INDIVIDUAL DISPLACEMENTS IN HASHING WITH COALESCED CHAINS

SVANTE JANSON

Abstract. We study the asymptotic distribution of the displacements in hashing with coalesced chains, for both late-insertion and early-insertion. Asymptotic formulas for means and variances follow. The method uses Poissonization and some stochastic calculus.

1. Introduction

The standard version of hashing with coalesced chains, due to Williams [10] can be described as follows, where n and m are integers with $0 \le n \le m$. (See further Knuth [6, Section 6.4, in particular Algorithm 6.4.C] and the monograph by Vitter and Chen [9].)

> A table with m cells $1, \dots, m$ is filled with $n \le m$ items $x_1, \dots, x_n$, by placing them sequentially using n integers $h_i \in \{1, \dots, m\}$. Each cell contains two fields, initially empty, one of which can hold an item and the other can link to another cell. Item x_i is inserted into cell h_i if it is empty; otherwise we follow the links from cell h_i until we reach the end of the chain (signalled by a null link), we add a link to an empty cell (which is chosen as the empty cell with largest index) and store the item there.

For our probabilistic treatment, we assume that each of the m^n possible hash sequences $(h_i)_1^n$ is equally likely; in other words, the hash addresses h_i are independent random numbers, uniformly distributed on $\{1, \dots, m\}$.

The *displacement* d_i of an item x_i is the number of links we have to follow from h_i until we find x_i. Large displacements make both insertion and searching less efficient, so it is desirable to keep the displacements small. (Two different but related quantities are used in other papers to measure the efficiency: *The number of probes* to find the item x_i in the table is $d_i + 1$. *The number of key comparisons* to find the item is also $d_i + 1$. This should be noted when comparing the results below with other papers.)

The items are thus arranged in linked chains in the hash table. If a new item hashes to an empty cell, a new chain with that single item is created. If a new item hashes to a cell in an existing chain, then that chain grows by addition of a formerly empty cell. It was shown by Chen and Vitter [3] and

Date: February 10, 2005; slightly revised January 18, 2006.
2000 *Mathematics Subject Classification.* 68P10 (60F05, 60G35, 68W40).

Knott [5] that the average performance could be improved by modifying the algorithm above, inserting the new item at a different place in an existing chain. We will therefore study two versions of hashing with coalesced chains:

- L *Late-insertion* (LISCH). The standard version described above where the new item is inserted last in its chain.
- E *Early-insertion* (EISCH) [3], [5], [9]. If cell h_i is occupied, item x_i is inserted into an empty cell as above, but this cell is linked into the chain immediately after h_i. (I.e., if the first free cell is j and the link from h_i points to k (null or not), then this link is reset to j, and the link field in j is set to k.) This method gives the smallest average displacement among all possible insertion schemes [9, Theorem 5.2].

Note that the insertion of a sequence of items results in the same set of occupied cells in both versions, and that this set is partitioned into chains in the same way, but that the order in the chains, and thus the individual displacements, may differ.

Our main result is Theorem 2.1 below (together with its refinement Theorem 2.5), which gives the asymptotic distribution of the displacements in a random hash table under both insertion methods; we consider also the case of unsuccessful searches. As corollaries we easily find earlier known asymptotic formulas for the average displacements and for the variances of them (some of the latter may be new). These asymptotic distributions are studied further in Section 3; some numerical values are given in Table 1. The proofs are given in Section 4. They are based on Poissonization, regarding the items as arriving at random times.

Remark 1.1. An interesting variation of the algorithm above [6, Exercise 6.4-43], discussed in detail by Vitter and Chen [9], is to choose a number $m_1 < m$ and reserve the last $m - m_1$ cells as a "cellar" for the undisturbed growth of the chains. We then assume that the hash addresses $h_i \in \{1, \dots, m_1\}$, and use exactly the same algorithms as above. (These versions are called LICH and EICH in [9].) It is shown in [9] that for given n and m, a suitable choice of m_1 will improve the average performance.

In this setting, it is also interesting to consider a third version *varied-insertion* (VICH) [9], which behaves like E except that when the chain from the hash address contains a cellar cell, the new item is inserted after the last cellar cell. It is shown in [9, Chapter 5] that this method gives the minimum average among all insertion methods satisfying a weak assumption.

We have not yet investigated the versions with cellar in detail, but it seems that our methods could be used with some additional work to find the asymptotic distributions of the displacements in these cases too. (The means are given in [9].)

Remark 1.2. Corresponding results for hashing with linear probing are given by Janson [4] and Viola [8]. Note that, as remarked in [6], the average displacement for linear probing tends to infinity if $n, m \to \infty$ with $n/m \to 1$, while for the chained hashing studied here, it stays bounded also in the extreme case $n = m$ of a full table.

Acknowledgements. This research was begun on the transatlantic flight to the Knuthfest at Stanford in honour of Donald Knuth's 64th birthday in January 2002, where some of the results were presented, and continued on another flight to the Analysis of Algorithms workshop at MSRI, Berkeley, June 2004; I thank KLM for providing long and quiet periods of work. I further thank Donald Knuth for helpful remarks.

2. Notation and results

By a hash table $\mathcal{T}$ we mean not only the final table, but also its construction history; moreover, we consider the two version above 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.

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 $\{1, \dots, m\}$.

We denote the two insertion policies defined in the introduction by L and E, and use Ξ to denote any of these.

Given a hash table $\mathcal{T}$ (with m cells and n items), random or not, and a policy $\Xi \in \{\mathsf{L}, \mathsf{E}\}$, 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

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

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

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

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. (Note that in an unsuccessful search starting at j, the number of key comparisons equals d_j^{U}, while the number of probes, $\widetilde{d}_j^{\mathsf{U}}$ say, is d_j^{U} if $d_j^{\mathsf{U}} \ge 1$ and 1 if $d_j^{\mathsf{U}} = 0$; i.e., $\widetilde{d}_j^{\mathsf{U}} := \max(d_j^{\mathsf{U}}, 1)$.) 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, cf. (2.2), given $\mathcal{T}$, $d^{\mathsf{U}}(\mathcal{T})$ has the distribution

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

Our main result is the following theorem, giving the asymptotic distribution of the displacement of a random item in a random hash table $\mathcal{T}_{m,n}$, together with the corresponding quantity for an unsuccessful search.

Theorem 2.1. *Suppose that $m, n \to \infty$ with $0 < n \le m$ and $n/m \to \alpha \in [0,1]$. Then, for every $k = 0, 1, \dots,$*
(i)

$$\mathbb{P}\big(d^{\mathsf{U}}(\mathcal{T}_{m,n}) = k\big) = \frac{1}{m}\,\mathbb{E}\big(n_k^{\mathsf{U}}(\mathcal{T}_{m,n})\big) \to p_\alpha^{\mathsf{U}}(k) := \begin{cases} 1-\alpha, & k=0,\\ \int_0^\alpha (1-\alpha+t)\big(1-e^{-t}\big)^{k-1}\,dt, & k \ge 1; \end{cases}$$

(ii)

$$\mathbb{P}\big(d^{\mathsf{L}}(\mathcal{T}_{m,n}) = k\big) = \frac{1}{n}\,\mathbb{E}\big(n_k^{\mathsf{L}}(\mathcal{T}_{m,n})\big) \to p_\alpha^{\mathsf{L}}(k) := \begin{cases} 1-\alpha/2, & k=0,\\ \frac{1}{\alpha}\int_0^\alpha \big(\alpha-t-(\alpha-t)^2/2\big)\big(1-e^{-t}\big)^{k-1}\,dt, & k \ge 1; \end{cases}$$

(iii)

$$\mathbb{P}\big(d^{\mathsf{E}}(\mathcal{T}_{m,n}) = k\big) = \frac{1}{n}\,\mathbb{E}\big(n_k^{\mathsf{E}}(\mathcal{T}_{m,n})\big) \to p_\alpha^{\mathsf{E}}(k) := \begin{cases} 1-\alpha/2, & k=0,\\ \frac{1}{\alpha}\int_0^\alpha (\alpha-t)e^{-t}\big(1-e^{-t}\big)^{k-1}\,dt, & k \ge 1. \end{cases}$$

(For $\alpha = 0$, $p_0^{\mathsf{L}}(k) = p_0^{\mathsf{E}}(k) = 0$ when $k \ge 1$.) For every $\Xi \in \{\mathsf{L}, \mathsf{E}, \mathsf{U}\}$ and $\alpha \in [0,1]$, $\{p_\alpha^\Xi(k)\}_{k=0}^\infty$ is a probability distribution on $\mathbb{N}$. If D_α^Ξ is a random variable with this distribution, i.e. $\mathbb{P}(D_\alpha^\Xi = k) = p_\alpha^\Xi(k)$, then these results can be written

$$d^\Xi(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{d}} D_\alpha^\Xi. \tag{2.4}$$

Moreover, all moments converge in (2.4), i.e., $\mathbb{E}\big(d^\Xi(\mathcal{T}_{m,n})\big)^r \to \mathbb{E}\big(D_\alpha^\Xi\big)^r$ for every $r \ge 0$.

It follows immediately that for the number of probes in an unsuccessful search, we have

$$\widetilde{d}^{\mathsf{U}}(\mathcal{T}_{m,n}) \xrightarrow{\mathrm{d}} \widetilde{D}_\alpha^{\mathsf{U}} := \max(D_\alpha^{\mathsf{U}}, 1), \tag{2.5}$$

again with convergence of all moments.

As a corollary, we find the asymptotics for the expectations; these have earlier been derived, together with exact formulas for $\mathbb{E}\, d^{\Xi}(\mathcal{T}_{m,n})$, by Knuth [6] and Vitter and Chen [9] (in equivalent forms for the number of probes or key comparisons).

Corollary 2.2. *Suppose that $m, n \to \infty$ with $0 < n \le m$ and $n/m \to \alpha \in [0, 1]$. Then*

$$\begin{aligned}\mathbb{E}(d^{\mathsf{U}}(\mathcal{T}_{m,n})) &\to \mathbb{E}\, D^{\mathsf{U}}_{\alpha} = \frac{1}{4}\big(e^{2\alpha} - 1\big) + \frac{\alpha}{2},\\ \mathbb{E}(d^{\mathsf{L}}(\mathcal{T}_{m,n})) &\to \mathbb{E}\, D^{\mathsf{L}}_{\alpha} = \frac{1}{8\alpha}\big(e^{2\alpha} - 1\big) + \frac{\alpha}{4} - \frac{1}{4},\\ \mathbb{E}(d^{\mathsf{E}}(\mathcal{T}_{m,n})) &\to \mathbb{E}\, D^{\mathsf{E}}_{\alpha} = \frac{1}{\alpha}\big(e^{\alpha} - 1 - \alpha\big).\end{aligned}$$

Remark 2.3. Note that the expected number of probes is $\mathbb{E}(d^{\mathsf{L}}(\mathcal{T}_{m,n})) + 1$ or $\mathbb{E}(d^{\mathsf{E}}(\mathcal{T}_{m,n})) + 1$ for a successful search, and, cf. (2.5),

$$\mathbb{E}(\widetilde{d}^{\mathsf{U}}(\mathcal{T}_{m,n})) = \mathbb{E}(d^{\mathsf{U}}(\mathcal{T}_{m,n})) + \frac{m-n}{m} \to \mathbb{E}(\widetilde{D}^{\mathsf{U}}_{\alpha}) = \mathbb{E}(D^{\mathsf{U}}_{\alpha}) + 1 - \alpha$$

for an unsuccessful search.

Theorem 2.1 similarly yields asymptotic formulas for higher moments too; in particular we have the following results for the variance.

Corollary 2.4. *Suppose that $m, n \to \infty$ with $0 < n \le m$ and $n/m \to \alpha \in [0, 1]$. Then*

$$\begin{aligned}\operatorname{Var}(d^{\mathsf{U}}(\mathcal{T}_{m,n})) &\to \operatorname{Var} D^{\mathsf{U}}_{\alpha}\\ &= -\tfrac{1}{16}e^{4\alpha} + \tfrac{4}{9}e^{3\alpha} - \big(\tfrac14\alpha + \tfrac18\big)e^{2\alpha} - \tfrac14\alpha^2 + \tfrac{5}{12}\alpha - \tfrac{37}{144},\\ \operatorname{Var}(d^{\mathsf{L}}(\mathcal{T}_{m,n})) &\to \operatorname{Var} D^{\mathsf{L}}_{\alpha} = -\tfrac{1}{64}\Big(\frac{e^{2a}-1}{a}\Big)^2 + \frac{64e^{2a} + 37e^{a} + 37}{432}\cdot\frac{e^{a}-1}{a}\\ &\qquad - \tfrac{1}{16}e^{2a} - \tfrac{1}{16}a^2 + \tfrac{5}{24}a - \tfrac{7}{36},\\ \operatorname{Var}(d^{\mathsf{E}}(\mathcal{T}_{m,n})) &\to \operatorname{Var} D^{\mathsf{E}}_{\alpha} = \frac{\alpha-2}{2}\Big(\frac{e^{\alpha}-1}{\alpha}\Big)^2 + 2\frac{e^{\alpha}-1}{\alpha} - 1,\end{aligned}$$

and, for the number of probes in an unsuccessful search,

$$\begin{aligned}\operatorname{Var}(\widetilde{d}^{\mathsf{U}}(\mathcal{T}_{m,n})) &\to \operatorname{Var}(\widetilde{D}^{\mathsf{U}}_{\alpha})\\ &= -\tfrac{1}{16}e^{4\alpha} + \tfrac{4}{9}e^{3\alpha} + \big(\tfrac14\alpha - \tfrac58\big)e^{2\alpha} - \tfrac14\alpha^2 - \tfrac{1}{12}\alpha + \tfrac{35}{144}.\end{aligned}$$

The asymptotic formula for $\operatorname{Var}(\widetilde{d}^{\mathsf{U}}(\mathcal{T}_{m,n}))$, together with an exact formula, is given in Knuth [6, Answer 6.4-40] and in Vitter and Chen [9]; the corresponding results for $\operatorname{Var}(d^{\mathsf{U}}(\mathcal{T}_{m,n}))$ follow easily. (The numerical result in [6, Answer 6.4-40] and [9] for the case $\alpha = 1$, when $\widetilde{D}^{\mathsf{U}}_{\alpha} = D^{\mathsf{U}}_{\alpha}$, should be 2.65.) The asymptotics of $\operatorname{Var}(d^{\mathsf{L}}(\mathcal{T}_{m,n}))$ are given in [9]. We do not know whether the asymptotic of $\operatorname{Var}(d^{\mathsf{E}}(\mathcal{T}_{m,n}))$ have been published earlier.

Consider a computer program where a large hash table is constructed once, and then used many times for finding the items. We assume that each item in the table is equally likely to be requested, and that each choice is independent of the previous ones. We therefore have two levels of randomness: First we construct a random hash table $\mathcal{T}$ with some displacements (d_i). Keeping $\mathcal{T}$ fixed and choosing a random index $I \in \{1, \dots, n\}$, we obtain the random displacement $d(\mathcal{T}) = d_I$. As the program runs with many searches in the hash table, the search times then are (functions of) 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. Note that this distribution depends on the hash table $\mathcal{T}$, which 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 $\operatorname{Var}(d(\mathcal{T})|\mathcal{T}) = \mathbb{E}(d(\mathcal{T})^2|\mathcal{T}) - \mathbb{E}(d(\mathcal{T})|\mathcal{T})^2$ are random variables. In other words, we study the conditional distribution of $d(\mathcal{T})$ given $\mathcal{T}$.

We can refine the results above by conditioning on $\mathcal{T}_{m,n}$. The following theorem says that we still have the same limits, now with convergence in probability. In other words, different realizations of $\mathcal{T}_{m,n}$ have (with large probability) almost the same distribution of the displacements, so a typical instance of the random hash table $\mathcal{T}_{m,n}$ has its displacements distributed as the average studied in Theorem 2.1.

Theorem 2.5. *Suppose that $m, n \to \infty$ with $0 < n \le m$ and $n/m \to \alpha \in [0, 1]$. Then, for every $k = 0, 1, \dots,$ with $p^{\Xi}_{\alpha}(k)$ defined in Theorem 2.1,*

(i) $\mathbb{P}\big(d^{\mathsf{U}}(\mathcal{T}_{m,n}) = k \bigm| \mathcal{T}_{m,n}\big) = m^{-1} n^{\mathsf{U}}_k(\mathcal{T}_{m,n}) \xrightarrow{\text{p}} p^{\mathsf{U}}_{\alpha}(k),$

(ii) $\mathbb{P}\big(d^{\mathsf{L}}(\mathcal{T}_{m,n}) = k \bigm| \mathcal{T}_{m,n}\big) = n^{-1} n^{\mathsf{L}}_k(\mathcal{T}_{m,n}) \xrightarrow{\text{p}} p^{\mathsf{L}}_{\alpha}(k),$

(iii) $\mathbb{P}\big(d^{\mathsf{E}}(\mathcal{T}_{m,n}) = k \bigm| \mathcal{T}_{m,n}\big) = n^{-1} n^{\mathsf{E}}_k(\mathcal{T}_{m,n}) \xrightarrow{\text{p}} p^{\mathsf{E}}_{\alpha}(k).$

Remark 2.6. A more fancy formulation of Theorem 2.5 is that the distribution of $d^{\Xi}(\mathcal{T}_{m,n})$ converges to p^{Ξ}_{α} 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 [2] for definitions.

Moment convergence holds in Theorem 2.5 too, i.e. conditioned on $\mathcal{T}_{m,n}$.

Theorem 2.7. *Suppose that $m, n \to \infty$ with $0 < n \le m$ and $n/m \to \alpha \in [0, 1]$. Then, for every $r \ge 0$ and $\Xi \in \{\mathsf{L}, \mathsf{E}, \mathsf{U}\}$,*

$$\mathbb{E}\big(d^{\Xi}(\mathcal{T}_{m,n})^r \mid \mathcal{T}_{m,n}\big) \xrightarrow{\text{p}} \mathbb{E}\big(D^{\Xi}_{\alpha}\big)^r.$$

In particular, the conditional mean and variance, given the hash table, converge in probability to the limits in Corollaries 2.2 and 2.4.

3. The asymptotic distributions

We give some further results on the probability distributions $p^{\Xi}_{\alpha}(k)$ defined in Theorem 2.1; we assume $\alpha > 0$. We omit the proofs. (Several of the results below were obtained with the help of `Maple`.)

It follows directly from the definitions in Theorem 2.1 that

$$p_\alpha^{\Xi}(k) = O\bigl(1-e^{-\alpha}\bigr)^k = O\bigl(1-e^{-1}\bigr)^k;$$

hence the probabilities decrease geometrically. More refined asymptotics can easily be derived (we omit the details); we have, as $k \to \infty$, for $\alpha > 0$,

$$\begin{aligned} p_\alpha^{\mathsf{U}}(k) &\sim e^{\alpha} k^{-1}\bigl(1-e^{-\alpha}\bigr)^k, \\ p_\alpha^{\mathsf{U}}(k) &\sim \alpha^{-1} e^{\alpha}(e^{\alpha}-1) k^{-2}\bigl(1-e^{-\alpha}\bigr)^k, \\ p_\alpha^{\mathsf{U}}(k) &\sim \alpha^{-1}(e^{\alpha}-1) k^{-2}\bigl(1-e^{-\alpha}\bigr)^k. \end{aligned}$$

In particular, the probability of an extremely large displacement is about e^{α} as large for late-insertion as for early-insertion.

The probability generating functions for D_α^{Ξ} follow also easily from the formulas in Theorem 2.1:

$$\begin{aligned} \mathbb{E}\, x^{D_\alpha^{\mathsf{U}}} &= \sum_{k=0}^{\infty} p_\alpha^{\mathsf{U}}(k) x^k = 1-\alpha + x\int_0^{\alpha} \frac{1-\alpha+t}{1-x+xe^{-t}}\,dt, \\ \mathbb{E}\, x^{D_\alpha^{\mathsf{L}}} &= \sum_{k=0}^{\infty} p_\alpha^{\mathsf{L}}(k) x^k = 1-\frac{\alpha}{2} + \frac{x}{\alpha}\int_0^{\alpha} \frac{\alpha-t-(\alpha-t)^2/2}{1-x+xe^{-t}}\,dt, \\ \mathbb{E}\, x^{D_\alpha^{\mathsf{E}}} &= \sum_{k=0}^{\infty} p_\alpha^{\mathsf{E}}(k) x^k = 1-\frac{\alpha}{2} + \frac{x}{\alpha}\int_0^{\alpha} \frac{\alpha-t}{(1-x)e^{t}+x}\,dt. \end{aligned}$$

These integrals can be evaluated in terms of the dilog function (and for L also polylog), but we do not know any simple form. The generating functions are analytic for $|x| < r(\alpha) := \bigl(1-e^{-\alpha}\bigr)^{-1}$, with a singularity at $r(\alpha)$.

The integrals defining $p_\alpha^{\Xi}(k)$ are easily evaluated for small k. We find, for example,

$$\begin{aligned} p_\alpha^{\mathsf{U}}(1) &= \alpha - \tfrac{1}{2}\alpha^2, & p_\alpha^{\mathsf{U}}(2) &= 2e^{-a} - 2 + 2\alpha - \tfrac{1}{2}\alpha^2, \\ p_\alpha^{\mathsf{L}}(1) &= \tfrac{1}{2}\alpha - \tfrac{1}{6}\alpha^2, & p_\alpha^{\mathsf{L}}(2) &= 2\frac{1-e^{-\alpha}}{\alpha} - 2 + \alpha - \tfrac{1}{6}\alpha^2, \\ p_\alpha^{\mathsf{E}}(1) &= \frac{e^{-\alpha}-1+\alpha}{\alpha}, & p_\alpha^{\mathsf{E}}(2) &= \frac{1}{2} - \frac{1-e^{-\alpha}}{\alpha} + \frac{1-e^{-2\alpha}}{4\alpha}. \end{aligned}$$

No simple pattern is seen, and we leave further investigation to the reader.

Numerical values for $p_\alpha^{\Xi}(0), \dots, p_\alpha^{\Xi}(10)$, the tail $\sum_{11}^{\infty} p_\alpha^{\Xi}(k)$, the mean $\mathbb{E}\, D_\alpha^{\Xi}$ and the variance $\operatorname{Var} D_\alpha^{\Xi}$ are given for $\alpha = 0.5$ and 1 (half-full and full tables) in Table 1.

4. Proofs

To prove the theorems, we randomize the times the items are inserted in the table by Poissonization: We assume that items with hash address i arrive according to a Poisson process with intensity 1, the m different Poisson processes being independent. We let $\mathcal{T}(t)$ denote the hash table at

k	$p^{\mathsf{U}}_{0.5}(k)$	$p^{\mathsf{L}}_{0.5}(k)$	$p^{\mathsf{E}}_{0.5}(k)$	$p^{\mathsf{U}}_{1}(k)$	$p^{\mathsf{L}}_{1}(k)$	$p^{\mathsf{E}}_{1}(k)$
0	0.5	0.75	0.75	0.0	0.5	0.5
1	0.375	0.2083	0.2130	0.5	0.3333	0.3679
2	0.0881	0.0322	0.0291	0.2358	0.0976	0.0840
3	0.0252	0.0070	0.0059	0.1200	0.0376	0.0280
4	0.0078	0.0018	0.0014	0.0638	0.0163	0.0110
5	0.0026	0.00049	0.00038	0.0349	0.0076	0.0048
6	0.00086	0.00014	0.00011	0.0194	0.0037	0.0022
7	0.00030	0.000043	0.000032	0.0110	0.0019	0.0011
8	0.00010	0.000014	0.000010	0.0063	0.0010	0.0005
9	0.00004	0.000004	0.000003	0.0036	0.0005	0.0003
10	0.00001	0.000001	0.000001	0.0021	0.0003	0.0001
≥ 11	0.000007	0.0000007	0.0000005	0.0031	0.0003	0.0002
$\mathbb{E}$	0.6796	0.3046	0.2974	2.0973	0.7986	0.7183
Var	0.7394	0.3565	0.3324	2.6533	1.2799	0.9603

TABLE 1. Some numerical values

time t, when there are $\mathrm{Po}(t)$ items with each hash address. For simplicity, we write $n^{\Xi}_k(t) := n^{\Xi}_k(\mathcal{T}(t))$.

Combining the m individual Poisson processes, we see that the items $x_1, x_2, \dots$ arrive according to a Poisson process with intensity m; we call the arrival times $\tau_1, \tau_2, \dots$ (we may assume that these are distinct). We really have to stop at τ_m, since the table then is full, but it is convenient to think of the hashing as continuing for ever, with the chains growing into a virtual, infinitely large attic; no new chains are created after τ_m.

The hash addresses of the items $x_1, x_2, \dots$ are independent and uniformly distributed, so except for the random time scale, this is the situation we want to study. More precisely, $\mathcal{T}(\tau_n) = \mathcal{T}_{m,n}$ for $0 \leq n \leq m$.

Note that $\tau_m \approx 1$; more precisely, $\tau_m/m \xrightarrow{\mathrm{p}} 1$ as $m \to \infty$, as shown in Lemma 4.3 below.

We will consider stochastic processes defined on $[0, \infty)$ (although we mainly are interested in $0 \leq t \leq 1$). We say that such a process $X(t)$ is increasing if $X(s) \leq X(t)$ whenever $s \leq t$. We let $\xrightarrow{\mathrm{ucp}}$ denote convergence *uniformly on compacts in probability* (ucp), i.e. $X_n \xrightarrow{\mathrm{ucp}} X$ if $\sup_{0\leq t\leq u} |X_n(t) - X(t)| \xrightarrow{\mathrm{p}} 0$ for every $u > 0$.

Lemma 4.1. *Let, for each n, $X_n(t)$, $t \geq 0$, be an increasing, stochastic process, and let $f(t)$ be a continuous function on $[0, \infty)$. If $X_n(t) \xrightarrow{\mathrm{p}} f(t)$ for every $t \geq 0$, then $X_n(t) \xrightarrow{\mathrm{ucp}} f(t)$.*

Proof. Fix $u > 0$. Let $\varepsilon > 0$, and let K be so large that if $\delta := u/K$, then $|f(s) - f(t)| < \varepsilon$ if $s - t \leq \delta$ and $0 \leq s \leq t \leq u$. Since each X_n is increasing, the limit $f(t)$ is too. Hence, if $(k-1)\delta \le t \le k\delta$,

$$X_n((k-1)\delta) - f((k-1)\delta) - \varepsilon \le X_n((k-1)\delta) - f(k\delta) \\ \le X_n(t) - f(t) \le X_n(k\delta) - f((k-1)\delta) \le X_n(k\delta) - f(k\delta) + \varepsilon,$$

and, consequently,

$$\sup_{0\le t\le u} |X_n(t) - f(t)| \le \sup_{0\le k\le K} |X_n(k\delta) - f(k\delta)| + \varepsilon.$$

We know that $X_n(t) - f(t) \xrightarrow{\mathrm{p}} 0$ for every $t \ge 0$. We apply this for $t = k\delta$, $k = 0, \dots, K$, and find that **whp** (i.e., with probability $\to 1$ as $n \to \infty$) $|X_n(k\delta) - f(k\delta)| < \varepsilon$ for $k = 0, \dots, K$, and thus $\sup_{0\le t\le u} |X_n(t) - f(t)| < 2\varepsilon$. Since $\varepsilon > 0$ is arbitrary, $\sup_{0\le t\le u} |X_n(t) - f(t)| \xrightarrow{\mathrm{p}} 0$. □

Let $N(t)$ be the number of items that have arrived at time t. Since each item is put into some empty cell, the number of empty cells at time t is $(m - N(t))_+$, i.e. $m - N(t)$ for $t \le \tau_m$ and then 0.

Lemma 4.2. *As $m \to \infty$, $N(t)/m \xrightarrow{\mathrm{ucp}} t$ for $t \ge 0$.*

Proof. We have $N(t) \sim \mathrm{Po}(mt)$, and thus $N(t)/m \xrightarrow{\mathrm{p}} t$ as $m \to \infty$ for every $t \ge 0$. The convergence ucp follows from Lemma 4.1. □

Lemma 4.3. *Suppose that $m \to \infty$ and $n/m \to \alpha$, with $0 \le n \le m$. Then $\tau_n \xrightarrow{\mathrm{p}} \alpha$. Consequently, if $X_m \xrightarrow{\mathrm{ucp}} X$ for some stochastic processes X_m and X, where $X(t)$ is continuous, then $X_m(\tau_n) \xrightarrow{\mathrm{p}} X(\alpha)$.*

Proof. τ_n is the sum of n i.i.d. waiting times, each $\mathrm{Exp}(1/m)$, so $\mathbb{E}\,\tau_n = n/m \to \alpha$ and $\operatorname{Var}\tau_n = n/m^2 \to 0$, whence $\tau_n \xrightarrow{\mathrm{p}} \alpha$ by Chebyshev's inequality. (Alternatively, this is a standard consequence of Lemma 4.2: For $\varepsilon > 0$, $\mathbb{P}(\tau_n > \alpha + \varepsilon) \le \mathbb{P}\big(N(\alpha+\varepsilon)/m < n/m\big) \to 0$. Similarly, $\mathbb{P}(\tau_n < \alpha - \varepsilon) \to 0$.)

The final assertion follows because **whp** $\tau_n < \alpha + 1$, and then $|X_m(\tau_n) - X(\alpha)| \le \sup_{s\le\alpha+1} |X_m(s) - X(s)| + |X(\tau_n) - X(\alpha)| \xrightarrow{\mathrm{p}} 0$ □

Chains. When an item arrives to an empty cell, a new chain of length 1 is created. The chain then grows one unit each time it is hit. Hence each chain, once created, grows according to a birth process where the transition $\ell \to \ell + 1$ has intensity ℓ, and different chains grow independently. (In order for this to hold for $t > \tau_m$ too, we may pretend that new items arrive also in the attic, but are ignored unless they hit an existing chain. Similar *ad hoc* modifications have to be made after τ_m for other quantities too, in order for the arguments below to be valid; a simple possibility is to redefine $n_k^{\mathsf{U}}(t)$ for $t > \tau_m$ so that (4.8) holds and redefine the processes Z and W in (4.10) and (4.12) to be constant for $t \ge \tau_m$. The details do not matter, since we will later consider only $t = \tau_n \le \tau_m$, so we will ignore them.)

The growth process of each chain thus is the same as the *Yule process*, or *binary splitting*, a branching process where each individual after a lifetime distributed as Exp(1) splits into two. It is well-known, see e.g. [1, Section III.5], that if we start a Yule process with a single particle at time 0, the number of particles at time t has the geometric distribution $\mathrm{Ge}(e^{-t})$ with mean e^t and

$$\mathbb{P}(k \text{ particles}) = e^{-t}\bigl(1-e^{-t}\bigr)^{k-1}, \qquad k \geq 1. \tag{4.1}$$

Let $C(\mathcal{T})$ be the number of chains in the hash table $\mathcal{T}$, and denote their lengths by $L_1(\mathcal{T}), \dots, L_{C(\mathcal{T})}(\mathcal{T})$. For $\mathcal{T}(t)$ we write $C(t)$ and $L_j(t)$.

Lemma 4.4. *Let $C_\ell(t)$ be the number of chains of length ℓ in $\mathcal{T}(t)$. Then, for each k and $t \geq 0$, as $m \to \infty$,*

$$\frac{1}{m}\sum_{l\geq k} C_\ell(t) \xrightarrow{\text{ucp}} \int_0^{t\wedge 1} (1-s)\bigl(1-e^{-(t-s)}\bigr)^{k-1}\,ds. \tag{4.2}$$

Proof. Informally, we observe that in a tiny time interval $[s, s+ds]$, about $m\,ds$ items arrive, and $(m-N(s))\,ds \approx m(1-s)\,ds$ of them create new chains, for $s \leq 1$. Of these chains, by (4.1), a proportion $(1-e^{-(t-s)})^{k-1}$ have grown to length at least k at time t, and the result follows by integration over s.

To be more formal, let, for $k \geq 1$, $u \geq 0$ and $j \geq 0$, $f_u^{(k)}(j)$ be the probability that a Yule process that starts with j particles at time 0 has reached at least k particles at time u; this is thus equal to the probability that a chain of length j at some instance s grows to length at least k at time $s+u$. We have $f_u^{(k)}(j) = 1$ for $j \geq k$ and all u, $f_0^{(k)}(j) = [\![j \geq k]\!]$, and, by (4.1),

$$f_u^{(k)}(1) = \bigl(1-e^{-u}\bigr)^{k-1}. \tag{4.3}$$

For fixed k and t, and $0 \leq s \leq t$, let

$$X(s) := \sum_{\ell\geq 1} C_\ell(s) f_{t-s}^{(k)}(\ell) = \sum_{j=1}^{C(s)} f_{t-s}^{(k)}\bigl(L_j(s)\bigr).$$

$X(s)$ is thus the expected number (given $\mathcal{T}(s)$) of the chains present at s that have grown to length at least k at t. In particular, $X(t) = \sum_{\ell\geq k} C_\ell(t)$.

Since new chains, all of length 1, are created with the rate $(m-N(s))_+$, it follows that the process

$$Y(s) := X(s) - \int_0^s f_{t-u}^{(k)}(1)\bigl(m-N(u)\bigr)_+\,du, \qquad 0 \leq s \leq t, \tag{4.4}$$

is a martingale. Moreover, $Y(s)$ has a jump $\Delta Y(s) = f_{t-s}^{(k)}(1)$ of size $|\Delta Y(s)| \leq 1$ each time a new chain is created, and $Y(s)$ is smooth with a bounded derivate between the jumps, so $s \mapsto Y(s)$ is of finite variation, Hence, see e.g. Protter [7, II.6], the quadratic variation $[Y,Y]_t = \sum_{s\leq t} \Delta Y(s)^2 \leq N(t)$, and, observing $Y(0) = X(0) = 0$,

$$\mathbb{E}\,Y(t)^2 = \mathbb{E}[Y,Y]_t \leq \mathbb{E}\,N(t) = mt.$$

In particular, as $m \to \infty$, $Y(t)/m \xrightarrow{\text{p}} 0$ by Chebyshev's inequality, i.e.

$$\frac{X(t)}{m} - \int_0^t f_{t-s}^{(k)}(1)\Bigl(1 - \frac{N(s)}{m}\Bigr)_+ ds = \frac{Y(t)}{m} \xrightarrow{\text{p}} 0.$$

Combined with Lemma 4.2, this shows

$$\frac{\sum_{\ell \ge k} C_\ell(t)}{m} = \frac{X(t)}{m} \xrightarrow{\text{p}} \int_0^t f_{t-s}^{(k)}(1)\bigl(1-s\bigr)_+ ds,$$

which by (4.3) proves (4.2) for fixed $t \ge 0$. Convergence ucp follows by Lemma 4.1. ☐

Lemma 4.5. (i) *For every $t \ge 0$ and $r > 0$, there exists a constant $K(t,r)$, not depending on m, such that*

$$\mathbb{E}\sum_{\ell=1}^{\infty} \ell^r C_\ell(t) = \mathbb{E}\sum_{j=1}^{C(t)} L_j(t)^r \le K(t,r)m.$$

(ii) *For every $r > 0$, there exists a constant $K(r)$, not depending on m or n, such that*

$$\mathbb{E}\sum_{j=1}^{C(\mathcal{T}_{m,n})} L_j(\mathcal{T}_{m,n})^r \le K(r)n.$$

Proof. (i): Since $Y(s)$ in (4.4) is a martingale with $Y(0) = 0$, we have $\mathbb{E}\, Y(t) = 0$ and

$$\begin{aligned}\mathbb{E}\, C_k(t) \le \mathbb{E}\, X(t) &= \mathbb{E}\int_0^t f_{t-u}^{(k)}(1)\bigl(m - N(u)\bigr)_+ du \\ &\le t f_t^{(k)}(1) m = t(1-e^{-t})^{k-1} m.\end{aligned}$$

Hence, if $a < (1-e^{-t})^{-1}$,

$$\mathbb{E}\sum_{\ell=1}^{\infty} a^\ell C_\ell(t) \le amt\sum_{\ell=1}^{\infty}\bigl(a(1-e^{-t})\bigr)^{\ell-1} = \frac{at}{1-a(1-e^{-t})} m. \tag{4.5}$$

Taking e.g. $a = 1 + e^{-t}$, the result follows, since $\sup_\ell \ell^r / a^\ell < \infty$.

(ii): Since $\sum_j L_j(t)^r$ is increasing and $\mathcal{T}(\tau_n) = \mathcal{T}_{m,n}$ is independent of τ_n, we have for every $t > 0$ and $a \ge 1$

$$\mathbb{E}\sum_j a^{L_j(t)} \ge \mathbb{E}\Bigl(\sum_j a^{L_j(\tau_n)} [\![\tau_n \le t]\!]\Bigr) = \mathbb{E}\Bigl(\sum_j a^{L_j(\mathcal{T}_{m,n})}\Bigr)\,\mathbb{P}(\tau_n \le t).$$

Choose $t := 2n/m \le 2$ and $a := 1 + e^{-2}$. Then, by (4.5),

$$\mathbb{E}\sum_j a^{L_j(t)} \le e^4 atm = 2e^2(e^2+1)n.$$

Moreover, $N(t) \sim \mathrm{Po}(mt) = \mathrm{Po}(2n)$, and thus

$$\mathbb{P}(\tau_n \le t) = \mathbb{P}\bigl(N(t) \ge n\bigr) = \mathbb{P}\bigl(\mathrm{Po}(2n) \ge n\bigr) \to 1, \qquad \text{as } n \to \infty;$$

hence, for some constant $c > 0$ and all $n \geq 1$, $\mathbb{P}(\tau_n \leq t) \geq c$. Consequently,

$$\mathbb{E}\Bigl(\sum_j a^{L_j(\mathcal{T}_{m,n})}\Bigr) \leq 2e^2(e^2+1)c^{-1}n,$$

and the result follows. □

Remark 4.6. The collection of chain lengths evolves as a generalized Pólya urn with balls of infinitely many types 0,1,...; we regard each empty cell as a ball of type 0 and each cell in a chain of length ℓ as a ball of type ℓ. The dynamics of the urn thus is that if a ball of type 0 is drawn, it is removed and replaced by a ball of type 1; if a ball of type $\ell \geq 1$ is drawn, ℓ balls of type ℓ are removed together with one ball of type 0, and $\ell+1$ balls of type $\ell+1$ are added. We start with n balls of type 0. We will, however, not use this urn representation.

U. In an unsuccessful search starting at address j in a hash table $\mathcal{T}$, the number $d_j^{\mathsf{U}}(\mathcal{T})$ of searched occupied cells is 0 if the cell j is empty; otherwise the cell belongs to a chain, and d_j^{U} equals 1 + the number of cells in the chain after j.

Hence, $n_0^{\mathsf{U}}(\mathcal{T})$ is the number of empty cells in $\mathcal{T}$, and for $\mathcal{T}(t)$,

$$n_0^{\mathsf{U}}(t) = \bigl(m - N(t)\bigr)_+. \tag{4.6}$$

By Lemma 4.2 thus

$$m^{-1}n_0^{\mathsf{U}}(t) \xrightarrow{\text{ucp}} p_t^{\mathsf{U}}(0) := (1-t)_+. \tag{4.7}$$

(We see also that $n_0^{\mathsf{U}}(\mathcal{T}_{m,n}) = m-n$, directly proving the case $k=0$ in Theorems 2.1(i) and 2.5(i).)

For $k \geq 1$, there is exactly one cell with d_j^{U} in each chain of length $\ell \geq k$, and thus, for $\mathcal{T}(t)$,

$$n_k^{\mathsf{U}}(t) = \sum_{\ell \geq k} C_\ell(t). \tag{4.8}$$

Consequently, Lemma 4.4 yields, for $k \geq 1$, using $u = t-s$,

$$\begin{aligned}\frac{1}{m}n_k^{\mathsf{U}}(t) \xrightarrow{\text{ucp}} p_t^{\mathsf{U}}(k) &:= \int_0^{t\wedge 1}(1-s)\bigl(1-e^{-(t-s)}\bigr)^{k-1}\,ds \\ &= \int_{(t-1)_+}^{t}(1-t+u)\bigl(1-e^{-u}\bigr)^{k-1}\,du.\end{aligned} \tag{4.9}$$

Theorem 2.5(i) follows by Lemma 4.3.

L. For the standard (late-insertion) method L, when a new item arrives with a hash address j, the insertion algorithm begins with an unsuccessful search for the item (followed by finding an empty cell). The displacement of the new item is thus the same as the number d_j^{U} for an unsuccessful search starting at j; note that for L, the displacement never changes after the item is inserted.

Consequently, for $\mathcal{T}(t)$, new items with displacement k are created at the rate $n_k^{\mathsf{U}}(t)$, and

$$Z(t) := n_k^{\mathsf{L}}(t) - \int_0^t n_k^{\mathsf{U}}(s)\,ds \tag{4.10}$$

is a martingale. The jumps are all 1, and we have, as for Y above, $[Z,Z]_t \le N(t)$, and $m^{-1}Z(t) \overset{\mathrm{p}}{\longrightarrow} 0$. Moreover, by Doob's inequality, see e.g. [7, p. 11],

$$\mathbb{E}\bigl(\sup_{s\le t} Z(s)^2\bigr) \le 4\,\mathbb{E}\,Z(t)^2 = 4\,\mathbb{E}[Z,Z]_t \le 4\,\mathbb{E}\,N(t) = 4mt,$$

and hence $m^{-1}Z(t) \overset{\mathrm{ucp}}{\longrightarrow} 0$. Consequently, by (4.10) and (4.9),

$$m^{-1} n_k^{\mathsf{L}}(t) \overset{\mathrm{ucp}}{\longrightarrow} \int_0^t p_s^{\mathsf{U}}(k)\,ds.$$

For $\alpha > 0$ we multiply by $m/n \to \alpha^{-1}$ and find by Lemma 4.3

$$n^{-1} n_k^{\mathsf{L}}(\tau_n) \overset{\mathrm{p}}{\longrightarrow} p_\alpha^{\mathsf{L}}(k) := \alpha^{-1} \int_0^\alpha p_s^{\mathsf{U}}(k)\,ds \tag{4.11}$$

as asserted in Theorem 2.5(ii). Explicitly we have, for $0 < \alpha \le 1$,

$$p_\alpha^{\mathsf{L}}(0) = \alpha^{-1} \int_0^\alpha (1-s)\,ds = 1 - \alpha/2$$

and, for $k \ge 1$,

$$\begin{aligned} p_\alpha^{\mathsf{L}}(k) &= \alpha^{-1} \int_{s=0}^\alpha \int_{t=0}^s (1-s+t)\bigl(1-e^{-t}\bigr)^{k-1}\,dt\,ds \\ &= \alpha^{-1} \int_{t=0}^\alpha \int_{s=t}^\alpha (1-s+t)\bigl(1-e^{-t}\bigr)^{k-1}\,ds\,dt \\ &= \alpha^{-1} \int_{t=0}^\alpha \bigl(\alpha - t - (\alpha-t)^2/2\bigr)\bigl(1-e^{-t}\bigr)^{k-1}\,dt. \end{aligned}$$

For $\alpha = 0$, we observe that $\mathbb{P}\bigl(d_i^{\mathsf{L}}(\mathcal{T}_{m,n}) \ne 0\bigr) \le (i-1)/m$, and thus

$$\mathbb{E}\,|n - n_0^{\mathsf{L}}(\mathcal{T}_{m,n})| \le \sum_{i=1}^n i/m \le n^2/m;$$

hence $\mathbb{E}\,|1 - n_0^{\mathsf{L}}(\mathcal{T}_{m,n})/n| \le n/m \to \alpha = 0$. This yields Theorem 2.5(ii) for $\alpha = 0$ with $p_0^{\mathsf{L}}(0) = 1$ and $p_0^{\mathsf{L}}(k) = 0$, $k \ge 1$.

E. For the early-insertion method E, a new item that hashes to an empty cell gets displacement 0, which remains unchanged for ever. Hence $n_0^{\mathsf{E}}(\mathcal{T}) = n_0^{\mathsf{L}}(\mathcal{T})$, and

$$n^{-1} n_0^{\mathsf{E}}(\mathcal{T}) = n^{-1} n_0^{\mathsf{L}}(\mathcal{T}) \overset{\mathrm{p}}{\longrightarrow} p_\alpha^{\mathsf{L}}(0) = 1 - \alpha/2$$

by the preceding case.

An item x that hashes to an occupied cell gets an initial displacement 1, and this displacement increases each time a new item hashes to one of the cells in the subchain beginning with the hash address of x and ending just before x; the number of such cells is the displacement, and thus the displacement grows according to the same Yule process as the chains. Fix t and k, let $f_u^{(k)}(j)$ be as in the proof of Lemma 4.4, and let, cf. (4.4),

$$W(s) := \sum_{j\geq 1} n_j^{\mathsf{E}}(s) f_{t-s}^{(k)}(j) - \int_0^s f_{t-u}^{(k)}(1) N(u)\, du. \qquad 0 \leq s \leq t, \tag{4.12}$$

Again, this is a martingale. This time, however, the jumps may be larger than 1, since more than one item can get its displacement increased when a new item is inserted. Clearly, the jump ΔW when a new item is inserted is at most the length of the chain where the new item was inserted, since only the items in this chain can have their displacements changed. Since $[W,W]_t$ equals the sum of the squares of all jumps up to t, it is at most the sum of the squares of the lengths of all chains that have existed during the process. A chain in $\mathcal{T}(t)$ of length ℓ is formed by ℓ insertions, and their contribution to the latter sum is $\sum_1^\ell k^2 \leq \ell^3$. Hence,

$$[W,W]_t = \sum_{s\leq t} |\Delta W(s)|^2 \leq \sum_{j=1}^{C(t)} L_j(t)^3 = \sum_{\ell=1}^{\infty} \ell^3 C_\ell(t),$$

and Lemma 4.5 shows that

$$\mathbb{E}\, W(t)^2 = \mathbb{E}[W,W]_t \leq K(t,3)m.$$

Hence, as above, $m^{-1}W(t) \xrightarrow{\text{p}} 0$, which together with Lemma 4.2 and (4.3) yields

$$m^{-1} \sum_{j\geq k} n_j^{\mathsf{E}}(t) - \int_0^t \bigl(1-e^{-(t-u)}\bigr)^{k-1} u\, du \xrightarrow{\text{p}} 0.$$

By Lemma 4.1, thus, with $s = t-u$,

$$m^{-1} \sum_{j\geq k} n_j^{\mathsf{E}}(t) \xrightarrow{\text{ucp}} \int_0^t \bigl(1-e^{-s}\bigr)^{k-1} (t-s)\, ds.$$

Replacing k by $k+1$ and subtracting, we find

$$m^{-1} n_k^{\mathsf{E}}(t) \xrightarrow{\text{ucp}} \int_0^t \bigl(1-e^{-s}\bigr)^{k-1} e^{-s} (t-s)\, ds,$$

and Theorem 2.5(iii) follows by Lemma 4.3 when $\alpha > 0$.

If $\alpha = 0$ we have $n_0^{\mathsf{E}}(\mathcal{T}_{m,n})/n = n_0^{\mathsf{L}}(\mathcal{T}_{m,n})/n \xrightarrow{\text{p}} 1$ by the case L, and thus also $n_k^{\mathsf{E}}(\mathcal{T}_{m,n}) \xrightarrow{\text{p}} 0$ for $k \geq 1$.

This completes the proof of Theorem 2.5.

Proof of Theorem 2.1. Theorem 2.1(i)–(iii) follow by taking expectations in Theorem 2.5 using dominated convergence.

We verify directly that $\{p_\alpha^{\Xi}(k)\}_{k=0}^{\infty}$ is a probability distribution by summing. For U we have, by (4.7) and (4.9) and summing the geometrical series

$\sum_1^\infty \bigl(1-e^{-s}\bigr)^{k-1} = e^s$,

$$\sum_{k=0}^{\infty} p_\alpha^{\mathsf{U}}(k) = 1-\alpha+\int_0^\alpha (1-\alpha+s)e^s\,ds = 1-\alpha+[(s-\alpha)e^s]_0^\alpha = 1.$$

For L and E, the case $\alpha=0$ is trivial. If $0<\alpha\le 1$ we have by (4.11)

$$\sum_{k=0}^{\infty} p_\alpha^{\mathsf{L}}(k) = \alpha^{-1}\int_0^\alpha \sum_{k=0}^{\infty} p_t^{\mathsf{U}}(k)\,dt = \alpha^{-1}\int_0^\alpha dt = 1$$

and, by the definition in Theorem 2.1,

$$\sum_{k=0}^{\infty} p_\alpha^{\mathsf{E}}(k) = 1-\alpha+\alpha^{-1}\int_0^\alpha (\alpha-t)\,dt = 1.$$

Finally, note that each chain of length ℓ contributes (for L, E and U) ℓ displacements which are all at most ℓ. Hence, for $r>0$,

$$\mathbb{E}\bigl(d^{\Xi}(\mathcal{T}_{m,n})^r \mid \mathcal{T}_{m,n}\bigr) \le \sum_j \frac{L_j}{n}L_j^r = \frac{1}{n}\sum_j L_j^{r+1}$$

(also for U), and thus, using Lemma 4.5

$$\mathbb{E}\bigl(d^{\Xi}(\mathcal{T}_{m,n})^r\bigr) \le n^{-1}\,\mathbb{E}\sum_j L_j^{r+1} \le K(r+1).$$

Replacing r by $r+1$ we see that the family $d^{\Xi}(\mathcal{T}_{m,n})^r$ is uniformly integrable, and thus (2.4) implies $\mathbb{E}(d^{\Xi}(\mathcal{T}_{m,n}))^r \to \mathbb{E}(D_\alpha^{\Xi})^r$. □

Proof of Corollary 2.2. It only remains to compute the expectation of D_α^{Ξ}.

$$\begin{aligned}
\mathbb{E}\,D_\alpha^{\mathsf{U}} &= \sum_{k=1}^{\infty} k p_\alpha^{\mathsf{U}}(k) = \int_0^\alpha (1-\alpha+t)\sum_{k=1}^{\infty} k\bigl(1-e^{-t}\bigr)^{k-1}\,dt \\
&= \int_0^\alpha (1-\alpha+t)\bigl(1-(1-e^{-t})\bigr)^{-2}\,dt = \int_0^\alpha (1-\alpha+t)e^{2t}\,dt \\
&= \Bigl[\Bigl(\frac{t}{2}-\frac{\alpha}{2}+\frac{1}{4}\Bigr)e^{2t}\Bigr]_0^\alpha = \frac{1}{4}e^{2\alpha}+\frac{\alpha}{2}-\frac{1}{4},
\end{aligned}$$

and, similarly, (for $\alpha>0$)

$$\begin{aligned}
\mathbb{E}\,D_\alpha^{\mathsf{L}} &= \frac{1}{\alpha}\int_0^\alpha \Bigl(\alpha-t-\frac{(\alpha-t)^2}{2}\Bigr)e^{2t}\,dt \\
&= \frac{1}{\alpha}\Bigl[\Bigl(-\frac{(\alpha-t)^2}{4}+\frac{\alpha-t}{4}+\frac{1}{8}\Bigr)e^{2t}\Bigr]_0^\alpha = \frac{1}{8\alpha}e^{2\alpha}+\frac{\alpha}{4}-\frac{1}{4}-\frac{1}{8\alpha}, \\
\mathbb{E}\,D_\alpha^{\mathsf{E}} &= \frac{1}{\alpha}\int_0^\alpha (\alpha-t)e^t\,dt = \alpha^{-1}\bigl[(\alpha-t+1)e^t\bigr]_0^\alpha = \alpha^{-1}\bigl(e^\alpha-1-\alpha\bigr).
\end{aligned}$$

□

Proof of Corollary 2.4. Similar to Corollary 2.2; we omit the details. (We used `Maple` to perform the integrations.) For the final part, note that

$$\begin{aligned}\operatorname{Var}(\widetilde{D}^{\mathsf{U}}_\alpha) &= \mathbb{E}((D^{\mathsf{U}}_\alpha)^2) + 1 - \alpha - \bigl(\mathbb{E}(D^{\mathsf{U}}_\alpha) + 1 - \alpha\bigr)^2 \\ &= \operatorname{Var}(D^{\mathsf{U}}_\alpha) - 2(1-\alpha)\,\mathbb{E}(D^{\mathsf{U}}_\alpha) + \alpha - \alpha^2.\end{aligned}$$

□

Proof of Theorem 2.7. An immediate consequence of Theorems 2.1 and 2.5 and the following general probabilistic lemma (with $X = d^{\Xi}(\mathcal{T}_{m,n})^r$, $Y = \mathcal{T}_{m,n}$, $Z = \bigl(D^{\Xi}_\alpha\bigr)^r$). □

Lemma 4.7. *Let X_n, Y_n and Z be random variables (with X_n and Y_n defined on the same probability space, and where Y_n may take values in any measure space), such that $X_n \geq 0$ and $Z \geq 0$, and, for every real x, as $n \to \infty$,*

$$\mathbb{P}(X_n \leq x \mid Y_n) \xrightarrow{\text{p}} \mathbb{P}(Z \leq x). \tag{4.13}$$

Suppose further $\mathbb{E}\,X_n \to \mathbb{E}\,Z$. Then $\mathbb{E}(X_n \mid Y_n) \xrightarrow{\text{p}} \mathbb{E}\,Z$.

Proof. Note first that, for every real x, by (4.13) and dominated convergence,

$$\mathbb{P}(X_n \leq x) = \mathbb{E}\bigl(\mathbb{P}(X_n \leq x \mid Y_n)\bigr) \to \mathbb{P}(Z \leq x),$$

and thus $X \xrightarrow{\text{d}} Z$.

For any fixed $K > 0$ we thus have $X_n \wedge K \xrightarrow{\text{d}} Z \wedge K$ and thus, by dominated convergence again, $\mathbb{E}(X_n \wedge K) \to \mathbb{E}(Z \wedge K)$. Hence also

$$\begin{aligned}\mathbb{E}(X_n - K)_+ &= \mathbb{E}(X_n - X_n \wedge K) = \mathbb{E}(X_n) - \mathbb{E}(X_n \wedge K) \\ &\to \mathbb{E}(Z) - \mathbb{E}(Z \wedge K) = \mathbb{E}(Z - K)_+.\end{aligned} \tag{4.14}$$

Moreover,

$$\begin{aligned}&\bigl|\mathbb{E}(X_n \wedge K \mid Y_n) - \mathbb{E}(Z \wedge K \mid Y_n)\bigr| \\ &\qquad = \left|\int_0^K \mathbb{P}(X_n > x \mid Y_n)\,dx - \int_0^K \mathbb{P}(Z > x)\,dx\right| \\ &\qquad \leq \int_0^K \bigl|\mathbb{P}(X_n > x \mid Y_n) - \mathbb{P}(Z > x)\bigr|\,dx\end{aligned}$$

and thus, by (4.13) and, yet again, dominated convergence (twice),

$$\begin{aligned}&\mathbb{E}\,|\,\mathbb{E}(X_n \wedge K \mid Y_n) - \mathbb{E}(Z \wedge K \mid Y_n)| \\ &\qquad \leq \int_0^K \mathbb{E}\bigl|\mathbb{P}(X_n > x \mid Y_n) - \mathbb{P}(Z > x)\bigr|\,dx \to 0.\end{aligned} \tag{4.15}$$

By the triangle inequality and $X = X \wedge K + (X - K)_+$,

$$\begin{aligned}|\,\mathbb{E}(X_n \mid Y_n) - \mathbb{E}(Z)| &\leq \mathbb{E}((X_n - K)_+ \mid Y_n) \\ &\qquad + |\,\mathbb{E}(X_n \wedge K \mid Y_n) - \mathbb{E}(Z \wedge K)| + \mathbb{E}(Z - K)_+.\end{aligned}$$

Taking expectations, we see by (4.15) and (4.14) that

$$\limsup_{n\to\infty} \mathbb{E}\,|\,\mathbb{E}(X_n \mid Y_n) - \mathbb{E}(Z)| \le \limsup_{n\to\infty} \mathbb{E}(X_n - K)_+ + \mathbb{E}(Z-K)_+ = 2\,\mathbb{E}(Z-K)_+.$$

Since K is arbitrary, we see by letting $K \to \infty$ that the left hand side is 0, i.e. $\mathbb{E}\,|\,\mathbb{E}(X_n \mid Y_n) - \mathbb{E}(Z)| \to 0$. □

References

[1] K.B. Athreya & P.E. Ney, *Branching Processes.* Springer-Verlag, Berlin, 1972.

[2] P. Billingsley, *Convergence of Probability Measures.* Wiley, New York, 1968.

[3] W.C. Chen & J.S. Vitter, Analysis of early-insertion standard coalesced hashing. *SIAM J. Comput.* **12**:4 (1983), 667–676.

[4] S. Janson, Individual displacements for linear probing hashing with different insertion policies. *ACM Transactions on Algorithms*, **1**:2 (2005), 177–213.

[5] G.D. Knott, Direct-chaining with coalescing lists. *J. Algorithms* **5**:1 (1984), 7–21.

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

[7] P.E. Protter, *Stochastic Integration and Differential equations.* 2nd ed., Springer, Berlin, 2004.

[8] A. Viola, Exact distributions of individual displacements in linear probing hashing. *ACM Transactions on Algorithms*, **1**:2 (2005), 214–242.

[9] J.S. Vitter & W.C. Chen, *Design and analysis of coalesced hashing.* International Series of Monographs on Computer Science, Oxford University Press, New York, 1987.

[10] F.A. Williams, Handling identifiers as internal symbols in language processors. *Communications of the ACM* **2**:6 (1959), 21–24.

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/