SIAM J. COMPUT.
Vol. 48, No. 2, pp. 736–741

LINEAR HASHING IS AWESOME[*]

MATHIAS BÆK TEJS KNUDSEN[†]

Abstract. We consider the hash function $h(x) = ((ax + b) \bmod p) \bmod n$ where a, b are chosen uniformly at random from $\{0, 1, \ldots, p-1\}$. We prove that when we use $h(x)$ in hashing with chaining to insert n elements into a table of size n the expected length of the longest chain is $\tilde{O}(n^{1/3})$. The proof also generalizes to give the same bound when we use the multiply-shift hash function by Dietzfelbinger et al. [*J. Algorithms*, 25 (1997), pp. 19–51].

Key words. hashing, linear hashing, hashing with chaining

AMS subject classifications. 68Q87, 68W20, 68W40, 68Q25, 68P05

DOI. 10.1137/17M1126801

1. Introduction.

In this paper, we study the hash function $h : [p] \to [m]$ (where $[m] = \{0, 1, \ldots, m-1\}$) defined by $h(x) = ((ax + b) \bmod p) \bmod m$, where $a, b \in [p]$ are chosen uniformly at random from $[p]$. The function $h(x)$ was first introduced by Carter and Wegman [2]. Here, p is a prime and $p \geq m$. We assume that we have a set $X \subset [p]$ of n *keys* with $n \leq m$ and use h to assign a hash value $h(x)$ to each key $x \in X$. We are interested in the frequency of the most popular hash value, i.e., we study the random variable $M(h, X)$ defined by

$$M(h, X) = \max_{y \in [m]} |\{x \in X \mid h(x) = y\}| \ . \tag{1}$$

In Theorem 3.3, we prove that $\mathrm{E}[M(h, X)] = O\big(\sqrt[3]{n \log n}\big)$. We also consider the hash function $\bar{h} : [q] \to [m]$ defined by $\bar{h}(x) = \big\lfloor \frac{(ax) \bmod q}{q/m} \big\rfloor$, where q, m are powers of 2, $q \geq m \geq n$, and a is chosen uniformly at random among the odd numbers from $[q]$. The function $\bar{h}(x)$ was first introduced by Dietzfelbinger et al. [3]. In Theorem 3.4 we prove that it also holds that $\mathrm{E}\big[M(\bar{h}, X)\big] = O\big(\sqrt[3]{n \log n}\big)$.

We note that when we use $h(x) = ((ax+b) \bmod p) \bmod m$ in hashing with chaining, M is the size of the largest chain. When scanning the hash table for an element the expected time used is $O(1)$ and the worst case time is at most $O(M(h, X))$.

1.1. Related work.

It is folklore that the size of the largest chain is $O(\sqrt{n})$. Alon et al. [1] consider the linear hash function $h_{m,k} : \mathcal{F}^m \to \mathcal{F}^k$, where $\mathcal{F}$ is a finite field and $n = |\mathcal{F}|^k$. The function is defined by $h_{m,k}(x_1, \ldots, x_m) = \sum_i x_i a_i$, where $a_i \in \mathcal{F}^k$ is chosen uniformly at random. For $m = 2, k = 1$, the hash function is $h_{2,1}(x, y) = ax + by$, where $a, b \in \mathcal{F}$ are chosen uniformly at random. It is shown in [1] that there exists a set $X \subset \mathcal{F}^2$ such that $\mathrm{E}[M(h_{2,1}, X)] > \sqrt{n}$ if n is a square and $\mathrm{E}[M(h_{2,1}, X)] = \Omega(\sqrt[3]{n})$ if n is a prime power that is not a square. In [1] it is also shown that when $\mathcal{F}$ is the field of two elements the expected length of the longest chain is $O(\log n \log \log n)$, improving the results in [4, 5].

[*] Received by the editors April 21, 2017; accepted for publication (in revised form) January 15, 2019; published electronically April 30, 2019. A preliminary version of this work appeared at FOCS 2016.

http://www.siam.org/journals/sicomp/48-2/M112680.html

Funding: This work is partly supported by the FNU project AlgoDisc - Discrete Mathematics, Algorithms, and Data Structures, and by Mikkel Thorup's Advanced Grant DFF-0602-02499B from the Danish Council for Independent Research under the Sapere Aude research career programme.

[†] Department of Computer Science, University of Copenhagen, Copenhagen 2200, Denmark (mathias@tejs.dk).

2. Notation.

In the paper, we use the following conventions.

The integers are denoted by $\mathbb{Z}$. For an integer r we let $\mathbb{Z}_r = \mathbb{Z}/r\mathbb{Z}$ denote the residue classes $\bmod\ r$, and $\mathbb{Z}_r^*$ denote the set of elements of $\mathbb{Z}_r$ that have a multiplicative inverse. For an integer $u \in \mathbb{Z}$ we let $[u]_r \in \mathbb{Z}_r$ denote the residue class of u mod r. We let $\iota_r : \mathbb{Z}_r \to [r]$ be the unique mapping that satisfies $[\iota_r(s)]_r = s$ for all $s \in \mathbb{Z}_r$. For $s \in \mathbb{Z}_r$ we let $\|s\|_r = \min\{\iota_r(s), \iota_r(-s)\}$. An *interval* of $\mathbb{Z}_r$ is a nonempty subset of $\mathbb{Z}_r$ that can be written as $\{[u]_r, [u+1]_r, \ldots, [u+t-1]_r\}$ for a suitable choice of $u, t \in \mathbb{Z}$ with $t \geq 0$.

Throughout the paper, n, m, p, q, r are integers such that $p, q, r \geq m \geq n$. It is always the case that n is the size of a set, p is a prime, and q is a power of 2. All residue classes considered will be either mod p, mod q, or mod r.

The integers a, b are random variables that are chosen independently and uniformly at random from either $\mathbb{Z}_p$, $\mathbb{Z}_q$, or $\mathbb{Z}_r$.

The symbols C, D, E denote sets of residues, i.e., they are subsets of either $\mathbb{Z}_p$, $\mathbb{Z}_q$ or $\mathbb{Z}_r$, and c, d, e will be elements of these sets such that $c \in C$, $d \in D$, $e \in E$.

The set T is a subset of $\mathbb{Z}$ and its elements are denoted $t \in T$.

The hash function h is defined in one of the two following ways depending on the context. Either $h : \mathbb{Z}_p \to \mathbb{Z}_m$ is defined by

$$h(x) = [\iota_p(ax+b)]_m ,$$

which is a formal way of writing $h(x) = ((ax+b) \bmod p) \bmod m$, or $h : \mathbb{Z}_q \to [m]$ is defined by

$$h(x) = \left\lfloor \iota_q(ax) \cdot \frac{m}{q} \right\rfloor ,$$

which is a formal way of writing h as a multiply shift hash function.

The set X consists of n elements and is a subset of either $\mathbb{Z}_p$ or $\mathbb{Z}_q$ depending on how h is defined. The longest chain $M = M(X)$ is a random variable defined by

$$M = M(X) = \max_y |\{x \in X \mid h(x) = y\}| .$$

3. Main result.

In this section, we prove the main results of this paper, namely Theorems 3.3 and 3.4. The proofs of the two theorems are very similar and both rely on Lemma 3.2, which contains the main technical contribution of the paper. In preparation for the proof, we provide some elementary facts about intervals in $\mathbb{Z}_r$ in the following lemma.

LEMMA 3.1. *Let r, t, t' be positive integers that are pairwise coprime such that $r \geq t, t'$. Let I be an interval in $\mathbb{Z}_r$. Then,*

1. *$[t]_r^{-1} I$ is the disjoint union of t intervals, each of size $\left\lfloor \frac{|I|}{t} \right\rfloor$ or $\left\lceil \frac{|I|}{t} \right\rceil$.*
2. *If $|I|(t+t') < r$, then $\left([t]_r^{-1} I\right) \cap \left([t']_r^{-1} I\right)$ is either empty or an interval.*

Proof. Let $s = |I|$, and for simplicity assume that $I = \{[0]_r, [1]_r, \ldots, [s-1]_r\}$. If I is a different interval we can argue in a very similar manner.

For part 1, for each $j \in [t]$ we define $I_{t,j}$ by

$$\begin{aligned} I_{t,j} &= [t]_r^{-1} \left\{ [it+j]_r \mid i = 0, 1, \ldots, \left\lceil \frac{s-j}{t} \right\rceil - 1 \right\} \\ &= \left\{ [i]_r + [j]_r [t]_r^{-1} \mid i = 0, 1, \ldots, \left\lceil \frac{s-j}{t} \right\rceil - 1 \right\} , \end{aligned}$$

which is clearly an interval. Furthermore, it is obviously true that $t^{-1}I$ is the disjoint union of $I_{t,0}, \dots, I_{t,t-1}$.

Now we prove part 2. Assume that $[t]_r^{-1} I \cap [t']_r^{-1} I$ contains at least two elements since the statement is trivial otherwise. Say the elements are c_0, c_1 and we write c_0 as $[t]_r^{-1}[i_0']_r = c_0 = [t]_r^{-1}[i_0']_r$ and c_1 similarly. Then $[t'i_0 - ti_0']_r$ and $[t'i_1 - ti_1']_r$ are both equal $[0]_r$, and hence $t'(i_1 - i_0) - t(i_1' - i_0')$ is divisible by r. Since we have

$$|t'(i_1 - i_0) - t(i_1' - i_0')| \le t' |i_1 - i_0| + t |i_1' - i_0'| < (t' + t) |I| < r ,$$

we conclude that $t'(i_1 - i_0) - t(i_1' - i_0') = 0$. Since t and t' are co-prime, this implies that t divides $i_1 - i_0$. Hence $[t]_r^{-1}[i_0]_r$ and $[t]_r^{-1}[i_1]_r$ must both be in the same intervals $I_{t,0}, \dots, I_{t,t-1}$ defined in part 1. So there exist some $j \in [t]$ such that the intersection is $I_{t,j} \cap [t']_r^{-1} I$. By symmetry there also exist a $j' \in [t']$ such that the intersection is $I_{t,j} \cap I_{t',j'}$. Hence it must be an interval as desired. □

We will now prove the main technical lemma.

Lemma 3.2. *Let t, m, r be positive integers satisfying $4t \le m \le r$, and let $C \subset \mathbb{Z}_r$. Let $D \subset \mathbb{Z}_r^*$ be a set of size $\le t$ satisfying the following conditions:*

1. *$t < \iota(d) < 2t$ for all $d \in D$.*
2. *$\iota(d)$ and $\iota(d')$ are co-prime for every $d, d' \in D$ with $d \ne d'$.*

Assume that for every $d \in D$ there exists an interval I_d of size $\lceil \frac{r}{m} \rceil$ such that $I_d \cap dC$ contains at least $4t$ elements. Then there are at least $t|D|$ ordered pairs of different elements $c, c' \in C$ such that $\|c - c'\|_r < \frac{r}{mt}$.

Proof. We note that for every d the set $d^{-1}I_d$ is the union of $\iota(d)$ disjoint intervals of size $\le \lceil \frac{r}{m\iota(d)} \rceil$, and we write it as such a union $d^{-1}I_d = \bigcup_{j=0}^{\iota(d)-1} I_{d,j}$ as in Lemma 3.1. From Lemma 3.1 we also see that for any $d, d' \in D, d \ne d'$, the set $d^{-1}I_d \cap d'^{-1}I_{d'}$ is either empty or an interval, and in fact, there exists at most one index $j \in [\iota(d)]$ such that the intersection $I_{d,j} \cap d'^{-1}I_{d'}$ is nonempty. For every $d \in D$ and $j \in [\iota(d)]$, let $\delta(d,j)$ denote the number of elements $d' \in D$ such that $I_{d,j} \cap d'^{-1}I_{d'}$ is nonempty. Note that $\delta(d,j) \ge 1$ since $d \in D$. Furthermore, $\sum_{j=0}^{\iota(d)-1} \delta(d,j) \le |D| + \iota(d) < 3t$ since each $d'^{-1}I_{d'}$ has a nonempty intersection with at most one of the sets $I_{d,j}, j \in [\iota(d)]$.

The number of ordered pairs of different elements $(c, c') \in (C \cap I_{d,j})^2$ such that $\|c - c'\|_r < \frac{r}{mt}$ is exactly $|C \cap I_{d,j}| \cdot (|C \cap I_{d,j}| - 1)$ since $I_{d,j}$ is an interval of size $\le \lceil \frac{r}{m\iota(d)} \rceil$ and $\iota(d) > t$. Let $\tau(d,j) = \max\{0, |C \cap I_{d,j}| - 1\}$, then the number of such pairs is at least $(\tau(d,j))^2$. We can lower bound the number of such pairs in C by considering the pairs in $(C \cap I_{d,j})^2$ for each $d \in D$ and $j \in [\iota(d)]$ and note that each pair in $(C \cap I_{d,j})^2$ we count is counted at most $\delta(d,j)$ times. This gives that the number of ordered pairs $(c, c') \in C$ such that $\|c - c'\|_r < \frac{r}{mt}$ is at least

$$\sum_{d \in D} \sum_{j \in [\iota(d)]} \frac{(\tau(d,j))^2}{\delta(d,j)} . \tag{2}$$

For any $d \in D$, by the Cauchy–Schwarz inequality we have that

$$\left(\sum_{j \in [\iota(d)]} \delta(d,j) \right) \left(\sum_{j \in [\iota(d)]} \frac{(\tau(d,j))^2}{\delta(d,j)} \right) \ge \left(\sum_{j \in [\iota(d)]} \tau(d,j) \right)^2 . \tag{3}$$

We clearly have that $\sum_{j\in[\iota(d)]}\tau(d,j)\ge 4t-\iota(d)\ge 2t$. Also recall that we have that $\sum_{j\in[\iota(d)]}\delta(d,j)\le 3t$. Combining this with (2) and (3) gives that C contains at least $\frac{4t|D|}{3}\ge t\,|D|$ of the desired pairs. ◻

We now state and prove our main result.

THEOREM 3.3. *Let n,m,p be integers with p a prime and $p\ge m\ge n$. Let $X\subset\mathbb{Z}_p$ be a set of n elements. Let $h:\mathbb{Z}_p\to\mathbb{Z}_m$ be defined by $h(x)=[\iota_p(ax+b)]_m$ where $a,b\in\mathbb{Z}_p$ are chosen uniformly at random. Let $M=M(X)$ be the random variable counting the number of elements $x\in X$ that hash to the most popular hash value, that is,*

$$M=M(X)=\max_{y\in\mathbb{Z}_m}\left|\{x\in X\mid h(x)=y\}\right|\,.$$

Then

$$\mathrm{E}[M]=O\Big(\sqrt[3]{n\log n}\Big)\,. \tag{4}$$

Proof. We note that $\mathrm{E}[M\mid a=0]=n$ since h is constant when $a=0$. Therefore,

$$\mathrm{E}[M]=\frac{p-1}{p}\,\mathrm{E}[M\mid a\neq 0]+\frac{1}{p}\,\mathrm{E}[M\mid a=0]<\mathrm{E}[M\mid a\neq 0]+1\,.$$

From now on, we assume that a is chosen uniformly at random from $\mathbb{Z}_p^*$, and due to the calculation above it is clearly enough to bound the expected value of M in this case.

The random variables a and $a^{-1}b$ are independent. Note that $h(x)$ can be rewritten as $h(x)=\big[\iota_p\big(a(x+a^{-1}b)\big)\big]_m$. It clearly suffices to bound the expected value of M conditioned on all possible values $a^{-1}b$. For any fixed c, the expected value of M conditioned on $a^{-1}b=c$ is the same as the expected value of $M(X+c)$ conditioned on $b=0$. Therefore, it suffices to give the proof under the assumption that $b=0$. So we assume that $b=0$.

Let $C=C(a)=[m]_p^{-1}\,aX$ (a random set). There exists an interval I_a of size at most $\lceil\frac{p}{m}\rceil$ that contains M elements of C for the following reason: Let $f:\mathbb{Z}_p\to\mathbb{Z}_m$ be defined by $z\mapsto[\iota_p(z)]_m$. By definition, there exists a random variable $y=y(a)\in\mathbb{Z}_p$ such that $\left|f^{-1}(y(a))\cap aX\right|\ge M$. And there exists an $i=i(a)\in[m]$ such that

$$f^{-1}(y(a))=\left\{[i(a)+km]_p\mid k\in\mathbb{Z},0\le k<\frac{p-i(a)}{m}\right\}\,,$$

and hence $I_a=m^{-1}f^{-1}(y(a))$ is an interval of size $\le\lceil\frac{p}{m}\rceil$ that contains M elements of A. If there are multiple ways to choose I_a, it is arbitrarily chosen among all possibilities.

Let $T=\left[1,\frac{n}{4}\right]\cap\mathbb{Z}$, and let $t\in T$ be some fixed integer. We are now going to bound the probability that $M\ge 4t$. Let $\delta=\Pr[M\ge 4t]$, and let $A=\left\{a\in\mathbb{Z}_p^*\mid M(a)\ge 4t\right\}$.

Let $E\subset\mathbb{Z}_p^*$ be the set of all elements $e\in\mathbb{Z}_p^*$ that satisfies that $\iota_p(e)$ is a prime in the interval $(t,2t)$. Let $D=D(a)\subset E$ be the set of all elements $e\in E$ such that $ae\in A$, that is, $D=D(a)=E\cap a^{-1}A$. By linearity of expectation and because as a is uniformly random on $\mathbb{Z}_p^*$ so is ae, we have that $\mathrm{E}[|D|]=|E|\,\delta$.

Recall that $C=C(a)=[m]_p^{-1}\,aX$. For any $d\in D$ we have that $ad\in A$. Therefore, as stated above, there exists an interval of size $\lceil\frac{p}{m}\rceil$ that contains at least

$4t$ elements of $dC = [m]_p^{-1}(ad)X$. By Lemma 3.2, this implies that there are $t\,|D|$ ordered pairs of different elements $x, x' \in X$ such that $\|ax - ax'\|_p < \frac{p}{mt}$. So the expected number of elements $x, x' \in X$ such that $\|a(x-x')\|_p < \frac{p}{mt}$ is at least $t\,\mathrm{E}[|D|] = t\delta\,|E|$. On the other hand, for each ordered pair of different elements $x, x' \in X$ the probability that $\|a(x-x')\|_p < \frac{p}{mt}$ is at most $\frac{2p}{mt(p-1)}$, and by linearity of expectation the expected number of such ordered pairs is at most

$$n(n-1)\cdot\frac{2p}{mt(p-1)} \le \frac{2n}{t}\,.$$

We conclude that $t\delta\,|E| \le \frac{2n}{t}$. By the prime number theorem, $|E| = \Theta\!\left(\frac{t}{\log t}\right) = \Omega\!\left(\frac{t}{\log n}\right)$. Reordering gives us that

$$\Pr[M \ge 4t] = \delta = O\!\left(\frac{n\log n}{t^3}\right)\,.$$

The expected value of M can now be bounded in the following manner:

$$\begin{aligned}
\mathrm{E}[M] &= \sum_{k=1}^{\infty}\Pr[M \ge k] \\
&= \sum_{k=1}^{\lfloor \sqrt[3]{n\log n}\rfloor}\Pr[M \ge k] + \sum_{k=\lfloor \sqrt[3]{n\log n}\rfloor+1}^{n}\Pr[M \ge k] \\
&\le \left\lfloor \sqrt[3]{n\log n}\right\rfloor + \sum_{k=\lfloor \sqrt[3]{n\log n}\rfloor+1}^{n} O\!\left(\frac{n\log n}{k^3}\right) \\
&= O\!\left(\sqrt[3]{n\log n}\right)\,,
\end{aligned}$$

which was what we wanted. ☐

The proof of Theorem 3.4 is very similar to the proof of Theorem 3.3 but we include it for completeness.

THEOREM 3.4. *Let n, ℓ, r, q, m be integers with $q = 2^r, m = 2^\ell$, and $q \ge m \ge n$. Let $X \subset \mathbb{Z}_q$ be a set of n elements. Let $h : \mathbb{Z}_q \to [m]$ be defined by $h(x) = \left\lfloor \iota_q(ax)\cdot 2^{\ell-r}\right\rfloor$ where $a \in \mathbb{Z}_q^*$ is chosen uniformly at random. Let $M = M(X) = \max_{y\in[m]}|\{x \in X \mid h(x) = y\}|$. Then*

$$\mathrm{E}[M] = O\!\left(\sqrt[3]{n\log n}\right)\,. \tag{5}$$

Proof. Let $y = y(a)$ be a random variable such that $\left|h^{-1}(y(a)) \cap X\right| = M$, and let $C = C(a) = aX$ be a random set. The set $ah^{-1}(y)$ is an interval of size $\frac{q}{m}$ that contains exactly M elements of C.

Let $T = \left[1, \frac{n}{4}\right] \cap \mathbb{Z}$, and let $t \in T$ be some fixed integer. We are now going to bound the probability that $M \ge 4t$. Let $\delta = \Pr[M \ge 4t]$, and let $A = \left\{a \in \mathbb{Z}_q^* \mid M(a) \ge 4t\right\}$.

Let $E \subset \mathbb{Z}_q^*$ be the set of all elements $e \in \mathbb{Z}_q^*$ that satisfies that $\iota_q(e)$ is a prime in the interval $(t, 2t)$. Let $D = D(a) \subset E$ be the set of all elements $e \in E$ such that $ae \in A$, that is, $D = D(a) = E \cap a^{-1}A$. By linearity of expectation and because as a is uniformly random on $\mathbb{Z}_q^*$ so is ae, we have that $\mathrm{E}[|D|] = |E|\,\delta$.

Recall that $C = C(a) = aX$. For any $d \in D$ we have that $ad \in A$. Therefore, as stated above, there exists an interval of size $\frac{q}{m}$ that contains at least $4t$ elements of $dC = (ad)X$. By Lemma 3.2, this implies that there are $t\,|D|$ ordered pairs of different elements $x, x' \in X$ such that $\|ax - ax'\|_q < \frac{q}{mt}$. So the expected number of elements $x, x' \in X$ such that $\|a(x - x')\|_q < \frac{q}{mt}$ is at least $t\,\mathrm{E}[|D|] = t\delta\,|E|$. On the other hand, for each ordered pair of different elements $x, x' \in X$ the probability that $\|a(x - x')\|_q < \frac{q}{mt}$ is at most $\frac{4}{mt}$, and by linearity of expectation the expected number of such ordered pairs is at most

$$n(n-1) \cdot \frac{4}{mt} \leq \frac{4n}{t}\,.$$

We conclude that $t\delta\,|E| \leq \frac{4n}{t}$, and now we can bound the expected value exactly as in Theorem 3.3. □

REFERENCES

[1] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos, *Linear hash functions*, J. ACM, 46 (1999), pp. 667–683.

[2] L. Carter and M. N. Wegman, *Universal classes of hash functions*, J. Comput. System Sci., 18 (1979), pp. 143–154, https://doi.org/10.1016/0022-0000(79)90044-8, see also STOC '77.

[3] M. Dietzfelbinger, T. Hagerup, J. Katajainen, and M. Penttonen, *A reliable randomized algorithm for the closest-pair problem*, J. Algorithms, 25 (1997), pp. 19–51.

[4] G. Markowsky, L. Carter, and M. N. Wegman, *Analysis of a universal class of hash functions*, in Mathematical Foundations of Computer Science 1978, Proceedings of the 7th Symposium, Zakopane, Poland, Springer, Berlin, New York, 1978, pp. 345–354.

[5] K. Mehlhorn and U. Vishkin, *Randomized and deterministic simulations of PRAMs by parallel machines with restricted granularity of parallel memories*, Acta Inform., 21 (1984), pp. 339–374.