{\mathsf{RH}}(\mathcal{T}_{m,n})\big)^2 &= \frac{1}{2}Q_1(m, n-1) - \frac{2}{3}Q_0(m, n-1) + \frac{1}{6} + \frac{n-1}{6m} \\ &= \frac{m}{n}\frac{3Q_1(m,n) - 7Q_0(m,n) + 4}{6} + \frac{1}{6} + \frac{n-1}{6m}, \end{aligned}$$

which easily is shown to be equivalent to the formula in [15, Answer 6.4-67].

It is easily seen that each integer moment of D_α^{RH} is a rational function in α, with denominator a power of $1-\alpha$. Hence each integer moment $\mathbb{E}\big(d^{\mathsf{RH}}(\mathcal{T}_{m,n})\big)^r$ can be expressed in Q functions (allowing negative indices; these terms form a polynomial in n/m and $1/m$).

We can also find exact formulas for point probabilities. For example, by Corollary 5.3 and (8.4),

$$\begin{aligned} \mathbb{P}\big(d^{\mathsf{RH}}(\mathcal{T}_{m,n}) = 0\big) &= \frac{(n-1)!}{m^{n-1}}[\alpha^{n-1}]\Big(e^{m\alpha}\frac{1-\alpha}{\alpha}(e^\alpha - 1)\Big) \\ &= \frac{1}{n}\Big((m+1-n)\Big(1+\frac{1}{m}\Big)^{n-1} - (m-n)\Big). \end{aligned}$$

By the formula in Theorem 8.1, $\psi_\alpha^{\mathsf{RH}}(z)$ is a meromorphic funcion of z in the entire complex plane. The poles are the roots of $z = e^{\alpha(z-1)}$, except $z = 1$. Since the Taylor coefficients p_α^{RH} are non-negative, the pole closest to the origin lies on the positive real axis; moreover, it is easily seen that all other poles have strictly larger absolute values. The following theorem follows easily.

Theorem 8.3. *Let $0 < \alpha < 1$. The radius of convergence $r^{\mathsf{RH}}(\alpha)$ of $\psi_\alpha^{\mathsf{RH}}(z)$ is the unique root $r > 1$ of $r = e^{\alpha(r-1)}$. If $\alpha^* > 1$ satisfies $\alpha^* e^{-\alpha^*} = \alpha e^{-\alpha}$, then $r^{\mathsf{RH}}(\alpha) = \alpha^*/\alpha$. Moreover, as $k \to \infty$,*

$$p_\alpha^{\mathsf{RH}}(k) \sim \frac{(1-\alpha)(\alpha^*-\alpha)}{\alpha\alpha^*(\alpha^*-1)} r^{\mathsf{RH}}(\alpha)^{-k}.$$

Note that $r^{\mathsf{RH}}(\alpha) > 1/\alpha > r^{\mathsf{FC}}(\alpha)$; another expression of the fact that large deviations are less likely for RH than for FC.

9. The limit distribution for LC

The Last Come policy seems to be the most difficult to analyse, and we are not able to give as explicit results as for the other policies. The simplest form of the probability generating function that we have been able to find is the following. It is quite possible that others may simplify the result, but the expression for the variance shows that this case is intrinsically more complicated than the two other policies considered here.

Theorem 9.1. *Define*

$$\begin{aligned}
u(\alpha,\beta,z) &:= \frac{1}{z+(1-z)e^{\alpha-\beta}} = \frac{e^\beta}{(1-z)(e^\alpha - e^\beta)+e^\beta},\\
v(\alpha,\beta,z) &:= u(\alpha,\beta,z)\beta e^{-\beta} = \frac{\beta}{ze^\beta + (1-z)e^\alpha},\\
w(\alpha,\beta,z) &:= \int_\beta^\alpha \frac{u(\alpha,\gamma,z)}{1-T(v(\alpha,\gamma,z))} d\gamma.
\end{aligned}$$

Then, for $0 < \alpha < 1$,

$$\begin{aligned}
\psi_\alpha^{\mathsf{LC}}(z) &= \frac{1-\alpha}{\alpha}\int_0^\alpha e^{zw(\alpha,\beta,z)} \frac{u(\alpha,\beta,z)}{T(v(\alpha,\beta,z))} \frac{\beta - T(v(\alpha,\beta,z))}{1-u(\alpha,\beta,z)}\, d\beta\\
&= \frac{1-\alpha}{\alpha}\int_0^\alpha e^{zw(\alpha,\beta,z)} \frac{v(\alpha,\beta,z)}{T(v(\alpha,\beta,z))} \int_0^1 T'\big((1-t+tu(\alpha,\beta,z))\beta e^{-\beta}\big)\, dt\, d\beta.
\end{aligned} \tag{9.1}$$

The first moments are

$$\mathbb{E}\,D_\alpha^{\mathsf{LC}} = \frac{\alpha}{2(1-\alpha)} = \frac{1}{2(1-\alpha)} - \frac{1}{2}, \tag{9.2}$$

$$\mathbb{E}(D_\alpha^{\mathsf{LC}})^2 = \frac{1}{2(1-\alpha)^2} - \frac{1}{6(1-\alpha)} - \frac{e^\alpha - 1}{3\alpha} - \frac{e^{\alpha-1}}{6\alpha}\big(\mathrm{Ei}(1) - \mathrm{Ei}(1-\alpha)\big) + \frac{1}{6}, \tag{9.3}$$

$$\mathrm{Var}(D_\alpha^{\mathsf{LC}}) = \frac{1}{4(1-\alpha)^2} + \frac{1}{3(1-\alpha)} - \frac{e^\alpha - 1}{3\alpha} - \frac{e^{\alpha-1}}{6\alpha}\big(\mathrm{Ei}(1) - \mathrm{Ei}(1-\alpha)\big) - \frac{1}{12}. \tag{9.4}$$

Here Ei is the exponential integral function, $\mathrm{Ei}(1) - \mathrm{Ei}(1-\alpha) = \int_{1-\alpha}^{1} \frac{e^x}{x}\,dx$.

Proof. We build heavily on the analysis of the first two moments by Poblete, Viola and Munro [23, 24, 28], who proved (9.3) and (9.4) (in a different form).

We use (5.3) and (5.4a). To keep track of the displacements and obtain formulas for $\Phi_\ell^{\mathsf{LC}}(z)$, we keep track also of the positions of the items in their blocks (in the final table). Thus, let $l_i(\mathcal{T})$ be the number of items with final position before item i in the same block in the hash table $\mathcal{T}$ with the LC rule, and consider the bivariate generating function $\Phi_\ell(z,y)$ equal to the sum of $z^{d_i^{\mathsf{LC}}(\mathcal{T})}y^{l_i(\mathcal{T})}$ over all $\ell^{\ell-2}$ confined almost full hash tables of length ℓ, and all $i \in \{1,\dots,\ell-1\}$. Hence, in the notation of (5.3) and (5.4a), $\Phi_\ell^{\mathsf{LC}}(z) = \Phi_\ell(z,1)$.

In the notation of [23],

$$\Phi_k(z,y) = \sum_{l+r+2=k} z^{-1}F_{l,r}(z)y^l \tag{9.5}$$

(the factor z^{-1} is because [23] studies the number of probes, i.e. $1+$ the displacement); moreover, their $C_i(z) := \sum_{l+r=i} F_{l,r}(z) = z\Phi_{i+2}(z,1) = z\Phi_{i+2}^{\mathsf{LC}}(z)$.

Further, define the trivariate generating function (note that $\Phi_1(z,y) = 0$)

$$\begin{aligned}\Psi(z,y,\lambda) &:= \sum_{i=0}^\infty \frac{\lambda^i}{i!}\Phi_{i+1}(z,y) = \sum_{i=0}^\infty \frac{\lambda^{i+1}}{(i+1)!}\Phi_{i+2}(z,y)\\ &= \sum_{l,r\ge 0} z^{-1}F_{l,r}(z)y^l \frac{\lambda^{l+r+1}}{(l+r+1)!}.\end{aligned} \tag{9.6}$$

The sums converge at least for $|z| \le 1$, $|y| \le 1$, $|\lambda| < e^{-1}$, and Ψ is continuous in that domain and analytic in its interior. By (5.4a) and (9.6), we have

$$\psi_\alpha^{\mathsf{LC}}(z) = \frac{1-\alpha}{\alpha}e^{-\alpha}\sum_{\ell=1}^\infty \frac{\big(\alpha e^{-\alpha}\big)^{\ell-1}}{(\ell-1)!}\Phi_\ell(z,1) = \frac{1-\alpha}{\alpha}e^{-\alpha}\Psi(z,1,\alpha e^{-\alpha}). \tag{9.7}$$

Poblete, Viola and Munro [23, (70)] give the recursion formula, for all $l, r \geq 0$ with the convention that $F_{l,r} = 0$ if $l < 0$ or $r < 0$,

$$\begin{aligned} F_{l,r}(z) = z \sum_{0\leq k\leq r} \binom{l+r}{k} (k+1)^{k-1} (l+r-k+1)^{l+r-k-1} \\ + \sum_{0\leq k\leq l+r} \binom{l+r}{k} (k+1)^{k-1} \bigl(F_{l-k-1,r}(z)(k+1) + lzF_{l-1,r-k}(z) \\ + (r-k)F_{l,r-k-1}(z) \bigr). \end{aligned}$$

Hence, using (9.6), with $1/i! = 0$ if $i < 0$,

$$\begin{aligned} z\frac{\partial \Psi(z,y,\lambda)}{\partial \lambda} &= \sum_{l,r\geq 0} \frac{\lambda^{l+r}}{(l+r)!} y^l F_{l,r}(z) \\ &= z \sum_{\substack{r\geq k\geq 0\\ l\geq 0}} \frac{\lambda^{l+r} y^l}{(l+r-k)!\,k!} (k+1)^{k-1} (l+r-k+1)^{l+r-k-1} \\ &\quad + \sum_{k,l,r} \frac{\lambda^{l+r} y^l}{(l+r-k)!\,k!} (k+1)^k F_{l-k-1,r}(z) \\ &\quad + z \sum_{k,l,r} \frac{\lambda^{l+r} y^l}{(l+r-k)!\,k!} (k+1)^{k-1} l F_{l-1,r-k}(z) \\ &\quad + \sum_{k,l,r} \frac{\lambda^{l+r} y^l}{(l+r-k)!\,k!} (k+1)^{k-1} (r-k) F_{l,r-k-1}(z) \\ &= zS_I + S_{II} + zS_{III} + S_{IV}. \end{aligned} \tag{9.8}$$

To evaluate the sums $S_I, S_{II}, S_{III}, S_{IV}$, we first rewrite (3.1) as

$$T(z) = \sum_{j=0}^{\infty} \frac{(j+1)^j z^{j+1}}{(j+1)!} = \sum_{j=0}^{\infty} \frac{(j+1)^{j-1} z^{j+1}}{j!} \tag{9.9}$$

and further differentiate to obtain, using (3.3),

$$\sum_{j=0}^{\infty} \frac{(j+1)^j z^{j+1}}{j!} = zT'(z) = \frac{T(z)}{1-T(z)}. \tag{9.10}$$

We now change summation indices, using $j = l + r - k$, and obtain by (9.9)

$$
\begin{aligned}
S_I &= \sum_{\substack{j,k,l\\ j\geq l\geq 0}} \frac{\lambda^{j+k}y^l}{j!\,k!}(k+1)^{k-1}(j+1)^{j-1}\\
&= \sum_{j,k\geq 0} \frac{\lambda^k}{k!}(k+1)^{k-1}\frac{\lambda^j}{j!}(j+1)^{j-1}\frac{1-y^{j+1}}{1-y}\\
&= \frac{1}{1-y}\sum_k \frac{\lambda^k}{k!}(k+1)^{k-1}\sum_j \frac{\lambda^{j+1}-(y\lambda)^{j+1}}{\lambda j!}(j+1)^{j-1}\\
&= \frac{1}{1-y}\frac{T(\lambda)}{\lambda}\frac{T(\lambda)-T(y\lambda)}{\lambda}.
\end{aligned}
$$

Similarly, using also $m = l - k - 1$ and (9.10), (9.6),

$$
\begin{aligned}
S_{II} &= \sum_{j,k,m} \frac{\lambda^{j+k}y^{m+k+1}}{j!\,k!}(k+1)^k F_{m,j-m-1}(z)\\
&= \lambda^{-1}\sum_k \frac{(\lambda y)^{k+1}}{k!}(k+1)^k \sum_{j,m}\frac{\lambda^j}{j!}y^m F_{m,j-m-1}(z)\\
&= \lambda^{-1}\frac{T(y\lambda)}{1-T(y\lambda)}z\Psi(z,y,\lambda),\\
S_{III} &= \sum_{j,k,l} \frac{\lambda^{j+k}y^l}{j!\,k!}(k+1)^{k-1} l F_{l-1,j-l}(z)\\
&= \frac{T(\lambda)}{\lambda}y\sum_{j,l}\frac{\lambda^j}{j!}ly^{l-1}F_{l-1,j-l}(z)\\
&= \frac{T(\lambda)}{\lambda}y\frac{\partial}{\partial y}\bigl(zy\Psi(z,y,\lambda)\bigr),\\
S_{IV} &= \sum_{j,k,l} \frac{\lambda^{j+k}y^l}{j!\,k!}(k+1)^{k-1}(j-l)F_{l,j-l-1}(z)\\
&= \frac{T(\lambda)}{\lambda}\Bigl(\lambda\frac{\partial}{\partial\lambda}-y\frac{\partial}{\partial y}\Bigr)\sum_{j,l}\frac{\lambda^j}{j!}y^l F_{l,j-l-1}(z)\\
&= \frac{T(\lambda)}{\lambda}\Bigl(\lambda\frac{\partial}{\partial\lambda}-y\frac{\partial}{\partial y}\Bigr)\Bigl(z\Psi(z,y,\lambda)\Bigr).
\end{aligned}
$$

Summing up, (9.8) yields

$$\begin{aligned}\frac{\partial\Psi(z,y,\lambda)}{\partial\lambda} &= \frac{1}{\lambda^2(1-y)}T(\lambda)\bigl(T(\lambda)-T(y\lambda)\bigr)+\frac{1}{\lambda}\frac{T(y\lambda)}{1-T(y\lambda)}\Psi(z,y,\lambda)\\ &\quad+\frac{1}{\lambda}T(\lambda)yz\Psi(z,y,\lambda)+\frac{1}{\lambda}T(\lambda)zy^2\frac{\partial\Psi(z,y,\lambda)}{\partial y}\\ &\quad+T(\lambda)\frac{\partial\Psi(z,y,\lambda)}{\partial\lambda}-\frac{y}{\lambda}T(\lambda)\frac{\partial\Psi(z,y,\lambda)}{\partial y},\end{aligned}$$

which after multiplication by λ can be rearranged to

$$\begin{aligned}\lambda\bigl(1-T(\lambda)\bigr)&\frac{\partial\Psi(z,y,\lambda)}{\partial\lambda}+T(\lambda)y(1-yz)\frac{\partial\Psi(z,y,\lambda)}{\partial y}\\ &=\Bigl(T(\lambda)yz+\frac{T(y\lambda)}{1-T(y\lambda)}\Bigr)\Psi(z,y,\lambda)+T(\lambda)\frac{T(\lambda)-T(y\lambda)}{\lambda(1-y)}\end{aligned}\tag{9.11}$$

We simplify a little by the change of variable $\lambda=\alpha e^{-\alpha}$. We then have $\alpha=T(\lambda)$ and $\partial/\partial\alpha=(1-\alpha)e^{-\alpha}\partial/\partial\lambda$. We write $\tilde\Psi(z,y,\alpha)=\Psi(z,y,\alpha e^{-\alpha})$, noting for later use that this conveniently also appears in (9.7), which can be written

$$\psi_\alpha^{\mathsf{LC}}(z)=\frac{1-\alpha}{\alpha}e^{-\alpha}\tilde\Psi(z,1,\alpha).\tag{9.12}$$

Returning to (9.11), we obtain after substitution, and division by α,

$$\begin{aligned}\frac{\partial\tilde\Psi(z,y,\alpha)}{\partial\alpha}&+y(1-yz)\frac{\partial\tilde\Psi(z,y,\alpha)}{\partial y}\\ &=\Bigl(yz+\frac{T(y\alpha e^{-\alpha})}{\alpha\bigl(1-T(y\alpha e^{-\alpha})\bigr)}\Bigr)\tilde\Psi(z,y,\alpha)+\frac{\alpha-T(y\alpha e^{-\alpha})}{\alpha e^{-\alpha}(1-y)},\end{aligned}\tag{9.13}$$

valid at least for $|z|\le 1$, $|y|\le 1$ and $0\le\alpha<1$. Note that this partial differential equation contains no $\partial/\partial z$; hence we may regard z as a constant. We solve the differential equation in the standard way: The characteristics of (9.13) are given by

$$\frac{dy}{d\alpha}=y(1-yz)\tag{9.14}$$

or

$$\frac{d\alpha}{dy}=\frac{1}{y(1-yz)}=\frac{1}{y}+\frac{z}{1-yz}$$

with the solutions

$$\alpha-\alpha_1=\ln y-\ln(1-yz)+\ln(1-z),$$

where α_1 is a constant (the value of α when $y=1$), or, equivalently,

$$y=\frac{1}{z+(1-z)e^{\alpha_1-\alpha}}=u(\alpha_1,\alpha,z).\tag{9.15}$$

Assume that $0\le z\le 1$ and $0\le\alpha_1<1$ and define $y(\alpha)$ by (9.15). Then (9.13) gives, renaming α to β,

$$\frac{d\tilde\Psi(z,y(\beta),\beta)}{d\beta}=g(\beta)\tilde\Psi(z,y(\beta),\beta)+h(\beta)\tag{9.16}$$

with, where $y = y(\beta) = u(\alpha_1, \beta, z)$,

$$g(\beta) = yz + \frac{T(y\beta e^{-\beta})}{\beta\bigl(1 - T(y\beta e^{-\beta})\bigr)} = yz + \frac{T(v(\alpha_1, \beta, z))}{\beta\bigl(1 - T(v(\alpha_1, \beta, z))\bigr)}$$

and

$$h(\beta) = \frac{\beta - T(y\beta e^{-\beta})}{\beta e^{-\beta}(1-y)} = \int_0^1 T'\bigl((1-t+ty)\beta e^{-\beta}\bigr)\,dt.$$

The solution to (9.16) is, because $\tilde{\Psi}(z, y, 0) = \Psi(z, y, 0) = 0$ by (9.6),

$$\tilde{\Psi}(z, y(\alpha), \alpha) = \int_0^\alpha e^{G(\alpha)-G(\beta)} h(\beta)\,d\beta \tag{9.17}$$

where $G' = g$. Since, by (9.14),

$$\frac{d}{d\beta}\bigl(y(\beta)\beta e^{-\beta}\bigr) = y(1-yz)\beta e^{-\beta} + ye^{-\beta} - y\beta e^{-\beta} = y(1-yz\beta)e^{-\beta}$$

we have, using (3.3), with $T = T(y\beta e^{-\beta}) = T(v(\alpha_1, \beta, z))$,

$$\begin{aligned}\frac{d}{d\beta}T\bigl(y(\beta)\beta e^{-\beta}\bigr) &= \frac{T}{1-T}\frac{y(1-yz\beta)e^{-\beta}}{y\beta e^{-\beta}} \\ &= \frac{T}{1-T}\Bigl(\frac{1}{\beta} - yz\Bigr) = g(\beta) - yz\Bigl(1 + \frac{T}{1-T}\Bigr).\end{aligned}$$

Hence, we can choose

$$G(\beta) = T\bigl(v(\alpha_1, \beta, z)\bigr) - zw(\alpha_1, \beta, z)$$

and (9.17) yields, with $\alpha = \alpha_1$ and dropping the subscript,

$$\tilde{\Psi}(z, 1, \alpha) = \int_0^\alpha e^{T(v(\alpha,\alpha,z)) - T(v(\alpha,\beta,z)) + zw(\alpha,\beta,z)} h(\beta)\,d\beta.$$

Finally, we use (9.12) and the facts that $T\bigl(v(\alpha,\alpha,z)\bigr) = T(\alpha e^{-\alpha}) = \alpha$ and $e^{-T(v)} = v/T(v)$. The formula (9.1) follows.

In principle, the moments can be computed by repeated differentiation of (9.1) with respect to z (under the integral signs) and then letting $z = 1$, noting that $u(\alpha, \beta, 1) = 1$, $v(\alpha, \beta, 1) = \beta e^{-\beta}$ and $w(\alpha, \beta, 1) = \ln(1-\beta) - \ln(1-\alpha)$. However, the expressions become very complicated. Indeed, even to verify the trivial $\psi_\alpha^{\mathsf{LC}}(1) = 1$ from (9.1) takes a little effort. We have verified the first moment (9.2) this way, using computer algebra (`Maple`), but failed to obtain the formulas (9.3), (9.4) for the second moment. (No doubt, a more skillful person would succeed.)

A much shorter proof is to simply refer to the asymptotic formulas in Poblete, Viola and Munro [23], in particular [23, Theorem 28], together with Corollary 4.9.

We will, however, give a third proof, which is closely related to the proof in [23] but formulated using the functions and equations used above; we hope that this may illuminate the arguments used in [23] and their connection to the present paper.

As a warm-up, we begin with the mean (9.2), although we already know the result because $\mathbb{E}\, D_\alpha^{\mathsf{LC}} = \mathbb{E}\, D_\alpha^{\mathsf{FC}} = \mathbb{E}\, D_\alpha^{\mathsf{RH}}$. We let D_α, D_z, D_y denote the partial differential operators. Applying D_z to (9.13) and then setting $z = y = 1$ we find

$$\mathrm{D}_z \mathrm{D}_\alpha \tilde\Psi(1,1,\alpha) - \mathrm{D}_y \tilde\Psi(1,1,\alpha) = \tilde\Psi(1,1,\alpha) + \Bigl(1 + \frac{1}{1-\alpha}\Bigr) \mathrm{D}_z \tilde\Psi(1,1,\alpha). \tag{9.18}$$

For convenience, we define

$$F(z,y,\alpha) := (1-\alpha)e^{-\alpha}\tilde\Psi(z,y,\alpha)$$

and write $f(\alpha) := F(1,1,\alpha)$, $f_z(\alpha) := \mathrm{D}_z F(1,1,\alpha)$, $f_y(\alpha) := \mathrm{D}_y F(1,1,\alpha)$, $f_{zz}(\alpha) := \mathrm{D}_z^2 F(1,1,\alpha)$, and so on. Then (9.18) yields, after simple calculations, the ordinary differential equation in the three functions $f(\alpha)$, $f_z(\alpha)$, $f_y(\alpha)$

$$\mathrm{D}_\alpha f_z(\alpha) - f_y(\alpha) = f(\alpha). \tag{9.19}$$

(It may be more natural to consider $\alpha^{-1}F$, cf. (9.12), which for fixed α is a bivariate probability generating function, but the extra factor α simplifies the differential equations.)

Note that, by (9.12), $\psi_\alpha^{\mathsf{LC}}(z) = \alpha^{-1}F(z,1,\alpha)$, and thus

$$f(\alpha) = \alpha\psi_\alpha^{\mathsf{LC}}(1) = \alpha,$$

$$f_z(\alpha) = \alpha\,\mathrm{D}_z\,\psi_\alpha^{\mathsf{LC}}(1) = \alpha\,\mathbb{E}\,D_\alpha^{\mathsf{LC}}, \tag{9.20}$$

$$f_{zz}(\alpha) = \alpha\,\mathrm{D}_z^2\,\psi_\alpha^{\mathsf{LC}}(1) = \alpha\,\mathbb{E}\,D_\alpha^{\mathsf{LC}}\bigl(D_\alpha^{\mathsf{LC}} - 1\bigr), \tag{9.21}$$

and so on.

To solve (9.19), we first have to find $f_y(\alpha)$. One possibility is to use the definition of Φ_ℓ or (9.5) so see

$$\Phi_\ell(1,y) = \ell^{\ell-2}\sum_{i=1}^{\ell-1} y^{i-1} = \ell^{\ell-2}\frac{1-y^{\ell-1}}{1-y}$$

and thus, by (9.6) and (3.1)

$$\Psi(1,y,\lambda) = \sum_{i=1}^\infty \frac{\lambda^{i-1}}{(i-1)!} i^{i-2}\frac{1-y^{i-1}}{1-y} = \frac{1}{1-y}\Bigl(\frac{T(\lambda)}{\lambda} - \frac{T(\lambda y)}{\lambda y}\Bigr)$$

which leads to

$$F(1,y,\alpha) = \frac{1-\alpha}{1-y}\Bigl(1 - \frac{T(\alpha e^{-\alpha} y)}{\alpha y}\Bigr).$$

A calculation using (3.3) yields

$$f_y(\alpha) = \mathrm{D}_y F(1,1,\alpha) = \frac{3\alpha^2 - 2\alpha^3}{2(1-\alpha)^2}.$$

Alternatively, we may observe that if we as in Section 6 use the superscript u for unsuccessful searches conditioned on starting at an existing item, (5.3) yields

$$\Phi_k^{\mathsf{u}}(y) = y\Phi_k(1,y).$$

Beneath our formalism, this simply reflects the fact that the number of items in a block to the right of a distinguished item has the same distribution as the number of items to the left; the factor y is because the distinguished element is included in the count defining d^{u} but not for $\Phi_k(z,y)$.

By (5.4a) and (9.6), we thus have

$$\psi^{\mathsf{u}}_\alpha(y) = \frac{1-\alpha}{\alpha} e^{-\alpha} \sum_{\ell=1}^{\infty} \frac{(\alpha e^{-\alpha})^{\ell-1}}{(\ell-1)!} \Phi^{\mathsf{u}}_\ell(y) = y \frac{1-\alpha}{\alpha} e^{-\alpha} \tilde{\Psi}(1,y,\alpha)$$

and so

$$yF(1,y,\alpha) = \alpha \psi^{\mathsf{u}}_\alpha(y). \tag{9.22}$$

Applying D_y we find at $y=1$, using (6.2),

$$f(\alpha) + f_y(\alpha) = \alpha \, \mathrm{D}_y \, \psi^{\mathsf{u}}_\alpha(y) = \alpha \, \mathbb{E}\, D^{\mathsf{u}}_\alpha = \alpha \frac{2-\alpha}{2(1-\alpha)^2}. \tag{9.23}$$

We can now integrate (9.19), noting $F(z,y,0) = 0$ and thus $f_z(0) = 0$:

$$f_z(\alpha) = \int_0^\alpha \big(f(\beta) + f_y(\beta)\big)\, d\beta = \frac{\alpha^2}{2(1-\alpha)}, \tag{9.24}$$

which by (9.20) yields (9.2).

For the second moment we argue similarly. We apply $\mathrm{D}_z \, \mathrm{D}_z$ to (9.13) and obtain for $z = y = 1$

$$\mathrm{D}_z \, \mathrm{D}_z \, \mathrm{D}_\alpha \, \tilde{\Psi} - 2 \, \mathrm{D}_z \, \mathrm{D}_y \, \tilde{\Psi} = 2 \, \mathrm{D}_z \, \tilde{\Psi} + \Big(1 + \frac{1}{1-\alpha}\Big) \, \mathrm{D}_z \, \mathrm{D}_z \, \tilde{\Psi}$$

which implies

$$\mathrm{D}_\alpha \, f_{zz}(\alpha) - 2 f_{zy}(\alpha) = 2 f_z(\alpha). \tag{9.25}$$

In order to solve this, we first need $f_{zy}(\alpha)$, which we find in the same way. We apply $\mathrm{D}_y \, \mathrm{D}_z$ to (9.13) and obtain for $z = y = 1$

$$\begin{aligned} &\mathrm{D}_y \, \mathrm{D}_z \, \mathrm{D}_\alpha \, \tilde{\Psi} - \mathrm{D}_z \, \mathrm{D}_y \, \tilde{\Psi} - \mathrm{D}_y^2 \, \tilde{\Psi} - 2 \, \mathrm{D}_y \, \tilde{\Psi} = \\ &\qquad \Big(1 + \frac{1}{1-\alpha}\Big) \, \mathrm{D}_y \, \mathrm{D}_z \, \tilde{\Psi} + \Big(1 + \frac{1}{(1-\alpha)^3}\Big) \, \mathrm{D}_z \, \tilde{\Psi} + \mathrm{D}_y \, \tilde{\Psi} + \tilde{\Psi} \end{aligned}$$

and hence

$$\mathrm{D}_\alpha \, f_{zy}(\alpha) = f_{zy}(\alpha) + f_{yy}(\alpha) + 3 f_y(\alpha) + \Big(1 + \frac{1}{(1-\alpha)^3}\Big) f_z(\alpha) + f(\alpha). \tag{9.26}$$

To solve this, we first need $f_{yy}(\alpha)$. We find from (9.22) and (6.2), (6.3)

$$f_{yy}(\alpha) = \alpha \, \mathbb{E}(D^{\mathsf{u}}_\alpha - 1)(D^{\mathsf{u}}_\alpha - 2) = \alpha \, \mathbb{E}\big(D^{\mathsf{u}}_\alpha\big)^2 - 3\alpha \, \mathbb{E}\, D^{\mathsf{u}}_\alpha + 2\alpha = \frac{16\alpha^3 - 19\alpha^4 + 6\alpha^5}{3(1-\alpha)^4}. \tag{9.27}$$

(Alternatively, it is possible to continue in the same manner as above by applying D_y^2 to (9.13) at $z = y = 1$; this yields a differential equation for $f_{yy}(\alpha)$ that can be solved. The same applies to $f_y(\alpha)$ above, which also can be found by applying D_y to (9.13).)

We can now put everything together. Noting that (9.27) and (9.23) imply $f_{yy}(\alpha) + 3 f_y(\alpha) + f(\alpha) = \alpha \, \mathbb{E}\big(D^{\mathsf{u}}_\alpha\big)^2$, we find the solution to (9.26), using (6.3) and (9.24),

$$\begin{aligned}
f_{zy}(\alpha) &= e^{\alpha}\int_0^{\alpha} e^{-t}\Bigl(t\,\mathbb{E}\bigl(D^{\mathsf{u}}_t\bigr)^2 + \Bigl(1+\frac{1}{(1-t)^3}\Bigr)f_z(t)\Bigr)\,dt\\
&= \frac{1}{2(1-\alpha)^3} - \frac{7}{12(1-\alpha)^2} - \frac{7}{12(1-\alpha)} + \frac{5}{6} + \frac{\alpha}{2} - \frac{e^{\alpha}}{6}\\
&\qquad - \frac{1}{12}e^{\alpha-1}\bigl(\operatorname{Ei}(1)-\operatorname{Ei}(1-\alpha)\bigr).
\end{aligned}$$

Next, (9.25) is solved by, using (9.26),

$$\begin{aligned}
f_{zz}(\alpha) &= 2f_{zy}(\alpha) + 2\int_0^{\alpha}\Bigl(f_z(t) - t\,\mathbb{E}\bigl(D^{\mathsf{u}}_t\bigr)^2 - \Bigl(1+\frac{1}{(1-t)^3}\Bigr)f_z(t)\Bigr)\,dt\\
&= \frac{1}{2(1-\alpha)^2} - \frac{7}{6(1-\alpha)} + 1 + \frac{2}{3}\alpha - \frac{e^{\alpha}}{3} - \frac{1}{6}e^{\alpha-1}\bigl(\operatorname{Ei}(1)-\operatorname{Ei}(1-\alpha)\bigr).
\end{aligned}$$

Finally, (9.3) and (9.4) follow from (9.21) and (9.2). $\square$

Remark 9.2. The reader who wants to make a detailed comparison with [23] should note that, by (9.6) and the definitions in [23], their

$$\begin{aligned}
\grave{c}_2(x,z) &= \frac{(1-x)e^{-x}}{x} z\tilde{\Psi}(z,1,x) = \frac{1}{x} zF(z,1,x),\\
\grave{g}_2(x,z) &= \frac{(1-x)e^{-x}}{x} z(\mathrm{D}_y+1)\tilde{\Psi}(z,1,x) = \frac{1}{x} z(\mathrm{D}_y+1)F(z,1,x),\\
\grave{h}_2(x) &= \mathrm{D}_z\,\grave{g}_2(x,1) = \frac{1}{x}(\mathrm{D}_z+1)(\mathrm{D}_y+1)F(1,1,x)\\
&= \frac{1}{x}\bigl(f_{zy}(x)+f_z(x)+f_y(x)+f(x)\bigr).
\end{aligned}$$

In principle, expressions for the probabilities $\mathbb{P}(D^{\mathsf{LC}}_\alpha = k)$ can be obtained from (9.1) by differentiating at $z=0$ (or making Taylor expansions). However, even for the simplest case $k=0$, we obtain a rather complicated formula. We leave it to the reader to find a simpler formula, and to treat $k\ge 1$.

Theorem 9.3. *For* $0<\alpha<1$,

$$p^{\mathsf{LC}}_\alpha(0) = \mathbb{P}(D^{\mathsf{LC}}_\alpha = 0) = \psi^{\mathsf{LC}}_\alpha(0) = \frac{1-\alpha}{\alpha}\int_0^{\alpha}\frac{(e^{\alpha-t}-1)e^{\alpha-t}(1-t)}{e^{\alpha-te^{\alpha-t}}-1}\,dt. \tag{9.28}$$

Proof. Taking $z=0$ in Theorem 9.1 we have $u(\alpha,\beta,0)=e^{\beta-\alpha}$ and $v(\alpha,\beta,0)=\beta e^{-\alpha}$. Hence (9.1) yields

$$\psi^{\mathsf{LC}}_\alpha(0) = \frac{1-\alpha}{\alpha}\int_0^{\alpha}\frac{e^{\beta-\alpha}}{T(\beta e^{-\alpha})}\,\frac{\beta - T(\beta e^{-\alpha})}{1-e^{\beta-\alpha}}\,d\beta.$$

The change of variables $t=T(\beta e^{-\alpha})$ yields $\beta = te^{\alpha-t}$ and $d\beta/dt = (1-t)e^{\alpha-t}$, and (9.28) follows by simple calculations. $\square$

Various substitutions are possible, but none seems to give a simpler integral.

10. Some numerical values

Numerical values of $p_\alpha^{\mathsf{FC}}(k)$ and $p_\alpha^{\mathsf{RH}}(k)$ for given α and k are easily computed from the formulas in Theorems 7.1 and 8.1, see Remark 8.2. As examples we give some values (computed by `Maple`) in Table 1. The last line gives the sum for all $k > 10$.

Note that the fact that FC gives higher probabilities than RH for zero displacement hardly is an advantage, as might be believed. Since the average displacements are the same, this is compensated by higher probability for large displacements, which is worse.

k	$p_{0.5}^{\mathsf{FC}}(k)$	$p_{0.5}^{\mathsf{RH}}(k)$	$p_{0.9}^{\mathsf{FC}}(k)$	$p_{0.9}^{\mathsf{RH}}(k)$
0	0.750	0.649	0.550	0.162
1	0.143	0.245	0.143	0.153
2	0.051	0.076	0.069	0.128
3	0.024	0.022	0.042	0.104
4	0.012	0.0062	0.029	0.085
5	0.0070	0.0017	0.021	0.069
6	0.0042	0.00050	0.016	0.056
7	0.0026	0.00014	0.013	0.046
8	0.0017	0.000041	0.011	0.037
9	0.0011	0.000011	0.009	0.031
10	0.0007	0.000003	0.008	0.024
≥ 11	0.0018	0.000001	0.090	0.106

Table 1. Some numerical values

In principle, $p_\alpha^{\mathsf{LC}}(k)$ can be calculated similarly from (9.1), but we have only done so for $k = 0$. We find, for example, by Theorem 9.3 and numerical integration, $p_{0.5}^{\mathsf{LC}}(0) \doteq 0.686$ and $p_{0.9}^{\mathsf{LC}}(0) \doteq 0.212$.

11. Monotonicity properties of the limit distributions

For FC, it follows immediately from (7.1) that the probabilities $p_\alpha^{\mathsf{FC}}(k)$ decrease, $p_\alpha^{\mathsf{FC}}(0) > p_\alpha^{\mathsf{FC}}(1) > \dots$. Moreover, we have monotonicity for finite m and n too; it follows from (7.6) that for each m and n, the probabilities $\mathbb{P}\bigl(d_i^{\mathsf{FC}}(\mathcal{T}_{m,n}) = k\bigr)$ are non-increasing in k.

This has the practical consequence that searching in a hash table constructed by the FC rule is best done in the standard way, probing at h, $h+1$, $h+2$, and so on, where h is the hash address of the searched item.

It has been suggested that for RH and LC hashing, where the variances of the individual displacements are smaller, the displacements might be concentrated about their mean $\mathbb{E}\, d$ so that it would be more efficient to start probing at locations close to $h + \mathbb{E}\, d$. This seems to be the case for random probing and double hashing, see [3, 22]. However, we will see that this hardly is the case for linear probing. It would be the case (for a large table) for an insertion policy Ξ if $p_\alpha^{\Xi}(k)$ is larger when k is close to $\mathbb{E}\, d$ than when k is close to 0, so we study these probabilities.

First, for Robin Hood hashing we have the following precise result; we postpone the proof.

Theorem 11.1. *Let $\alpha_1 \doteq 0.931$ be the unique positive root of*

$$e^{2\alpha} - (2+\alpha)e^{\alpha} + 1 = 0$$

If $\alpha \in [0, \alpha_1]$, then $p_\alpha^{\mathsf{RH}}(0) \geq p_\alpha^{\mathsf{RH}}(1) > p_\alpha^{\mathsf{RH}}(2) > \dots$.
Conversely, if $\alpha \in (\alpha_1, 1)$, then $p_\alpha^{\mathsf{RH}}(0) < p_\alpha^{\mathsf{RH}}(1)$.

Hence, it is only for α close to 1 that a different probing sequence might be better. However, even for $\alpha > \alpha_1$, small displacements seems to be more likely than displacements close to $\mathbb{E}\,d$. One reason is that it follows easily from Theorem 8.1 that as $\alpha \to 1$, $(1-\alpha)D_\alpha^{\mathsf{RH}} \xrightarrow{\mathrm{d}} \mathrm{Exp}(1/2)$. Hence, for α close to 1, the distribution of D_α^{RH} is approximatively exponential. Although this does not imply corresponding asymptotics of individual probabilities, it implies that the average of the values of $p_\alpha^{\mathsf{RH}}(k)$ for k in a suitable interval close to $\mathbb{E}\,d$ is about e^{-1} times the average of the values for k in an interval close to 0. This suggests that the standard probing sequence is close to optimal when α is close to 1, when the differences between different methods ought to be greatest.

Nevertheless, it is interesting to ask for the mode of the distribution of D_α^{RH}, i.e. the value of k that maximizes $p_\alpha^{\mathsf{RH}}(k)$. Theorem 11.1 shows that the mode is 0 for $\alpha < \alpha_1$, but not for larger α.

Numerical calculations with `Maple` suggest the following; we have, however, no rigorous proof.

Conjecture 11.2. *The mode of D_α^{RH} is*

$$\begin{cases} 0, & 0 < \alpha < \alpha_1 \doteq 0.9308 \\ 1, & \alpha_1 < \alpha < \alpha_2 \doteq 0.9888 \\ 2, & \alpha_2 < \alpha < \alpha_3 \doteq 0.9989 \\ 3, & \alpha_3 < \alpha < \alpha_4 \doteq 0.999995 \\ 4, & \alpha_4 < \alpha < 1. \end{cases}$$

Of course, these results have merely theoretical interest; hashing with linear probing with $n/m > .93$ should probably be avoided. (In particular, this applies to the last case; if $\alpha > \alpha_4$, then the expectation of the displacement is over 100000.) Moreover, the differences are rather minor in this range; according to numerical calculations, no probability is ever more than 16.4% higher than $p_\alpha^{\mathsf{RH}}(0)$, and the absolute differences $p_\alpha^{\mathsf{RH}}(k) - p_\alpha^{\mathsf{RH}}(0)$ are less than 0.004. Hence, even when another probing sequence is better, the difference in performance seems to be small, and probably out-weighted by the additional steps required in the program.

It is somewhat surprising that the conjectured sequence of modes stops at 4. However, we have a another asymptotic result explaining this.

Theorem 11.3. *As $\alpha \to 1$, $(1-\alpha)^{-1}p_\alpha^{\mathsf{RH}}(k) \to L_{k+1}$ for every $k \geq 0$, where L_k have the generating function*

$$\sum_{k=1}^{\infty} L_k z^k = z\frac{1-e^{z-1}}{e^{z-1}-z}.$$

Proof. Theorem 8.1 implies that as $\alpha \to 1$, at least for $|z| \leq 1$,

$$(1-\alpha)^{-1}\psi_\alpha^{\mathsf{RH}}(z) \to \frac{e^{1-z}-1}{1-ze^{1-z}} = \frac{1-e^{z-1}}{e^{z-1}-z} = \sum_{k=0}^{\infty} L_{k+1}z^k,$$

and the result follows. □

This generating function equals the one in Knuth [15, 5.1.3-(25)], and thus the limits L_k coincide with the numbers L_k defined there as the average lengths of the successive increasing runs in a random sequence. As shown in Knuth [15, Section 5.1.3], these numbers L_k converge rapidly to 2 as $k \to \infty$, but the convergence is not monotone; the smallest is $L_1 = e - 1 \doteq 1.718$ and the largest is $L_5 \doteq 2.00006$. This and Theorem 11.3 strongly suggest (although do not strictly prove) that for α close to 1, the largest value of $p_\alpha^{\mathsf{RH}}(k)$ is for $k+1 = 5$, i.e. $k = 4$, as asserted in Conjecture 11.2.

Proof of Theorem 11.1. By (4.1), we can study $\mathbb{E}\, n_k^{\mathsf{RH}}(\mathcal{T}_\alpha)$ instead of $p_\alpha^{\mathsf{RH}}(k)$.

Let $k \geq 0$. By Lemmas 2.1 and 3.4, $\mathbb{E}\, n_k^{\mathsf{RH}}(\mathcal{T}_\alpha)$ and $\mathbb{E}\, n_{k+1}^{\mathsf{RH}}(\mathcal{T}_\alpha)$ are the expected number of visits of the random walk S_i to k and $k+1$, respectively. We compare these numbers by studying the excursions of $\{S_i\}$ above $k-1$. Since the only negative step is -1, every such excursion ends at k. At each visit to k, the probability of exiting is $e^{-\alpha}$, and if the random walk does not exit, it will sooner or later (perhaps immediately) return to k. Consequently, the number of visits to k in an excursion has a geometric distribution with expectation $1/e^{-\alpha} = e^\alpha$.

Furthermore, at each visit to k, the probability of going to $k+1$ or higher is $\mathbb{P}\big(\mathrm{Po}(\alpha) \geq 2\big) = 1 - e^{-\alpha} - \alpha e^{-\alpha}$; each such step leads, by the argument above, to an average of e^α visits to $k+1$ before the first return to k. In an excursion beginning at k, there are thus on the average

$$e^\alpha\big(1-e^{-\alpha}-\alpha e^{-\alpha}\big)e^\alpha = e^\alpha\big(e^\alpha - 1 - \alpha\big)$$

visits to $k+1$.

In an excursion beginning at $l \geq k+1$, there are on the average an additional e^α visits to $k+1$ before the first visit to k, and thus in total $e^\alpha\big(e^\alpha-\alpha\big)$ visits to $k+1$.

The average number of excursions above $k-1$ starting at $l \geq k$ is

$$\mathbb{P}(\xi_0 = l+1) + \sum_{j=0}^{k-1} \mathbb{E}\, n_j^{\mathsf{RH}}(\mathcal{T}_\alpha)\, \mathbb{P}(\xi_1 = l-j+1). \tag{11.1}$$

For $k=0$, we thus find, in agreement with (4.1) and (8.4), (8.5),

$$\mathbb{E}\, n_0^{\mathsf{RH}}(\mathcal{T}_\alpha) = \mathbb{P}(\xi_0 \geq 1)e^\alpha = e^\alpha - 1, \tag{11.2}$$

$$\mathbb{E}\, n_1^{\mathsf{RH}}(\mathcal{T}_\alpha) = \mathbb{P}(\xi_0 = 1)e^\alpha(e^\alpha - 1 - \alpha) + \mathbb{P}(\xi_0 \geq 2)e^\alpha(e^\alpha - \alpha) = e^\alpha(e^\alpha - 1 - \alpha), \tag{11.3}$$

and hence

$$\mathbb{E}\, n_0^{\mathsf{RH}}(\mathcal{T}_\alpha) - \mathbb{E}\, n_1^{\mathsf{RH}}(\mathcal{T}_\alpha) = \mathbb{P}(\xi_0 = 1)A - \mathbb{P}(\xi_0 \geq 2)B,$$

with $A := e^\alpha(2+\alpha-e^\alpha) > 0$ and $B := e^\alpha(e^\alpha - 1 - \alpha) > 0$. Consequently, $\mathbb{E}\, n_0^{\mathsf{RH}}(\mathcal{T}_\alpha) \geq \mathbb{E}\, n_1^{\mathsf{RH}}(\mathcal{T}_\alpha)$ if and only if $\mathbb{P}(\xi_0 \geq 2)/\,\mathbb{P}(\xi_0 = 1) \leq A/B$.

Similarly, for $k \geq 1$, using (11.1),

$$\begin{aligned}&\mathbb{E}\, n_k^{\mathsf{RH}}(\mathcal{T}_\alpha) - \mathbb{E}\, n_{k+1}^{\mathsf{RH}}(\mathcal{T}_\alpha)\\ &\quad= \sum_{j=0}^{k-1}\bigl(\mathbb{E}\, n_j^{\mathsf{RH}}(\mathcal{T}_\alpha) + \mathbf{1}[j=0]\bigr)\bigl(\mathbb{P}(\xi_0 = k-j+1)A - \mathbb{P}(\xi_0 \geq k-j+2)B\bigr).\end{aligned} \tag{11.4}$$

For a Poisson distributed variable ξ, $\mathbb{P}(\xi = j+1)/\,\mathbb{P}(\xi = j)$ is a decreasing function of j, and thus so is $\mathbb{P}(\xi \geq j+1)/\,\mathbb{P}(\xi = j)$. Consequently, if $\mathbb{E}\, n_0^{\mathsf{RH}}(\mathcal{T}_\alpha) \geq \mathbb{E}\, n_1^{\mathsf{RH}}(\mathcal{T}_\alpha)$, then for each $j \geq 2$

$$\mathbb{P}(\xi \geq j+1)/\,\mathbb{P}(\xi = j) < \mathbb{P}(\xi \geq 2)/\,\mathbb{P}(\xi = 1) \leq A/B,$$

and thus, by (11.4), $\mathbb{E}\, n_k^{\mathsf{RH}}(\mathcal{T}_\alpha) - \mathbb{E}\, n_{k+1}^{\mathsf{RH}}(\mathcal{T}_\alpha) > 0$ for every $k \geq 1$.

Finally, by (11.2) and (11.3),

$$\mathbb{E}\, n_1^{\mathsf{RH}}(\mathcal{T}_\alpha) - \mathbb{E}\, n_0^{\mathsf{RH}}(\mathcal{T}_\alpha) = e^{2\alpha} - (2+\alpha)e^\alpha + 1.$$

Denote this function by $f(\alpha)$. It is easily seen that $f(0) = 0$, $f'(0) < 0$, $f(1) > 0$ and $f''(\alpha) > 0$ for $\alpha > 0$. Hence f is a convex function on $[0,1]$ with exactly one zero $\alpha_1 > 0$, and $f(\alpha) < 0$ for $0 < \alpha < \alpha_1$ while $f(\alpha) > 0$ for $\alpha_1 < \alpha < 1$. The result follows. $\square$

For LC hashing, we do not have any similar exact results, in view of the complicated expressions in Theorem 9.1. However, as will be shown in a sequel to this paper, as $\alpha \to 1$, the difference between RH and LC becomes negligible; in particular, just as for RH, $(1-\alpha)D_\alpha^{\mathsf{LC}} \xrightarrow{\mathrm{d}} \operatorname{Exp}(1/2)$. This suggests that the standard probing sequence is almost optimal for LC hashing too. Moreover, it follows from Theorem 9.3 that as $\alpha \to 1$,

$$(1-\alpha)^{-1}p_\alpha^{\mathsf{LC}}(0) \to \int_0^1 \frac{(e^{1-t}-1)e^{1-t}(1-t)}{e^{1-te^{1-t}}-1}\,dt \doteq 2.647.$$

This limit is larger than the corresponding limit $L_1 = e-1$ for RH found above. Hence, for α close to 1 at least, $p_\alpha^{\mathsf{LC}}(0) > p_\alpha^{\mathsf{RH}}(0)$, which suggests that 0 may be the mode of D_α^{LC} too for a wide range of α. It would be interesting to have any results on the mode of D_α^{LC}. We conjecture, in contrast to RH, that it is 0 for all α.

References

[1] P. Billingsley, *Convergence of Probability Measures*. John Wiley & Sons, New York, 1968.

[2] É. Borel, Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'une infinité de coefficients. Application au problème de l'attente à un guichet. *C. R. Acad. Sci. Paris* **214** (1942), 452–456.

[3] S. Carlsson, J.I. Munro & P.V. Poblete, On linear probing hashing. Unpublished manuscript, 1987.

[4] P. Celis, P.-Å. Larson, J.I. Munro, Robin Hood Hashing (Preliminary Report). *FOCS* **26** (1985), 281-288.

[5] P. Chassaing & G. Louchard, Phase transition for parking blocks, Brownian excursion and coalescence. *Rand. Struct. Alg.* **21** (2002), no. 1, 76–119.

[6] M. Dwass, The total progeny in a branching process and a related random walk. *J. Appl. Probab.* **6** (1969), 682–686.

[7] P. Flajolet, P. Poblete & A. Viola, On the analysis of linear probing hashing. *Algorithmica* **22** (1998), no. 4, 490–515.

[8] R.L. Graham, D.E. Knuth & O. Patashnik, *Concrete Mathematics*. 2nd ed., Addison–Wesley, Reading, Mass., 1994.

[9] S. Janson, Moment convergence in conditional limit theorems. *J. Appl. Probab.* **38** (2001), no. 2, 421–437.

[10] S. Janson, Asymptotic distribution for the cost of linear probing hashing. *Random Struct. Alg.* **19** (2001), no. 3–4, 438–471.

[11] J.H.B. Kemperman, *The General One-Dimensional Random Walk with Absorbing Barriers with Applications to Sequential Analysis*. Thesis. Excelsiors Foto-Offset, The Hague, 1950.

[12] J.H.B. Kemperman, *The Passage Problem for a Stationary Markov Chain*. University of Chicago Press, Chicago, Ill., 1961.

[13] D.G. Kendall, Some problems in the theory of queues. *J. Roy. Statist. Soc. Ser. B.* **13** (1951), 151–185.

[14] D.E. Knuth, Notes on "open" addressing. Unpublished notes, 1963. Available at `http://www.wits.ac.za/helmut/first.ps`

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

[16] D.E. Knuth, Linear probing and graphs. *Algorithmica* **22** (1998), no. 4, 561–568.

[17] V.F. Kolchin, *Random Mappings*. Nauka, Moscow, 1984 (Russian). English transl.: Optimization Software, New York, 1986.

[18] R. Otter, The multiplicative process. *Ann. Math. Statist.* **20** (1949), 206–224.

[19] Yu. L. Pavlov, The asymptotic distribution of maximum tree size in a random forest. *Teor. Verojatnost. i Primenen.* **22** (1977), no. 3, 523–533 (Russian). English transl.: *Th. Probab. Appl.* **22** (1977), no. 3, 509–520.

[20] Yu. L. Pavlov, *Random forests*. Karelian Centre Russian Acad. Sci., Petrozavodsk, 1996 (Russian). English transl.: VSP, Zeist, The Netherlands, 2000.

[21] J. Pitman, Enumerations of trees and forests related to branching processes and random walks. *Microsurveys in discrete probability (Princeton, NJ, 1997)*, 163–180, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 41, Amer. Math. Soc., Providence, RI, 1998.

[22] P.V. Poblete & J.I. Munro, Last-come-first-served hashing. *J. Algorithms* **10** (1989), no. 2, 228–248.

[23] P.V. Poblete, A. Viola & J.I. Munro, The Diagonal Poisson Transform and its application to the analysis of a hashing scheme. *Random Struct. Alg.* **10** (1997), no. 1-2, 221–255.

[24] P.V. Poblete, A. Viola & J.I. Munro, Analyzing the LCFS linear probing hashing algorithm with the help of Maple. *MapleTech* **4** (1997), no. 1, 8–13.

[25] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*. John Wiley & Sons, New York, 1967.

[26] L. Takács, Ballots, queues and random graphs. *J. Appl. Probab.* **26** (1989), no. 1, 103–112.

[27] J.C. Tanner, A derivation of the Borel distribution. *Biometrika* **48** (1961), 222–224.

[28] A. Viola, Analysis of hashing algorithms and a new mathematical transform. Ph.D. thesis. Technical report CS-95-50, Computer Science Department, University of Waterloo, Waterloo, ON, Canada, November 1995.

[29] A. Viola, Exact distributions of individual displacements in linear probing hashing. *ACM Trans. Algorithms*, to appear.

Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden

E-mail address: `svante.janson@math.uu.se`

URL: `http://www.math.uu.se/~svante/`