Combinatorics, Probability and Computing (2013) 22, 455–476. © Cambridge University Press 2013
doi:10.1017/S0963548312000582

The Maximum Displacement for Linear Probing Hashing

NICLAS PETERSSON

Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden
(e-mail: niclas.petersson@gmail.com)

Received 4 March 2008; revised 11 November 2012; first published online 24 January 2013

In this paper we study the maximum displacement for linear probing hashing. We use the standard probabilistic model together with the insertion policy known as First-Come-(First-Served). The results are of asymptotic nature and focus on dense hash tables. That is, the number of occupied cells n and the size of the hash table m tend to infinity with ratio $n/m \to 1$. We present distributions and moments for the size of the maximum displacement, as well as for the number of items with displacement larger than some critical value. This is done via process convergence of the (appropriately normalized) length of the largest block of consecutive occupied cells, when the total number of occupied cells n varies.

AMS 2010 *Mathematics subject classification*: Primary 68W40
Secondary 08P10, 60F05, 60C05

1. Introduction

The standard version of linear probing hashing with insertion policy *First-Come-*(*First-Served*) can be described as follows, where n and m are integers with $0 \leqslant n \leqslant m$ (see Knuth [16, Sec. 6.4, Algorithm 6.4.L]):

n items $x_1, \ldots, x_n$ are placed sequentially into a table with m cells $1, \ldots, m$ according to some hash sequence $\{h_i\}_1^n$ where $h_i \in \{1, \ldots, m\}$. That is, item x_i is inserted into cell h_i if it is empty. If not, then x_i probes cell h_i+1, h_i+2, *etc.* until an empty cell is found. All positions in the table can be interpreted modulo m as any item that is rejected from the last cell m, continues to probe cell 1, 2, *etc.* until it is inserted in an empty cell.

The insertion policy specifies how conflicts are solved when a new item probes an already occupied cell. Other common insertion policies are the so-called *Last-Come-*(*First-Served*) (see Poblete and Munro [19]) and *Robin Hood* (see Celis, Larson and Munro [3] and Knuth [16, Answer 6.4-67]). In this paper, though, only the First-Come policy described above will be considered.

Throughout this paper, all hash sequences follow the standard probabilistic model. That is, the hash addresses h_i are independent random numbers, uniformly distributed on

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

$\{1,\ldots,m\}$. Hence all m^n possible hash sequences $\{h_i\}_1^n$ are equally likely when choosing a sequence at random.

Displacements. If item x_i is inserted into cell q_i, then its *displacement* or *individual displacement* is

$$d_i := (q_i - h_i) \mod m.$$

Thus, the displacement equals the number of unsuccessful probes item x_i makes before being inserted into an empty cell. The displacement is a measure of time (or cost) it takes to insert an item into the table. Once the hash table has been made, the only way to find x_i is to probe cell h_i, h_i+1, *etc.*, until x_i is located. The search time needed to allocate x_i in the final table is therefore proportional to d_i+1.

The *total displacement*, *i.e.*, the sum of all individual displacements, has been studied in detail by Flajolet, Poblete and Viola [9]. The individual displacement of an item chosen at random has been studied by Janson [11] for the three different insertion policies First-Come, Last-Come and Robin Hood. Some results for individual displacements have also been found independently by Viola: see [21].

The main purpose of this paper is to establish distribution and moments of the *maximum displacement*,

$$d_{\max} := \max_{1\leqslant i\leqslant n} d_i,$$

when the number of occupied cells n and the size of the hash table m tend to infinity with ratio $n/m \to 1$; see Theorems 5.1 and 5.2.

A secondary purpose is to study the number of items with displacements very close to the maximum. Such items are from now on said to have a *critical displacement*. The exact definition of critical displacement will be postponed until later, although the principle is to define

$$\text{'item } x_i \text{ has a critical displacement'} \quad \Leftrightarrow \quad d_i \geqslant \zeta(m),$$

for some (appropriate) function $\zeta(m)$ that only depends on m. It is possible to choose $\zeta(m)$ such that the number of items with a critical displacement stays bounded in probability when m and n tend to infinity. The distribution, expectation and variance of the number of items with critical displacements are stated in Theorem 4.9, respectively Corollary 4.10.

For further reading, Gonnet [10] has performed asymptotic analysis for the expected maximum displacement with the First-Come policy. Devroye [7] has also contributed to the same topic but with the difference that the hashing is done with non-uniform probabilities. Devroye, Morin and Viola [8] analyse the expected maximum displacement for the Robin Hood policy.

Remark. This paper adopts many of the basic notations used in [11] by Janson. However, the reader should be informed that there is not a complete one-to-one correspondence between the set of notations in [11] and those presented here.

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

2. Preliminaries

A hash table is completely described by the number of cells m, the number of items n, and the hash sequence $\{h_i\}_1^n$. When m and n are finite, the random hash table is denoted by $\mathcal{T}_{m,n}$. The hash table is called *full* if $n = m$ and *almost full* if $n = m - 1$.

Confined hash tables. A hash table is called *confined* if the last cell in the table is empty. Confined hash tables are from now on distinguished with a prime $\mathcal{T}'_{m,n}$. Due to the circular symmetry of the construction procedure, all cells are equally likely to be empty in the final table. Thus, a confined hash table is simply a standard table conditioned on the last cell being empty. So any (non-full) hash table can be given the confined property by choosing an empty cell with uniform probability (among all the empty cells) and then relabelling the table so that the chosen cell becomes the last. It follows that the sequence of displacements has the same distributions whether or not the table is confined. Further analysis can therefore be restricted to confined hash tables without loss of generality.

A hash table $\mathcal{T}_{m,n}$ can be constructed by m^n different hash sequences, all equally likely when hashing with uniform probability. Each of these tables has $m - n$ empty cells, so by symmetry, the probability that the last cell is empty is $(m - n)/m$, and thus the number of hash sequences that yield a confined hash table $\mathcal{T}'_{m,n}$ is $(m - n)m^{n-1}$. In particular, there are k^{k-2} different ways to construct a confined almost full hash table $\mathcal{T}'_{k,k-1}$.

Blocks. A *block* in a hash table $\mathcal{T}_{m,n}$ is defined as a sequence of consecutive cells $\{i + 1, \ldots, j\}$ (modulo m) such that cells i and j are empty and all cells in between are occupied. Thus, every block contains exactly one empty cell (the last one) and the table has as many blocks as it has empty cells. Note that the smallest block size is 1, since an empty cell following the empty cell of the previous block constitutes a block of its own. The internal structures of different blocks are also independent of each other. Thus, any block of a given length k may therefore be regarded as a confined almost full hash table $\mathcal{T}'_{k,k-1}$ in itself.

Coalescence of blocks. During the construction of a hash table every new inserted item merges two blocks into one. When a new item is inserted it will target a particular block with probability proportional to the size of that block. Thus, the probability that the given block coalesces with its neighbour to the right is given by the block's own size divided by m. A particular block can, in other words, become larger either by assimilating its neighbour to the right or by being assimilated by its neighbour to the left. The probability of this occurring when a single new item is inserted into the table is therefore proportional to the sum of the block's own size and the size of the block to the left. It means that large blocks and blocks with large neighbours to the left are the ones most likely to grow even larger. Chassaing and Louchard [5] proved that a giant block will emerge if the table is dense enough when m and n tend to infinity.

Fragmentation of hash tables. An alternative approach to the evolution of blocks and displacements during the construction of a hash table is to start with the final table and then go backwards in time. Consider a confined almost full random hash table $\mathcal{T}'_{m,m-1}$ constructed by some unknown hash sequence $\{h_i\}_{i=1}^{m-1}$, where each h_i was chosen with

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

uniform probability. By removing the items one by one in reversed sequential order, the evolution of the table can be reconstructed in an algorithmic way.

First, determine the shape of the hash table before x_{m-1} was inserted. That is, remove the last inserted item x_{m-1} from the table. The hash table will be fragmented into two blocks, a left one called B_L and a right one called B_R. The probability mass function for the size of B_R (and thus also B_L) can be calculated directly from equations (37) and (38) in Knuth [16, Sec. 6.4, Algorithm 6.4.L]. The simple calculation is omitted in this paper, but the resulting function is given by

$$\mathbb{P}(B_R = k) = \frac{1}{k}\binom{m-2}{k-1}\left(\frac{k}{m}\right)^{k-1}\left(1-\frac{k}{m}\right)^{m-k-1}, \quad k = 1,\ldots,m-1.$$

This is the probability mass function of the Consul distribution, which is a basic Lagrangian distribution of the first kind (BLD_1), according to the classification in Johnson, Kemp and Kotz [13, Sections 2.5 and 7.5].

Second, determine the displacement of x_{m-1}. The removed item must have targeted one of the cells in the left block B_L as it was inserted into the table. This fact conditions h_{m-1} to be uniformly distributed among the $(m-k)$ cells of block B_L (assuming $B_R = k$). So the displacement of x_i can be determined by picking a number between 0 and $m-k-1$ with uniform probability.

Third, determine which block the next item to be removed belongs to. The internal block structures are independent of other blocks. So the probability that a particular block will get chosen is size-biased with respect to the number of items. That is, in the case of only two blocks, $B_L = m-k$ and $B_R = k$, the probability of picking B_L is given by $(m-k-1)/(m-2)$.

The procedure can now be repeated by treating the chosen block as a confined almost full hash table in itself. The first two steps of the algorithm are identical, but only restricted to the chosen block. The third step, however, requires that the new block that is about to be fragmented is picked with a size-biased probability from the entire table.

Remark. Note that the Consul distribution converges to the Borel distribution (with parameter equal to 1, Bo(1)) when m tends to infinity but k is finite:

$$\mathbb{P}(B_R = k) \to \frac{k^{k-1}}{k!}e^{-k}, \quad \text{as } m \to \infty,\ k = 1,2,\ldots$$

Hash tables and random forests. There is a one-to-one correspondence between linear probing hash tables and Pavlov's random rooted forests [17]. Knuth [16, Sec. 6.4, Exercise 31] observed this and included it as an exercise to 'Find an interesting connection between parking sequences and trees'. A common description of this connection follows here.

As mentioned above, there are k^{k-2} different ways to construct a confined hash table $\mathcal{T}'_{k,k-1}$.

The number of different rooted trees with $(k-1)$ non-root labelled vertices is also equal to k^{k-2}: see [17]. Thus, letting the root represent the empty cell in $\mathcal{T}'_{k,k-1}$ and the labelled vertices represent the (numbered) items in $\mathcal{T}'_{k,k-1}$, it is possible to link every hash table

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

$\mathcal{T}'_{k,k-1}$ to a unique labelled rooted tree. How this representation is carried out is irrelevant in this particular paper, but the reader may find it interesting to know that there exists a 'canonical representation', which Chassaing, Janson, Louchard and Marckert [4, 5, 6] have proved to be very useful.

A confined hash table $\mathcal{T}'_{m,n}$, which is not necessarily almost full, consists of $(m-n)$ ordered blocks. These blocks may each be regarded as confined almost full hash tables in themselves. Thus, every confined table $\mathcal{T}'_{m,n}$ can be linked to a unique rooted forest by the same representation used for the individual rooted tree. The only addition is that the roots of the forest must be given their own set of labels so that the information about the order in which the blocks appear is not lost.

This one-to-one correspondence between hash tables and random forests will be used to gain access to Pavlov's asymptotic results on the size of the largest tree in a random forest [17]. The internal structure of a particular block will mainly be studied by the fragmentation approach, which is why no explicit representation scheme needs to be introduced.

There is also a very elegant way to construct the random forests described above via the theory for branching processes [17]. Start with $(m-n)$ labelled roots. Let every root be the ancestor to a genealogical tree with $\mathrm{Po}(\lambda)$-distributed offspring, and condition these $(m-n)$ Galton–Watson branching processes to have a total progeny of n children. (The parameter $\lambda > 0$ is irrelevant, but it is natural to choose $\lambda = n/m$.) The children can then be labelled randomly to get a unique representation of some random confined hash table $\mathcal{T}'_{m,n}$.

Remark. Note that the total progeny of a Galton–Watson branching process where each individual has $\mathrm{Po}(\alpha)$ children and $0 < \alpha \leqslant 1$, is a $\mathrm{Bo}(\alpha)$-distributed random variable. This means that the sizes of the different blocks in the random hash table $\mathcal{T}_{m,n}$ is a string of $(m-n)$ $\mathrm{Bo}(n/m)$-distributed random variables conditioned on the total size being m.

3. The size of the largest block in dense hash tables

Let $N = m - n$ denote the number of different blocks in a confined hash table $\mathcal{T}'_{m,n}$. Chassaing and Louchard [5] proved that a phase transition of the largest block occurs when m and n tend to infinity in such a way that $N = \Theta(\sqrt{m})$. That is, the length of the largest block in $\mathcal{T}'_{m,n}$, which from now on is denoted $\eta(m,n)$, is subject to the following change when $m \to \infty$:

$$\begin{aligned} \frac{\eta(m,n)}{m} &\xrightarrow{\mathrm{p}} 0 \quad \text{if } \sqrt{m} = o(N), \\ \frac{\eta(m,n)}{m} &\xrightarrow{\mathrm{p}} 1 \quad \text{if } N = o(\sqrt{m}). \end{aligned} \tag{3.1}$$

Consider the construction of an almost full hash table $\mathcal{T}'_{m,m-1}$ when m tends to infinity. After the phase transition occurs there are still some $o(\sqrt{m})$ items left to hash into the table. Most of these items will target the largest block, which by now dominates the table. The probability of getting large individual displacements is therefore significantly increased by the phase transition. The size of the largest block also determines the maximum possible

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

size of the next inserted item's displacement. Thus, it is reasonable to assume that the item with the maximum displacement will be hashed into the table first after the phase transition has occurred. This motivates the use of the fragmentation approach described in Section 2 for further analysis.

The probability items having a large displacement decrease as the largest block is fragmented. At some point the size of the largest block will be less than the maximum displacement found so far. The fragmentation can then be stopped since the remaining items can no longer beat the current maximum. As the results will show, it imposes no restrictions to assume that this happens while N is still of order $o(\sqrt{m})$.

As a short digression, consider the one-to-one correspondence between hash tables and random forests again. The following theorem for the maximum size of a tree in a random forest was proved by Pavlov; see [17, Chapter 2, Theorem 5] for the proof.

Theorem 3.1. *Let η be the size of the largest tree in a random forest with m vertices, of which $N = o(\sqrt{m})$ are roots. Then, for any fixed positive x,*

$$\mathbb{P}\left(\frac{m-\eta}{N^2} \leqslant x\right) \to \frac{1}{\sqrt{2\pi}} \int_0^x y^{-3/2}\, e^{-1/2y}\, \mathrm{d}y \quad \text{as } m, N \to \infty. \tag{3.2}$$

The limit distribution given by equation (3.2) is identical to the marginal distribution of the stable subordinator of index 1/2, for a fixed time point. (See for instance Karatzas and Shreve [15, Chapter 6.2.C and Theorem 6.2.1].) Throughout the rest of the paper, the stable subordinator of index 1/2 will be denoted by X_t, $t > 0$. X_t is an increasing process with stationary independent increments, such that, for any fixed time $t > 0$,

$$\mathbb{P}\big(X_t \leqslant x\big) = \frac{1}{\sqrt{2\pi}} \int_0^{x/t^2} y^{-3/2}\, e^{-1/2y}\, \mathrm{d}y, \quad 0 < x < \infty.$$

Note that the process is scale-invariant, $X_t \stackrel{\mathrm{d}}{=} t^2 X_1$.

Now, return to the fragmentation of the almost full hash table $\mathcal{T}'_{m,m-1}$. Let $r(m)$ be an arbitrary increasing function of m such that $r(m) \to \infty$ and $r(m) = o(\sqrt{m})$, and let the number of empty cells in $\mathcal{T}'_{m,m-N}$ be a function of time and $r(m)$,

$$N_t^m = \lceil t\, r(m) \rceil, \quad t > 0.$$

The size of the largest block then becomes a continuous time process,

$$\eta_t^m = \eta\big(m,\, m - \lceil t\, r(m) \rceil\big), \quad t > 0.$$

Let X_t^m be a normalized process of η_t^m defined by

$$X_t^m = \frac{m - \eta_t^m}{r(m)^2}, \quad t > 0.$$

Then, for any fixed time $t > 0$,

$$X_t^m \xrightarrow{\mathrm{d}} X_t \quad \text{as } m \to \infty, \tag{3.3}$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

since

$$\mathbb{P}\big(X_t^m \leqslant x\big) = \mathbb{P}\left(\frac{m-\eta_t^m}{r(m)^2} \leqslant x\right) = \mathbb{P}\left(\frac{m-\eta_t^m}{(N_t^m)^2} \leqslant \frac{x\, r(m)^2}{\lceil t\, r(m)\rceil^2}\right)$$
$$\to \frac{1}{\sqrt{2\pi}} \int_0^{x/t^2} y^{-3/2}\, e^{-1/2y}\, \mathrm{d}y \quad \text{as } m \to \infty,$$

as a result of Theorem 3.1.

Theorem 3.2. *For any normalizing sequence $r(m) = o(\sqrt{m})$ with $r(m) \to \infty$, the sequence of random processes $\{X_t^m\}$ converges in distribution to the Lévy process known as the stable subordinator of index 1/2,*

$$X_t^m \xrightarrow{\mathrm{d}} X_t \quad \textit{as } m \to \infty,$$

in the usual Skorokhod topology on $\mathcal{D}[0, \infty)$.

For a definition of the Skorokhod topology, see Billingsley [2].

Proof. Since the full proof is rather extensive, the principal method is presented here and the most technical parts are proved separately in the three consecutive lemmas below.

If $\{X_t^m\}$ is tight and the finite-dimensional distributions converge, then the entire process converges. To prove this within the current time scale causes some resistance, especially for the convergence of the finite-dimensional distributions. To get around this problem, the original time t will be replaced with the random time-deformation $\Delta^m(t)$, defined below.

Recall that X_t^m is a measure of the size of the largest block when $\lceil t\, r(m)\rceil$ items have been removed from the table $\mathcal{T}'_{m,m-1}$. The items that belong to the currently largest block at the time of their removal are from now on called *marked* items. Define $\Delta^m(t)$ as the minimum time needed to remove $\lceil t\, r(m)\rceil$ marked items.

The difference between the original time t and the time-deformation $\Delta^m(t)$ is uniformly small when m tends to infinity (Lemma 3.3). Thus, $\{X_t^m\}$ converges if and only if $\{X^m_{\Delta^m(t)}\}$ converges in the usual Skorokhod topology on $\mathcal{D}[0, \infty)$: see Billingsley [2, Chapter 3]. Since $\{X^m_{\Delta^m(t)}\}$ is tight (Lemma 3.4) and the finite-dimensional distributions converge (Lemma 3.5), the entire process converges. □

Remark. It is true that the time-deformation $\Delta^m(t)$ introduces some new problems of its own (like the entire Lemma 3.3 for instance). However, $\Delta^m(t)$ will also be useful below, and is therefore worth the extra effort needed in this proof.

Lemma 3.3. *$\{\Delta^m(t)\}$ converges uniformly to t in probability, on compact intervals. That is, for any fixed time t and any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{s \leqslant t} |\Delta^m(s) - s| > \varepsilon\right) \to 0 \quad \textit{as } m \to \infty. \tag{3.4}$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

Proof. Define ${}^\star\Delta^m(t)$ such that ${}^\star\Delta^m(t)\, r(m)$ is the number of marked items that have been removed at time t. Thus, ${}^\star\Delta^m(t)$ is a kind of generalized inverse to $\Delta^m(t)$. Since ${}^\star\Delta^m(t)\, r(m)$ is bounded by $\lceil t\, r(m)\rceil$ it would be convenient if it were possible to show something similar to

$$\mathbb{P}\left(\sup_{s\leqslant t}|s-{}^\star\Delta^m(s)|>\varepsilon\right)\to 0 \quad \text{as } m\to\infty,$$

instead of proving equation (3.4). With a slight adjustment, this is possible.

Let $\varepsilon>0$ and assume that

$$\sup_{s\leqslant t}|\Delta^m(s)-s|>\varepsilon.$$

Note first that by the definitions above, $\Delta^m(s)r(m) > N^m_{\Delta^m(s)} - 1 \geqslant \lceil sr(m)\rceil \geqslant sr(m)$, and thus $\Delta^m(s)\geqslant s$. Thus there exists an $s\leqslant t$ such that $(s+\varepsilon)<\Delta^m(s)$. Since $\Delta^m(s)$ is the minimum time needed to remove $\lceil s\, r(m)\rceil$ marked items, there are strictly fewer marked items removed at time $(s+\varepsilon)$, *i.e.*,

$${}^\star\Delta^m(s+\varepsilon)\, r(m) < \lceil s\, r(m)\rceil.$$

The number of marked items removed at time $(t+\varepsilon)$ must therefore satisfy the inequality

$$\begin{aligned}{}^\star\Delta^m(t+\varepsilon)\, r(m) &< \lceil s\, r(m)\rceil + \lceil (t-s)\, r(m)\rceil \leqslant \lceil t\, r(m)\rceil + 1\\ &\leqslant \lceil (t+\varepsilon)\, r(m)\rceil - \lfloor \varepsilon\, r(m)\rfloor + 1,\end{aligned}$$

or equivalently

$$\lceil (t+\varepsilon)\, r(m)\rceil - {}^\star\Delta^m(t+\varepsilon)\, r(m) \geqslant \lfloor \varepsilon\, r(m)\rfloor. \tag{3.5}$$

The left-hand side of equation (3.5) denotes the number of unmarked items removed at time $(t+\varepsilon)$. Thus,

$$\begin{aligned}&\sup_{s\leqslant t}|\Delta^m(s)-s|>\varepsilon \quad \text{for some } \varepsilon>0 \quad \Longrightarrow\\ &\#\text{ unmarked items removed at time } (t+\varepsilon) \geqslant \lfloor \varepsilon\, r(m)\rfloor.\end{aligned}$$

Now, let $\delta(m) = r(m)/\sqrt{m}$; thus $\delta(m)\to 0$ and $r(m)^2/(m\delta(m))\to 0$. If an item is removed at time u, it has probability less than

$$\frac{m-\eta^m_u}{m} \tag{3.6}$$

of being unmarked. Let us stop the process at the time when this ratio becomes larger than $\delta(m)$, $\tau = \inf\{u : (m-\eta^m_u)/m \geqslant \delta(m)\}$. Since there are $\lceil (t+\varepsilon)\, r(m)\rceil$ items removed up to time $(t+\varepsilon)$, the total number of unmarked items removed at time $\min(t+\varepsilon,\tau)$ is stochastically smaller than a binomial random variable $W^m_{t+\varepsilon} \in \mathrm{Bin}\big(\lceil (t+\varepsilon)\, r(m)\rceil,\, \delta(m)\big)$.

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

It follows that

$$\begin{aligned}\mathbb{P}\left(\sup_{s\leqslant t}|\Delta^m(s)-s|>\varepsilon\right) &\leqslant \mathbb{P}\big(W^m_{t+\varepsilon}\geqslant \lfloor \varepsilon\, r(m)\rfloor\big)+\mathbb{P}(\tau\leqslant t+\varepsilon)\\ &\leqslant \frac{\mathbb{E}\big[W^m_{t+\varepsilon}\big]}{\lfloor \varepsilon\, r(m)\rfloor}+\mathbb{P}\left(\frac{m-\eta^m_{t+\varepsilon}}{m}\geqslant \delta(m)\right)\\ &\leqslant \frac{\lceil (t+\varepsilon)\, r(m)\rceil \delta(m)}{\lfloor \varepsilon\, r(m)\rfloor}+\mathbb{P}\left(\frac{m-\eta^m_{t+\varepsilon}}{r(m)^2}\geqslant \frac{m\delta(m)}{r(m)^2}\right)\\ &\to 0 \quad \text{as } m\to\infty,\end{aligned}$$

since $\delta(m)\to 0$, $m\delta(m)/r(m)^2\to\infty$ and $(m-\eta^m_{t+\varepsilon})/r(m)^2$ converges in distribution: see Theorem 3.1 and (3.3). □

Lemma 3.4. *The sequence* $\{X^m_{\Delta^m(t)}\}$ *is tight.*

Proof. Tightness can be proved via equicontinuity and Aldous's tightness criterion: see Kallenberg [14, Theorem 16.11]. That is, $\{X^m_{\Delta^m(t)}\}$ is tight if, for any $\varepsilon>0$ and any bounded sequence $\{\tau_m\}$ of stopping times and any sequence of positive constants $\{h_m\}$ such that $h_m\to 0$ as m tends to infinity,

$$\mathbb{P}\big(|X^m_{\Delta^m(\tau_m+h_m)}-X^m_{\Delta^m(\tau_m)}|>\varepsilon\big)\to 0 \quad \text{as } m\to\infty.$$

A total of $\lceil h_m\, r(m)\rceil$ (or possibly $\lfloor h_m\, r(m)\rfloor$) marked items are removed in the time interval $(\Delta^m(\tau_m),\, \Delta^m(\tau_m+h_m)]$. In order to prove the desired convergence, it is enough to consider the 'worst' of these two cases, that is, when $\lceil h_m\, r(m)\rceil$ marked items are removed. Now, it is only the size of the largest block at time $\Delta^m(\tau_m)$ that matters when considering the removal of these additional $\lceil h_m\, r(m)\rceil$ marked items. Thus, the increment

$$X^m_{\Delta^m(\tau_m+h_m)}-X^m_{\Delta^m(\tau_m)} \tag{3.7}$$

does not depend on whether τ_m is a stopping time or a deterministic time.

Condition increment (3.7) on $X^m_{\Delta^m(\tau_m)}$ being larger, respectively smaller, than $\sqrt{m}/r(m)$ to obtain the inequality

$$\begin{aligned}\mathbb{P}\big(|X^m_{\Delta^m(\tau_m+h_m)}-X^m_{\Delta^m(\tau_m)}|>\varepsilon\big) &\leqslant \mathbb{P}\left(X^m_{\Delta^m(\tau_m)}>\frac{\sqrt{m}}{r(m)}\right)\\ &+\mathbb{P}\left(|X^m_{\Delta^m(\tau_m+h_m)}-X^m_{\Delta^m(\tau_m)}|>\varepsilon \;\middle|\; X^m_{\Delta^m(\tau_m)}\leqslant\frac{\sqrt{m}}{r(m)}\right).\end{aligned}$$

The sequence τ_m is bounded, and thus there exists T such that $\tau_m\leqslant T$. Since $r(m)=o(\sqrt{m})$, it follows from (3.3) and Lemma 3.3 that

$$\mathbb{P}\left(X^m_{\Delta^m(\tau_m)}>\frac{\sqrt{m}}{r(m)}\right)\leqslant \mathbb{P}\left(X^m_{T+1}>\frac{\sqrt{m}}{r(m)}\right)+\mathbb{P}\big(\Delta^m(\tau_m)>T+1\big)\to 0$$

as $m\to\infty$. Thus, it only remains to prove that for any $k=k(m)$ such that

$$(m-r(m)\sqrt{m})\leqslant k\leqslant m,$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

the conditional random variable

$$\big(X^m_{\Delta^m(\tau_m+h_m)} - X^m_{\Delta^m(\tau_m)} \mid \eta^m_{\Delta^m(\tau_m)} = k\big)$$

converges to zero in probability. Let $\{X^{m,k}_t\}$ denote the auxiliary sequence of random processes defined by

$$X^{m,k}_t = \frac{k - \eta\big(k,\, k - \lceil t\, r(m)\rceil\big)}{r(m)^2}, \quad t > 0.$$

Then

$$X^{m,k}_{\Delta^m(h_m)} \overset{\mathrm{d}}{=} \big(X^m_{\Delta^m(\tau_m+h_m)} - X^m_{\Delta^m(\tau_m)} \mid \eta^m_{\Delta^m(\tau_m)} = k\big),$$

since both sides of the equation describe the difference (normalized by $r(m)$) of the largest block size after $\lceil h_m r(m)\rceil$ marked items have been removed from the subtable $\mathcal{T}'_{k,k-1}$.

Moreover, $k/m \to 1$ as m tends to infinity, so $r(m) = o(\sqrt{k})$. This means that $r(m)$ is a valid scaling and normalization function for $\mathcal{T}'_{k,k-1}$ as k tends to infinity. Thus, it is enough to prove that

$$\mathbb{P}\big(X^m_{\Delta^m(h_m)} > \varepsilon\big) \to 0 \quad \text{as } m \to \infty.$$

Let $\delta > 0$ and condition $X^m_{\Delta^m(h_m)}$ on whether $(\Delta^m(h_m) - h_m) > \delta$, to obtain the inequality

$$\mathbb{P}\big(X^m_{\Delta^m(h_m)} > \varepsilon\big) \leqslant \mathbb{P}\big(\Delta^m(h_m) - h_m > \delta\big) + \mathbb{P}\big(X^m_{h_m+\delta} > \varepsilon\big). \tag{3.8}$$

The difference between $\Delta^m(h_m)$ and h_m is uniformly small as m tends to infinity (Lemma 3.3), *i.e.*, for any $\delta > 0$,

$$\mathbb{P}\big(\Delta^m(h_m) - h_m > \delta\big) \to 0 \quad \text{as } m \to \infty.$$

It remains to prove that the last term in (3.8) can be made arbitrarily small for large m. For any $\delta > 0$, there exists an $M \in \mathbb{N}$ such that $h_m < \delta$ for all $m \geqslant M$. Thus,

$$\limsup_{m\to\infty} \mathbb{P}\big(X^m_{h_m+\delta} > \varepsilon\big) \leqslant \limsup_{m\to\infty} \mathbb{P}\big(X^m_{2\delta} > \varepsilon\big).$$

Define an auxiliary sequence of fragmentation processes $\{\bar{X}^m_t\}$ that is scaled and normalized by $2\delta r(m)$ instead of just $r(m)$. Then

$$\begin{aligned} X^m_{2\delta} &= \frac{m - \eta(m,\, m - \lceil 2\delta \cdot r(m)\rceil)}{r(m)^2} \\ &= (2\delta)^2 \cdot \frac{m - \eta(m,\, m - \lceil 1 \cdot 2\delta r(m)\rceil)}{\big(2\delta r(m)\big)^2} = 4\delta^2 \bar{X}^m_1. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}\big(X^m_{2\delta} > \varepsilon\big) = \mathbb{P}\left(\bar{X}^m_1 > \frac{\varepsilon}{4\delta^2}\right) &\to \frac{1}{\sqrt{2\pi}} \int_{\varepsilon/4\delta^2}^{\infty} y^{-3/2}\, e^{-1/2y}\, \mathrm{d}y \\ &\leqslant \frac{1}{\sqrt{2\pi}} \big[-2y^{-1/2}\big]_{\varepsilon/4\delta^2}^{\infty} = \frac{4\delta}{\sqrt{2\pi\varepsilon}}, \end{aligned}$$

as m tends to infinity. So for any $\varepsilon > 0$ it follows that

$$\mathbb{P}\big(X^m_{\Delta^m(h_m)} > \varepsilon\big) \to 0 \quad \text{as } m \to \infty,$$

since $\delta > 0$ can be chosen arbitrarily small. Thus, the proof is complete. □

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

Lemma 3.5. *The finite-dimensional distributions of* $\{X^m_{\Delta^m(t)}\}$ *converge,*

$$X^m_{\Delta^m(t)} \xrightarrow{\text{fd}} X_t \quad \text{as } m \to \infty.$$

Proof. The finite-dimensional distributions of $\{X^m_{\Delta^m(t)}\}$ converge if and only if all linear combinations converge (the Cramér–Wold device: see for instance Kallenberg [14, Corollary 5.5]). Recall that the supposed limit process X_t has independent increments and that independent increments ensure convergence of all linear combinations. Thus, it is sufficient to prove that $\{X^m_{\Delta^m(t)}\}$ has independent increments in the limit.

Let s, t, x and y be fixed positive constants and define the set

$$J = \{k \in \mathbb{N} \,:\, m - x\,r(m)^2 \leqslant k \leqslant m\}.$$

Note that

$$k \in J \quad \Rightarrow \quad k/m = 1 + o(1).$$

Start with two increments,

$$\begin{aligned}
&\mathbb{P}\big(X^m_{\Delta^m(s+t)} - X^m_{\Delta^m(s)} \leqslant y\,,\, X^m_{\Delta^m(s)} \leqslant x\big) \\
&\quad = \mathbb{P}\left(\frac{\eta^m_{\Delta^m(s)} - \eta^m_{\Delta^m(s+t)}}{r(m)^2} \leqslant y\,,\, \frac{m - \eta^m_{\Delta^m(s)}}{r(m)^2} \leqslant x\right) \\
&\quad = \sum_{k\in J} \mathbb{P}\left(\frac{\eta^m_{\Delta^m(s)} - \eta^m_{\Delta^m(s+t)}}{r(m)^2} \leqslant y \;\middle|\; \eta^m_{\Delta^m(s)} = k\right) \mathbb{P}\big(\eta^m_{\Delta^m(s)} = k\big) \\
&\quad \overset{(1)}{=} \sum_{k\in J} \mathbb{P}\left(\frac{k - \eta^k_{\Delta^k(t^\circ)}}{r(m)^2} \leqslant y\right) \mathbb{P}\big(\eta^m_{\Delta^m(s)} = k\big) \\
&\quad = \sum_{k\in J} \mathbb{P}\left(X^k_{\Delta^k(t^\circ)} \cdot \frac{r(k)^2}{r(m)^2} \leqslant y\right) \mathbb{P}\big(\eta^m_{\Delta^m(s)} = k\big) \\
&\quad \overset{(2)}{=} \sum_{k\in J} \big(\mathbb{P}\big(X_t \leqslant y\big) + o(1)\big)\, \mathbb{P}\big(\eta^m_{\Delta^m(s)} = k\big) \\
&\quad = \big(\mathbb{P}\big(X_t \leqslant y\big) + o(1)\big)\, \mathbb{P}\big(X^m_{\Delta^m(s)} \leqslant x\big) \\
&\quad \to \mathbb{P}\big(X_t \leqslant y\big)\, \mathbb{P}\big(X_s \leqslant x\big) \quad \text{as } m \to \infty,
\end{aligned}$$

where (1) either $\lceil t\,r(m)\rceil$ or $\lfloor t\,r(m)\rfloor$ marked items are removed in the time interval

$$(\Delta^m(s),\, \Delta^m(s+t)],$$

so t° satisfies the inequality

$$\frac{\lfloor t\,r(m)\rfloor}{r(k)} \leqslant t^\circ \leqslant \frac{t\,r(m)}{r(k)};$$

and (2) follows from Cramér's theorem, *i.e.*, for any $k \in J$,

$$X^k_{\Delta^k(t^\circ)} \cdot \frac{r(k)^2}{r(m)^2} \xrightarrow{\text{d}} X_t \quad \text{as } m \to \infty,$$

since

$$\frac{r(k)}{r(m)} \to 1, \quad t^\circ \to t \quad \text{and} \quad X^k_{\Delta^k(t)} \xrightarrow{\text{d}} X_t$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

when m tends to infinity.

The induction principle ensures that the calculation above can be extended to any finite number of increments. To see this, add a third increment

$$\mathbb{P}\big(X^m_{\Delta^m(s+t+u)} - X^m_{\Delta^m(s+t)} \leqslant z\,,\, X^m_{\Delta^m(s+t)} - X^m_{\Delta^m(s)} \leqslant y\,,\, X^m_{\Delta^m(s)} \leqslant x\big)$$

and condition on all possible values for $\eta^m_{\Delta^m(s)}$ and $\eta^m_{\Delta^m(s+t)}$. In contrast to the original process $\{\eta^m_t\}$, the random time process $\{\eta^m_{\Delta^m(t)}\}$ has the Markov property. The third (conditioned) increment can therefore be rewritten as

$$\begin{aligned}&\mathbb{P}\big(X^m_{\Delta^m(s+t+u)} - X^m_{\Delta^m(s+t)} \leqslant z \mid \eta^m_{\Delta^m(s)} = k\,,\, \eta^m_{\Delta^m(s+t)} = i\big)\\ &\quad = \mathbb{P}\big(X^m_{\Delta^m(s+t+u)} - X^m_{\Delta^m(s+t)} \leqslant z \mid \eta^m_{\Delta^m(s+t)} = i\big)\\ &\quad = \cdots = \mathbb{P}\big(X_u \leqslant z\big) + o(1).\end{aligned}$$

with the technique constituted by (1) and (2). The rest of the proof relies on induction since it only remains to show that

$$\mathbb{P}\big(X^m_{\Delta^m(s+t)} - X^m_{\Delta^m(s)} \leqslant y\,,\, X^m_{\Delta^m(s)} \leqslant x\big) \to \mathbb{P}\big(X_t \leqslant y\big)\,\mathbb{P}\big(X_s \leqslant x\big)$$

when m tends to infinity. The procedure can be repeated by adding new increments, one by one, and then resolving the calculation using induction.

Thus, $\{X^m_{\Delta^m(t)}\}$ has independent increments in the limit. □

4. Critical displacements in almost full hash tables

Every item in the hash table $\mathcal{T}'_{m,m-1}$ is categorized to have either a *critical* or a *non-critical* displacement. As mentioned in the Introduction, item x_i has a critical displacement d_i if it is of such magnitude that

$$d_i \geqslant \zeta(m),$$

for some function $\zeta(m)$. Although the choice of $\zeta(m)$ will first become evident in the proof of Lemma 4.2, it will be defined here for a better overview:

$$\zeta(m) = m - z\,r(m)^2, \quad z > 0. \tag{4.1}$$

Note that $r(m)$ is the same function that is used to scale and normalize the fragmentation process X^m_t, a function that from now on is set to

$$r(m) = m^{1/3}. \tag{4.2}$$

Again, the explanation for this choice will be postponed to the proof of Lemma 4.2. With $r(m)$ fixed, the constant z is left to provide some freedom to the concept of critical displacements (if viewed as a parameter).

Let $\{U^m_t(z)\}$ be a sequence of counting processes defined by

$$\begin{aligned}U^m_t(z) = &\text{'The number of items removed from } \mathcal{T}'_{m,m-1} \text{ in the}\\ &\text{time interval } (0,\, t] \text{ that has a critical displacement.'}\end{aligned}$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

Thus, every process $U_t^m(z)$ is associated to the corresponding fragmentation process X_t^m of table $\mathcal{T}'_{m,m-1}$. It is convenient to think of X_t^m and $U_t^m(z)$ as a pair of processes,

$$\big(X_t^m,\, U_t^m(z)\big),$$

where $U_t^m(z)$ depends on the path of X_t^m.

Theorem 4.1. *For any fixed $z > 0$, there exists a limit process $U_t(z)$ such that the pair*

$$\big(X_t^m,\, U_t^m(z)\big) \xrightarrow{\mathrm{d}} \big(X_t,\, U_t(z)\big) \quad \text{as } m \to \infty,$$

in the usual Skorokhod topology on $\mathcal{D}[0,\infty)$. Moreover, conditioned on the path of the limit fragmentation process X_t, the limit counting process $U_t(z)$ is an inhomogeneous Poisson process with intensity

$$(z - X_t)_+, \quad t \geqslant 0.$$

Proof. Distributional process convergence of the sequence $\{X_t^m\}$ to the stable subordinator of index $1/2$, in the usual Skorokhod topology on $\mathcal{D}[0,\infty)$, has already been established in Theorem 3.2,

$$X_t^m \xrightarrow{\mathrm{d}} X_t \quad \text{as } m \to \infty.$$

By the Skorokhod representation theorem there exists a process $\hat{X}_t$ and a sequence of processes $\{\hat{X}_t^m\}$ such that

$$\hat{X}_t \overset{\mathrm{d}}{=} X_t, \quad \hat{X}_t^m \overset{\mathrm{d}}{=} X_t^m \quad \text{for } m \geqslant 1$$

and

$$\hat{X}_t^m \xrightarrow{\text{a.s.}} \hat{X}_t \quad \text{as } m \to \infty.$$

Let $\{\hat{U}_t^m(z)\}$ define the sequence attained by conditioning $\{U_t^m(z)\}$ on the paths of $\hat{X}_t$ and $\{\hat{X}_t^k\}_{k\geqslant 1}$,

$$\hat{U}_t^m(z) = \big(U_t^m(z) \,|\, \hat{X}_t,\, \{\hat{X}_t^k\}_{k\geqslant 1}\big) \quad \text{for } m \geqslant 1.$$

Since

$$\hat{U}_t^m(z) \overset{\mathrm{d}}{=} \big(U_t^m(z) \,|\, X_t^m\big),$$

it is enough to show that $\{\hat{U}_t^m(z)\}$ converges to an inhomogeneous Poisson process with intensity

$$(z - \hat{X}_t)_+, \quad t \geqslant 0, \tag{4.3}$$

to prove the theorem.

Now, recall that the random time-deformation $\Delta^m(t)$ is defined as the minimum time needed to remove $\lceil t\, r(m) \rceil$ marked items. Since only marked items can have a critical displacement there is reason to substitute t with $\Delta^m(t)$ again (as was done in the proof of Theorem 3.2). By Lemma 3.3, $\{\hat{U}_t^m(z)\}$ and $\{\hat{U}_{\Delta^m(t)}^m(z)\}$ have the same limit process

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

(provided a limit process exists) in the usual Skorokhod topology on $\mathcal{D}[0,\infty)$ when m tends to infinity.

Three important steps remain before the proof is complete. Similarly to the proof of Theorem 3.2, the principal method is presented here and the technical parts are proved separately in the three consecutive lemmas below. The first step is to prove that for any fixed time t, the sequence of random variables $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ converges in distribution to a

$$\mathrm{Po}\left(\int_0^t (z-\hat{X}_s)_+ \,\mathrm{d}s\right)$$

distributed random variable when m tends to infinity (Lemma 4.2). The second step is to prove that $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ is tight (Lemma 4.3), and the third that the finite-dimensional distributions converge (Lemma 4.4). □

Lemma 4.2. *Given the paths of $\hat{X}_t$ and $\{\hat{X}^k_t\}_{k\geqslant 1}$, the conditional sequence $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ converges in distribution to a*

$$\mathrm{Po}\left(\int_0^t (z-\hat{X}_s)_+ \,\mathrm{d}s\right)$$

distributed random variable, for any fixed time $t>0$, as m tends to infinity.

Proof. Exactly $\lceil t\,r(m)\rceil$ marked items have been removed from $\mathcal{T}'_{m,m-1}$ at time $\Delta^m(t)$. Relabel these marked items $\{x^m_1,\ldots,x^m_{\lceil t\,r(m)\rceil}\}$ according to the order of removal and let $\mathcal{D}^m_k$ denote the event that item x^m_k has a critical displacement.

Recall that an item is marked if it belongs to the largest block at the time of its removal. Now, consider the original direction of time, *i.e.*, items are hashed into the table (instead of removed from the final table). It is not certain that a marked item targeted the currently largest block when it was hashed into the table. The item may have targeted the block to the left of the largest block (which coalesces the two blocks into one) and thus become a marked item.

Let $0<s\leqslant t$ and consider item $x^m_{\lceil s\,r(m)\rceil}$, *i.e.*, the last marked item in the time interval $(0,\Delta^m(s)]$. Let $\Delta^m(s-)$ denote the time when exactly $\lceil s\,r(m)\rceil-1$ marked items have been removed. Thus, the size of the largest block decreases from $\eta^m_{\Delta^m(s-)}$ to $\eta^m_{\Delta^m(s)}$ when $x^m_{\lceil s\,r(m)\rceil}$ is removed. The probability that $x^m_{\lceil s\,r(m)\rceil}$ targeted the largest block when it was hashed into the table (and thereby assimilated the block to the right) is proportional to the size of the largest block before insertion, *i.e.*,

$$\frac{\eta^m_{\Delta^m(s)}}{\eta^m_{\Delta^m(s-)}}.$$

In order to get a critical displacement item $x^m_{\lceil s\,r(m)\rceil}$ must not only target the largest block when inserted but it must target the very beginning of the largest block. Consequently, the probability that the last-removed item in the time interval $(0,\Delta^m(s)]$ has a critical

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

displacement is given by

$$\mathbb{P}\big(\mathcal{D}^m_{\lceil s\,r(m)\rceil}\big) = \frac{\big(\eta^m_{\Delta^m(s)} - \zeta(m)\big)_+}{\eta^m_{\Delta^m(s-)}} = \frac{r(m)^2\,\big(z - \hat{X}^m_{\Delta^m(s)}\big)_+}{m - r(m)^2\,\hat{X}^m_{\Delta^m(s-)}}$$
$$= \frac{r(m)^2}{m}\big(z - \hat{X}^m_{\Delta^m(s)}\big)_+\left(1 + O\left(\frac{r(m)^2}{m}\right)\right).$$

The second step motivates equation (4.1), where $\zeta(m)$ is defined in terms of the function $r(m)$, and the last step holds since the path of $\hat{X}^m_{\Delta^m(t)}$ is given.

Note that $\mathbb{P}\big(\mathcal{D}^m_{\lceil s\,r(m)\rceil}\big)$ only depends on the value of $\hat{X}^m_t$ at the time points $\Delta^m(s)$ and $\Delta^m(s-)$. Since $\hat{U}^m_t$ is defined as a counting process conditioned on the path of $\hat{X}^m_t$, all events $\mathcal{D}^m_1, \ldots, \mathcal{D}^m_{\lceil t\,r(m)\rceil}$ are independent of each other. The number of critical displacements can therefore be written as a sum of independent indicator variables

$$\hat{U}^m_{\Delta^m(t)}(z) = \sum_{k=1}^{\lceil t\,r(m)\rceil} \mathbf{1}_{\{\mathcal{D}^m_k\}}.$$

This sum suggests the use of the Poisson 'law of small numbers'. Define

$$\mathcal{A}^m_t(z) = \sum_{k=1}^{\lceil t\,r(m)\rceil} \mathbb{P}\big(\mathcal{D}^m_k\big) = \int_0^{t\,r(m)} \mathbb{P}\big(\mathcal{D}^m_{\lceil y\rceil}\big)\,\mathrm{d}y + O\left(\frac{r(m)^2}{m}\right)$$
$$= \{y \mapsto s\,r(m)\} = \int_0^t r(m)\;\mathbb{P}\big(\mathcal{D}^m_{\lceil s\,r(m)\rceil}\big)\,\mathrm{d}s + O\left(\frac{r(m)^2}{m}\right)$$
$$= \frac{r(m)^3}{m}\left(1 + O\left(\frac{r(m)^2}{m}\right)\right)\int_0^t \big(z - \hat{X}^m_{\Delta^m(s)}\big)_+\,\mathrm{d}s + O\left(\frac{r(m)^2}{m}\right)$$
$$= \big(1 + O(m^{-1/3})\big)\int_0^t \big(z - \hat{X}^m_{\Delta^m(s)}\big)_+\,\mathrm{d}s + O(m^{-1/3}).$$

The last step motivates equation (4.2) where the normalization and scaling function $r(m)$ is set to $m^{1/3}$.

Let $\mathcal{A}_t(z)$ denote the parameter for the hypothetical Poisson limit distribution. By the dominated convergence theorem

$$\mathcal{A}_t(z) = \lim_{m\to\infty} \mathcal{A}^m_t(z) = \int_0^t (z - \hat{X}_s)_+\,\mathrm{d}s,$$

since the integrands are dominated by z for all $m \geqslant 1$ and

$$\hat{X}^m_{\Delta^m(s)} \xrightarrow{\text{a.s.}} \hat{X}_s \quad \text{as } m \to \infty.$$

Let $\mathcal{L}(\hat{U}^m_{\Delta^m(t)})$ denote the distribution of the random variable $\hat{U}^m_{\Delta^m(t)}$. The total variation distance between $\mathcal{L}(\hat{U}^m_{\Delta^m(t)})$ and $\mathrm{Po}(\mathcal{A}_t(z))$ can be estimated by Le Cam's inequality, which

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

is labelled (1) below (see Barbour, Holst and Janson [1]):

$$d_{TV}\big(\mathcal{L}(\hat{U}^m_{\Delta^m(t)}),\, \mathrm{Po}(\mathcal{A}_t(z))\big) \overset{(1)}{\leqslant} \sum_{k=1}^{\lceil t\, r(m)\rceil} \mathbb{P}\big(\mathcal{D}^m_k\big)^2 \leqslant \lceil t\, r(m)\rceil\, \mathbb{P}\big(\mathcal{D}^m_1\big)^2$$
$$\leqslant \lceil t\, r(m)\rceil \left(\frac{r(m)^2}{m} z\right)^2 \leqslant \frac{(t+1)\, z^2}{m^{1/3}} \to 0 \quad \text{as } m \to \infty.$$

Convergence in total variation implies distributional convergence and the proof is complete. □

Lemma 4.3. *Given the paths of $\hat{X}_t$ and $\{\hat{X}^k_t\}_{k\geqslant 1}$, the conditional sequence $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ is tight.*

Proof. Tightness can be proved via equicontinuity and Aldous's tightness criterion: see Kallenberg [14, Theorem 16.11]. That is, $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ is tight if, for any $\varepsilon > 0$,

$$\mathbb{P}\big(|\hat{U}^m_{\Delta^m(\tau_m+h_m)}(z) - \hat{U}^m_{\Delta^m(\tau_m)}(z)| > \varepsilon\big) \to 0 \quad \text{as } m \to \infty,$$

where $\{\tau_m\}$ is any bounded sequence of stopping times and $\{h_m\}$ any sequence of positive constants such that $h_m \to 0$ as m tends to infinity. This is indeed the case since

$$\begin{aligned}\mathbb{P}\big(|\hat{U}^m_{\Delta^m(\tau_m+h_m)}(z) - \hat{U}^m_{\Delta^m(\tau_m)}(z)| > \varepsilon\big) &\leqslant 1 - \big(1 - \mathbb{P}\big(\mathcal{D}^m_1\big)\big)^{\lceil h_m\, r(m)\rceil} \\ &= 1 - \exp\{\lceil h_m\, r(m)\rceil \ln\big(1 - \mathbb{P}\big(\mathcal{D}^m_1\big)\big)\} \\ &= 1 - \exp\{-h_m\, r(m)\, \mathbb{P}\big(\mathcal{D}^m_1\big) + o(1)\} \\ &= 1 - \exp\{-h_m\, z + o(1)\} \to 0 \quad \text{as } m \to \infty.\end{aligned}$$

□

Lemma 4.4. *Given the paths of $\hat{X}_t$ and $\{\hat{X}^k_t\}_{k\geqslant 1}$, the finite-dimensional distributions of the conditional sequence $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ converge.*

Proof. All processes in the sequence $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ have independent increments since

$$\hat{U}^m_{\Delta^m(t)}(z) = \sum_{k=1}^{\lceil t\, r(m)\rceil} \mathbf{1}_{\{\mathcal{D}^m_k\}},$$

and all events $\mathcal{D}^m_1, \ldots, \mathcal{D}^m_{\lceil t\, r(m)\rceil}$ only depend on the path of $\hat{X}^m_t$. Independent increments ensure that all linear combinations converge and consequently that the finite-dimensional distributions of $\{\hat{U}^m_{\Delta^m(t)}(z)\}$ converge. □

Let $\{U^m(z)\}$ be the sequence of counting processes defined by

$$U^m(z) = \text{'The total number of items with critical displacements in the table } \mathcal{T}'_{m,m-1}\text{.'}$$

Theorem 4.5. *There exists a limit variable $U(z)$ such that*

$$U^m(z) \xrightarrow{\mathrm{d}} U(z) \quad \text{as } m \to \infty.$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

Moreover, conditioned on the path of the limit fragmentation process X_t, the limit variable $U(z)$ is Poisson-distributed with parameter

$$\mathcal{A}(z) = \int_0^\infty (z - X_t)_+ \, \mathrm{d}t.$$

Proof. Let T be the stopping time when the intensity for critical displacements of the limit process $U_t(z)$ becomes zero, *i.e.*,

$$T = \inf\{t > 0 \, : \, X_t > z\}.$$

Conditioned on the path of X_t, the limit process $U_t(z)$ should manage to count the total number of critical displacements only if it is run for long enough. Thus, provided that $T < \infty$, then for all $t \geqslant T$

$$\big(U_t(z) \,|\, X_t\big) = \big(U_T(z) \,|\, X_t\big) \in \mathrm{Po}\big(\mathcal{A}_T(z)\big) = \mathrm{Po}\big(\mathcal{A}(z)\big).$$

Consequently, if there exists a limit variable $U(z)$ to $\{U^m(z)\}$ it must be $\mathrm{Po}\big(\mathcal{A}(z)\big)$-distributed, when conditioned on the path of X_t.

Since the sequences $\{U^m(z)\}$ and $\{\lim_{t\to\infty} U_t^m(z)\}$ are equivalent, a limit variable $U(z)$ exists if the order in which m and t tend to infinity can be interchanged. A sufficient condition for this is given by

$$\lim_{t\to\infty} \limsup_{m\to\infty} \mathbb{P}\big(|U^m(z) - U_t^m(z)| > \varepsilon\big) = 0$$

(see Billingsley [2, Theorem 3.2]). This is indeed the case since

$$\begin{aligned} \mathbb{P}\big(|U^m(z) - U_t^m(z)| > \varepsilon\big) &\leqslant \mathbb{P}\big(X_t^m < z\big) \\ &\to \mathbb{P}\big(X_t < z\big) \quad \text{as } m \to \infty \\ &\to 0 \quad \text{as } t \to \infty. \end{aligned}$$

□

The parameter $\mathcal{A}(z)$ can be visualized as the random area bounded by the y-axis and the two graphs $y(t) = z$ and $y(t) = X_t$. Alternatively, by exchanging coordinate axis, $\mathcal{A}(z)$ can be visualized as the random area covered by the inverse process

$$S_\tau = \inf\{t > 0 \, : \, X_t > \tau\},$$

in the 'time interval' $0 < \tau \leqslant z$, that is,

$$\mathcal{A}(z) = \int_0^z S_\tau \, \mathrm{d}\tau.$$

This inverse process of the stable subordinator of index $1/2$ has the same distribution as the supremum process of standard Brownian motion B_s,

$$S_\tau \overset{\mathrm{d}}{=} \max_{0\leqslant s\leqslant \tau} B_s, \quad \tau \geqslant 0$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

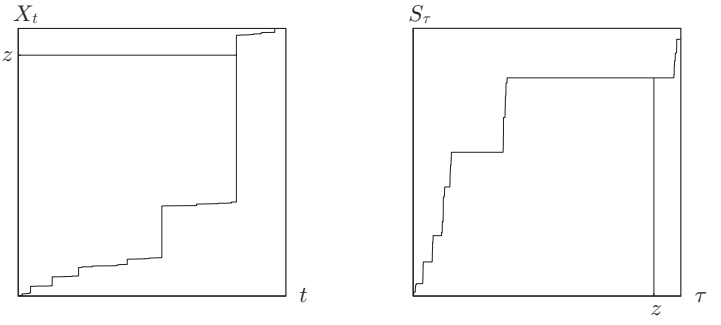


Figure 1. Two ways to visualize the random area $\mathcal{A}(z)$.

(see for instance Karatzas and Shreve [15, Chapters 2.8 and 6.2]). For any fixed $c > 0$, the usual Brownian scaling

$$B_{c\tau} \overset{\mathrm{d}}{=} c^{1/2} B_\tau, \quad \tau \geqslant 0,$$

implies that

$$S_{c\tau} \overset{\mathrm{d}}{=} c^{1/2} S_\tau, \quad \tau \geqslant 0.$$

For ease of notation, let $\mathcal{A} := \mathcal{A}(1)$. By Brownian scaling,

$$\mathcal{A}(z) = \int_0^z S_\tau \,\mathrm{d}\tau = z \int_0^1 S_{z\tau} \,\mathrm{d}\tau \overset{\mathrm{d}}{=} z^{3/2} \int_0^1 S_\tau \,\mathrm{d}\tau \overset{\mathrm{d}}{=} z^{3/2} \mathcal{A},$$

and it is enough to consider $\mathcal{A}$.

Properties of $\mathcal{A}$ were studied separately in [12] by the author and Svante Janson. The most important results for this paper are reproduced here.

Theorem 4.6. *The nth moment of $\mathcal{A}$ is*

$$\mathbb{E}\,\mathcal{A}^n = \frac{n!\,\Gamma(n+2/3)}{\Gamma(2/3)\,\Gamma(3n/2+1)} \left(\frac{3\sqrt{2}}{4}\right)^n, \quad n \in \mathbb{N}. \qquad \square$$

Theorem 4.7. *The Laplace transform of $\mathcal{A}$ is*

$$\psi(s) = \frac{1}{\Gamma(2/3)} \sum_{n=0}^{\infty} \frac{\Gamma(n+2/3)}{\Gamma(3n/2+1)} \left(\frac{-3\sqrt{2}\,s}{4}\right)^n, \quad s \in \mathbb{C}. \qquad \square$$

Theorem 4.8. *Let ψ be the Laplace transform of $\mathcal{A}$. For all $\alpha, \lambda > 0$, the (so-called) double Laplace transform of $\mathcal{A}$ is given by*

$$\int_0^\infty \psi\left(\alpha s^{3/2}\right) e^{-\lambda s} \,\mathrm{d}s = \int_0^\infty \left(1 + \frac{3\alpha s}{2\sqrt{2\lambda}}\right)^{-2/3} e^{-\lambda s} \,\mathrm{d}s. \qquad \square$$

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

For the proofs of Theorem 4.6–4.8, see [12].

Theorem 4.9. *The total number of items with a critical displacement is given by the probability function*

$$\mathbb{P}\big(U(z)=k\big)=\frac{(-z^{3/2})^k}{k!}\,\psi^{(k)}(z^{3/2}),\quad k\in\mathbb{N},$$

where ψ is the Laplace transform of $\mathcal{A}$. In particular,

$$\mathbb{P}\big(U(z)=0\big)=\psi(z^{3/2}).$$

Proof.

$$\begin{aligned}\mathbb{P}\big(U(z)=k\big)&=\int_0^\infty \mathbb{P}\big(U(z)=k\,|\,\mathcal{A}(z)=x\big)\,\mathrm{d}F_{\mathcal{A}(z)}(x)\\&=\int_0^\infty\frac{x^k}{k!}e^{-x}\,\mathrm{d}F_{\mathcal{A}(z)}(x)=\{x\mapsto x\,z^{3/2}\}\\&=\frac{(z^{3/2})^k}{k!}\int_0^\infty x^k e^{-x\,z^{3/2}}\,\mathrm{d}F_{\mathcal{A}}(x)\\&=\frac{(-z^{3/2})^k}{k!}\,\psi^{(k)}(z^{3/2})\end{aligned}$$

□

Corollary 4.10. *The expectation and variance of the number of critical displacements are given by*

$$\begin{aligned}\mathbb{E}\big[U(z)\big]&=\frac{4}{3\sqrt{2\pi}}\,z^{3/2},\\\mathrm{Var}\big[U(z)\big]&=\frac{4}{3\sqrt{2\pi}}\,z^{3/2}+\left(\frac{5}{12}-\frac{8}{9\pi}\right)z^3.\end{aligned}$$

Proof. Since $U(z)\in\mathrm{Po}\,\big(\mathcal{A}(z)\big)$, it follows from Theorem 4.6 that

$$\mathbb{E}\big[U(z)\big]=\mathbb{E}\big[\mathbb{E}\big[U(z)\,|\,\mathcal{A}(z)\big]\big]=\mathbb{E}\big[\mathcal{A}(z)\big]=\frac{4}{3\sqrt{2\pi}}\,z^{3/2}$$

and

$$\begin{aligned}\mathrm{Var}\big[U(z)\big]&=\mathbb{E}\big[\mathrm{Var}\big[U(z)\,|\,\mathcal{A}(z)\big]\big]+\mathrm{Var}\big[\mathbb{E}\big[U(z)\,|\,\mathcal{A}(z)\big]\big]\\&=\mathbb{E}\big[\mathcal{A}(z)\big]+\mathrm{Var}\big[\mathcal{A}(z)\big]=\frac{4}{3\sqrt{2\pi}}\,z^{3/2}+\left(\frac{5}{12}-\frac{8}{9\pi}\right)z^3.\end{aligned}$$

□

5. The maximum displacement in almost full hash tables

In order to find the limit distribution for the maximum displacement in an almost full hash table $\mathcal{T}'_{m,m-1}$, the random variable $d_{\max}$ must be properly normalized. The definition

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

of a critical displacement, *i.e.*,

$$\text{'item } x_i \text{ has a critical displacement'} \quad \Leftrightarrow \quad d_i \geqslant m - z\,m^{2/3},$$

motivates the following normalization for the maximum displacement,

$$Y^m = \frac{m - d_{\max}}{m^{2/3}}.$$

Define Y as a random variable with distribution function

$$\mathbb{P}\big(Y \leqslant z\big) = 1 - \psi(z^{3/2}), \quad 0 < z < \infty,$$

where ψ is the Laplace transform of $\mathcal{A}$ given by Theorem 4.7.

Theorem 5.1. *The maximum displacement Y^m in an almost full hash table $\mathcal{T}'_{m,m-1}$ converges in distribution to Y,*

$$Y^m \xrightarrow{\mathrm{d}} Y \quad \text{as } m \to \infty.$$

Proof. From the definition of critical displacements it follows that

$$Y^m \leqslant z \quad \Leftrightarrow \quad U^m(z) > 0,$$

for any positive z. Thus,

$$\mathbb{P}\big(Y^m \leqslant z\big) = \mathbb{P}\big(U^m(z) > 0\big) \to \mathbb{P}\big(U(z) > 0\big) = 1 - \psi(z^{3/2})$$

as m tends to infinity. □

Theorem 5.2. *The limit random variable Y has expectation*

$$\mathbb{E}\,Y = \frac{2\,\Gamma(1/3)}{3^{2/3}}$$

but higher-order moments diverge.

Proof. Let $k \in \mathbb{N} \setminus \{0\}$. The expectation is given by

$$\begin{aligned}
\mathbb{E}\,Y &= \int_0^\infty \mathbb{P}\big(Y > z\big)\,\mathrm{d}z = \int_0^\infty \psi(z^{3/2})\,\mathrm{d}z \\
&\overset{(1)}{=} \lim_{k\to\infty} \int_0^\infty \psi(z^{3/2}) e^{-z/k}\,\mathrm{d}z \\
&\overset{(2)}{=} \lim_{k\to\infty} \int_0^\infty \left(1 + \frac{3\sqrt{k}}{2\sqrt{2}}\,z\right)^{-2/3} e^{-z/k}\,\mathrm{d}z \\
&\overset{(3)}{=} \lim_{a\to 0+} \frac{2e^a}{3^{2/3}} \int_a^\infty z^{-2/3} e^{-z}\,\mathrm{d}z \\
&= \frac{2}{3^{2/3}} \int_0^\infty z^{-2/3} e^{-z}\,\mathrm{d}z = \frac{2\,\Gamma(1/3)}{3^{2/3}}.
\end{aligned}$$

Here (1) follows from the monotone convergence theorem, where $\{\psi(z^{3/2})e^{-z/k}\}$ is a non-decreasing sequence of functions converging to $\psi(z^{3/2})$; (2) follows by the double Laplace

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

transform of $\mathcal{A}$, see Theorem 4.8; and (3) follows by a change of variables. Let $a = \frac{2\sqrt{2}}{3k\sqrt{k}}$ and make the substitution $\frac{z}{k} \mapsto (z - a)$. Moments of higher order diverge since

$$\begin{aligned}
\mathbb{E}\, Y^2 &= 2\int_0^\infty z\, \mathbb{P}\big(Y > z\big)\, \mathrm{d}z = 2\int_0^\infty z\, \psi(z^{3/2})\, \mathrm{d}z \\
&\overset{(4)}{=} \lim_{k\to\infty} 2k \int_0^\infty \psi(z^{3/2})\big(1 - e^{-z/k}\big)\, \mathrm{d}z \\
&\overset{(5)}{=} \lim_{a\to 0+} \frac{4}{(3a)^{2/3}} \left(\frac{2\,\Gamma(1/3)}{3^{2/3}} - \frac{2e^a}{3^{2/3}} \int_a^\infty z^{-2/3} e^{-z}\, \mathrm{d}z \right) \\
&\overset{(6)}{=} \lim_{a\to 0+} 2(3a)^{1/3} \left(-\frac{2e^a}{3^{2/3}} \int_a^\infty z^{-2/3} e^{-z}\, \mathrm{d}z + \frac{2}{(3a)^{2/3}} \right) \\
&= \infty.
\end{aligned}$$

Here (4) by the monotone convergence theorem, where $\{k\,\psi(z^{3/2})(1 - e^{-z/k})\}$ is a non-decreasing sequence of functions converging to $z\,\psi(z^{3/2})$; (5) follows from (2) and (3) together with the fact that $\mathbb{E}\, Y$ is known; and (6) follows by l'Hôpital's rule. □

Acknowledgements

Many thanks to Svante Janson for helpful discussions and valuable advice during the course of this project. I thank an anonymous referee for careful comments, and *Combinatorics, Probability and Computing* for their patience and understanding.†

References

[1] Barbour, A. D., Holst, L. and Janson, S. (2003) *Poisson Approximation*, first edition, Oxford University Press.

[2] Billingsley, P. (1999) *Convergence of Probability Measures*, second edition, Wiley.

[3] Celis, P., Larson, P-Å. and Munro, J. I. (1985) Robin Hood hashing. In *26th Annual Symposium on Foundations of Computer Science*, IEEE Computer Society, pp. 281–288.

[4] Chassaing, P. and Janson, S. (2001) A Vervaat-like path transformation for the reflected Brownian bridge conditioned on its local time at 0. *Ann. Probab.* **29** 1755–1779.

[5] Chassaing, P. and Louchard, G. (2002) Phase transition for parking blocks, Brownian excursion and coalescence. *Random Struct. Alg.* **21** 76–119.

[6] Chassaing, P. and Marckert, J.-F. (2001) Parking functions, empirical processes, and the width of rooted labeled trees. *Electron. J. Combin.* **8** #R14.

[7] Devroye, L. (1985) The expected length of the longest probe sequence when the distribution is not uniform. *J. Algorithms* **6** 1–9.

[8] Devroye, L., Morin, P. and Viola, A. (2004) On worst-case Robin Hood hashing. *SIAM. J. Comput.* **33** 923–936.

[9] Flajolet, P., Poblete, P. and Viola, A. (1998) On the analysis of linear probing hashing. *Algorithmica* **22** 490–515.

[10] Gonnet, H. (1981) Expected length of the longest probe sequence in hash code searching. *J. Assoc. Comput. Mach.* **28** 289–304.

† The author has suffered from health problems and the publication of this paper has been delayed several times.

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press

[11] Janson, S. (2005) Individual displacements for linear probing hashing with different insertion policies. *ACM Trans. Algorithms* **1** 177–213.
[12] Janson, S. and Petersson, N. (2009) The integral of the supremum process of Brownian motion. *J. Appl. Probab.* **46** 593–600.
[13] Johnson, N. L., Kemp, A. W. and S. Kontz, S. (2005) *Univariate Discrete Distributions*, third edition, Wiley,
[14] Kallenberg, O. (2002) *Foundations of Modern Probability*, second edition, Springer.
[15] Karatzas, I. and Shreve, S. E. (2004) *Brownian Motion and Stochastic Calculus*, second edition, Springer.
[16] Knuth, D. E. (1998) *The Art of Computer Programming*, Vol. 3, *Sorting and Searching*, second edition, Addison-Wesley.
[17] Pavlov, Y. L. (2000) *Random Forests*, first edition, VSP.
[18] Perman, M. and Wellner, J. A. (1996) On the distribution of Brownian areas. *Ann. Appl. Probab.* **6** 1091–1111.
[19] Poblete, P. V. and Munro, J. I. (1989) Last-come-first-served hashing. *J. Algorithms* **10** 228–248.
[20] Poblete, P. V., Viola, A. and Munro, J. I. (1997) The diagonal Poisson transform and its application to the analysis of a hashing scheme. *Random Struct. Alg.* **10** 221–255.
[21] Viola, A. (2005) Exact distributions of individual displacements in linear probing hashing. *ACM Trans. Algorithms* **1** 214–242.

https://doi.org/10.1017/S0963548312000582 Published online by Cambridge University Press