2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Linear Hashing is Awesome

Mathias Bæk Tejs Knudsen
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
Email: mathias@tejs.dk

***Abstract*—The most classic textbook hash function, e.g. taught in CLRS [MIT Press '09], is**

$$h(x) = ((ax + b) \bmod p) \bmod m \ , \qquad (\diamond)$$

where $x, a, b \in \{0, 1, \ldots, p-1\}$ and a, b are chosen uniformly at random. It is known that $(\diamond)$ is 2-independent and almost uniform provided p is a prime and $p \gg m$. This implies that when using $(\diamond)$ to build a hash table with chaining that contains $n \leqslant m$ keys, the expected query time is $O(1)$ and the expected length of the longest chain is $O(\sqrt{n})$. This result holds for any 2-independent hash function. No hash function can improve on the expected query time, but the upper bound on the expected length of the longest chain is not known to be tight for $(\diamond)$. Partially addressing this problem, Alon et al. [STOC '97] proved the existence of a class of linear hash functions such that the expected length of the longest chain is $\Omega(\sqrt{n})$ and leave as an open problem to decide which non-trivial properties $(\diamond)$ has. We make the first progress on this fundamental problem, by showing that the expected length of the longest chain is at most $n^{1/3+o(1)}$ which means that the performance of $(\diamond)$ is similar to that of a 3-independent hash function for which we can prove an upper bound of $O\left(n^{1/3}\right)$.

As a lemma we show that within a fixed set of integers there are few pairs such that the height of the ratio of the pairs are small. Given two non-zero coprime integers $n, m \in \mathbb{Z}$ with the height of $\frac{n}{m}$ is $\max\{|n|, |m|\}$, and the height is a way of measuring how complex a fraction is. This is proved using a mixture of techniques from additive combinatorics and number theory, and we believe that the result might be of independent interest.

For a natural variation of $(\diamond)$, we show that it is possible to apply second order moment bounds even when a hash value is fixed. As a consequence:

- **For min-wise hashing it was known that any key from a set of n keys has the smallest hash value with probability $O\left(\frac{1}{\sqrt{n}}\right)$. We improve this to $n^{-1+o(1)}$.**
- **For linear probing it was known that the worst case expected query time is $O(\sqrt{n})$. We improve this to $n^{o(1)}$.**

***Keywords*-hashing, linear hashing, hashing with chaining, additive combinatorics.**

I. Introduction

Hash functions are widely used and well studied within theoretical computer science. There are a vast number of different hash functions in the literature ranging from simple to complex, and some hash functions are constructed to perform well when used for specific purposes. In contrast a very fundamental class of hash functions is $\overline{h} : [p] \to [m]$ (where $[m] = \{0, 1, \ldots, m-1\}$) defined by $\overline{h}(x) = ((ax + b) \bmod p) \bmod m$, where $a, b \in [p]$ are chosen uniformly at random. Here p is a prime and $p \geqslant m$. Almost all students of computer science will come across it at least once in their curriculum, and it is e.g. the only randomized hash function mentioned in CLRS [1]. This makes the study of this particular hash function especially interesting.

The study of $\overline{h}(x)$ boils down to the following two properties:

(1) The collision probability is low, that is $\Pr(\overline{h}(x) = \overline{h}(y)) = O\left(\frac{1}{m}\right)$ for $x \neq y$. This follows from the 2-independence of $\overline{h}$.
(2) $\overline{h}(x)$ is almost affine, in the following sense: $\overline{h}(x) + \overline{h}(y) - \overline{h}(x+y) - \overline{h}(0)$ can only attain a constant number of different values.

In 1979 Property (1) led Carter and Wegman [2] to introduce the concept of k-independence, which has since been a prominent part of the studies of hash functions. This was done by replacing the degree one polynomial in the definition of $\overline{h}$ with a random degree $k-1$ polynomial. From Property (1) one can obtain a number of results when using $\overline{h}$ as the hash function:

(i) When inserting n keys into a table of size n using hashing with chaining the expected time used to insert any key is $O(1)$.
(ii) When inserting n keys into a table of size n using hashing with chaining the expected size of the longest chain is $O(\sqrt{n})$. This corresponds to the expected worst case performance over the insertion of all n keys.
(iii) Letting $X = \{x_1, \ldots, x_n\}$ be a set of n keys, $\overline{h}(x_1)$ is the minimum hash value of all $\overline{h}(x_i), x_i \in X$ with probability $O\left(\frac{1}{\sqrt{n}}\right)$.
(iv) In a hash table implemented by linear probing using a table of size n and containing $(1-\varepsilon)n$ keys, the worst-case expected query time is $O(\sqrt{n})$.

Property (2) is used by Baran et al. [3] to solve the 3SUM problem faster than $\Theta(n^2)$, but Property (1) seems to have more far reaching consequences than Property (2).

Result (i) is tight as no matter which hash function is used the expected time to insert any element must be $\Omega(1)$. However, it is not clear whether Results (ii), (iii) and (iv)

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.45

are tight. There are no results showing that the bounds in Results (ii)–(iv) are tight for $\overline{h}$. Instead it is shown in [4], [5] that there exist 2-independent hash functions for which the upper bounds in Results (ii)–(iv) are tight. This shows that we either have to use Property (2) or find a new property in order to improve these bounds. Furthermore, Alon et al. [6] have shown that there exist linear hash functions meeting the bounds in Result (ii). These linear hash functions can easily be altered to be affine hash functions that satisfy Property (1) as well showing that even a combination of Property (1) and (2) will not suffice in order to improve on Result (ii). However in [6] it is also shown that there exists linear hash function (over the field of 2 elements) for which the maximum load is $O(\log n \log\log n)$.

The discussion above leads to one of two conclusions. Either Property (1) and (2) captures the essence of $\overline{h}$ and we need to prove that Results (ii)–(iv) cannot be improved. Otherwise Results (ii)–(iv) can be improved and we need to find a new property of $\overline{h}$ that must be entirely different of Property (1) and (2).

A. Our Results

In the following let p be a prime and m an integer not larger than p. Let $a, b \in [p]$ be independent and uniformly random integers. Let $h : [p] \to [p]$ be defined by $h(x) = (ax + b) \bmod p$. We let $\overline{h}(x) = h(x) \bmod m$ and $\tilde{h}(x) = \left\lfloor \frac{h(x)\cdot m}{p} \right\rfloor$. We can think of $\overline{h}$ and $\tilde{h}$ as extracting the least and the most significant bits of h respectively. The main result of this paper is the proof of the following property of $\overline{h}$ and $\tilde{h}$:

Theorem 1. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$ and $m \geqslant n$. Let $x, y, z \in X$ be three keys chosen from X independently and uniformly at random, then:*

$$P\left(\overline{h}(x) = \overline{h}(y) = \overline{h}(z)\right) = O\left(m^{-2} + n^{-2+o(1)}\right),$$

and the same result holds when $\overline{h}$ is replaced with $\tilde{h}$.

Comparing Theorem 1 with the performance of a uniformly random hash function we see that if $\overline{h}$ had been uniformly random then we could have gotten the bound $O(m^{-2} + n^{-2})$ instead, which follows from the fact that for any 3 distinct keys the probability that they have same hash value is exactly m^{-2}. In fact this holds for any 3-independent hash function. Except for the multiplicative term $n^{o(1)}$ the main difference is that for a 3-independent hash function it holds that for *every* triple of distinct keys (x, y, z) the probability that they all have the same hash value is small. Theorem 1 instead states that this property holds for *most* such triples. Considering the three keys $(x, 2x, 3x)$ it becomes apparent that it is not possible to prove that it holds for every triple as if we condition on $\overline{h}(x) = \overline{h}(2x)$ then with probability $\approx \frac{1}{2}$ it also holds that $\overline{h}(2x) = \overline{h}(3x)$.

Whenever we are interested in the behaviour of all keys together the it might not be important whether the collision probability of a triple is small for every triple or for most triples of keys. As a direct corollary of Theorem 1 we prove the upper bound in Result (2) can be improved from $O(\sqrt{n})$ to $n^{1/3+o(1)}$:

Corollary 1. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$. Assume that $m \geqslant n$ and let M be the number of keys that hash to the most popular hash value, i.e. $M = \max_{v\in[m]} \left|\left\{x \in X \mid \overline{h}(x) = v\right\}\right|$. Then $E(M) \leqslant n^{1/3+o(1)}$. The same result holds when $\overline{h}$ is replaced with $\tilde{h}$ in the definition of M.*

In order to improve on Results (iii) and (iv) we prove the following result, that says that even if we fix one hash value, we can still use a certain type of second order moment bounds.

Theorem 2. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$. Let $v \in [m]$, $I \subset [m]$ be a set of consecutive[1] integers $\bmod m$ and $x_0 \in U\backslash X$. Let $A = \sum_{x\in X}[\tilde{h}(x) \in I]$ and $\mu = E(A)$. Here the expression $[\tilde{h}(x) \in I]$ is the Iverson bracket, which denotes a a number that is 1 if $\tilde{h}(x) \in I$ and 0 otherwise. Then for every $\delta > 0$ it holds that:*

$$\begin{aligned} &P\left(|A - \mu| \geqslant \delta\sqrt{\mu} \mid \tilde{h}(x_0) = v\right) \\ \leqslant &\frac{1}{\delta^2} \cdot \left(n^{o(1)} + O\left(\frac{n^2}{p\mu}\right)\right). \end{aligned}$$

Using Theorem 2 we improve on Results (iii) and (iv).

Corollary 2. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$. Let $x_0 \in X$. Then:*

$$P\left(\tilde{h}(x_0) < \min_{x\in X\setminus\{x_0\}}\left\{\tilde{h}(x)\right\}\right) \leqslant n^{-1+o(1)}.$$

Corollary 3. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$. Assume that $m \geqslant (1+\varepsilon)n$ for some constant $\varepsilon > 0$, and say that all of the keys from X are inserted into a table of size m using $\tilde{h}$ with linear probing. Then the expected query time for any element $u \in U$ is $n^{o(1)}$.*

B. Technical contribution

The main technical contribution is an upper bound on the number of pairs within a set where the ratio of the pairs have small *height*. For a non-zero rational number $x \in \mathbb{Q}$ the *height* $H(x)$ of x is the smallest number such that there exists integers $n, m \in \mathbb{Z}$ with $|n|, |m| \leqslant H(x)$ and $x = \frac{n}{m}$. Equivalently, if $x = \frac{m}{n}$ for co-prime integers n, m

[1]Meaning that for suitable values of $r, s \in \mathbb{N}$ we have $I = \{r \bmod m, (r+1) \bmod m, \ldots, (r+s-1) \bmod m\}$.

then $H(x) = \max\{|n|, |m|\}$, see e.g. [7]. We note that that $H(x) = H(x^{-1})$. The main technical contribution is captured in the following theorem:

Theorem 3. *Let $A \subset \mathbb{Q}\setminus\{0\}$ be a set of n non-zero rational numbers and let k be an integer. The number of pairs $(a, a') \in A^2$ such that $H\left(\frac{a}{a'}\right) \leqslant k$ is at most $nk \cdot 2^{O\left(\sqrt{(\log n)\cdot(\log\log 3k)}\right)}$.*

The theorem is proved using a combination of tools from additive combinatorics and number theory. The use of this theorem comes from the fact that we can essentially quantify "how correlated" $h(x), h(y), h(z)$ are by looking at the fraction $\frac{y-x}{z-x}$. If the fraction has small height, then the hash values are very correlated and vice versa.[2]

C. Structure of the Paper

The structure of the paper is as follows. In Section II we introduce the necessary notation. In Section III we prove Theorem 3, and in Section IV we show how to use this result to prove new properties for $\overline{h}$ and $\tilde{h}$. Finally, in Section V we show that these properties imply a number of interesting results.

II. Preliminaries

$\mathbb{Z}$ denotes the integers, $\mathbb{N}$ the positive integers, $\mathbb{Q}$ the rational numbers, and $\mathbb{Q}^+$ the positive rational numbers. $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ denotes the integers $\bmod n$. $\mathbb{Z}_n^*$ is the set of elements of $\mathbb{Z}_n$ having a multiplicative inverse; $(\mathbb{Z}_n^*, \cdot)$ is an abelian group. $[n]$ is shorthand for $\{0, 1, 2, \ldots, n-1\}$. For a pair of integers $n, m \in \mathbb{Z}$ such that $(n, m) \neq (0, 0)$ we let $\gcd(n, m)$ denote the greatest common divisor of n and m. If $\gcd(n, m) = 1$ then n and m are *coprime*. We write $a \mid b$ to mean that a divides b and $a \nmid b$ if a does not divide b. We use the Iverson bracket notation meaning that for a condition P we let $[P]$ denote 1 if P is satisfied and 0 otherwise. $\log$ denotes the natural logarithm.

Throughout the paper p will be a prime and m a positive integer smaller than p. We let $h : [p] \to [p]$ be a random hash function defined by $h(x) = (ax + b) \bmod p$ where $a, b \in [p]$ are chosen independently and uniformly at random. We will often consider h to be a mapping from $\mathbb{Z}_p$ to $\mathbb{Z}_p$ in the obvious way. We define the hash functions $\overline{h}, \tilde{h} : [p] \to [m]$ by:

$$\overline{h}(x) = h(x) \bmod p, \quad \tilde{h}(x) = \left\lfloor \frac{h(x)\cdot m}{p} \right\rfloor .$$

For an integer $x \in \mathbb{Z}$ we let $[x]_p \in \mathbb{Z}_p$ denote its residue class $\bmod p$. We let $\iota : \mathbb{Z}_p \to [p]$ be the unique mapping that satisfies $[\iota(x)]_p = x$.

[2]This is not entirely true as we need to introduce a concept of height $\bmod p$ in order to justify this claim, but for the sake of the simplicity of the introduction this concept is not defined before needed.

A non-zero rational number $x \in Q$ can be written as $x = \frac{m}{n}$ where $m, n \in \mathbb{Z}$ are coprime. The *height* of x is defined as $H(x) = \max\{|m|, |n|\}$.

For a positive integer n we let $d_2(n)$ be the number of divisors of n. In general, for a positive integer r we let $d_r(n)$ be the number of ways n can be written as a product of a sequence of r positive integers, i.e.:

$$d_r(n) = \left|\{(a_1, a_2, \ldots, a_r) \in \mathbb{N}^r \mid a_1 a_2 \ldots a_r = n\}\right| .$$

Let G be a finite abelian group. A *bi-character* $e : G \times G \to \mathbb{C}\setminus\{0\}$ is a function that satisfies

$$e(xx', \xi) = e(x, \xi)e(x', \xi), \quad e(x, \xi\xi') = e(x, \xi')e(x, \xi'),$$

for all $x, x', \xi, \xi' \in G$ and such that for any $x \in G$ other than the identity there exists $\xi \in G$ such that $e(x, \xi) \neq 1$, and for any $\xi \in G$ other than the identity there exists $x \in G$ such that $e(x, \xi) \neq 1$. It is well-known [8] that such a bi-character exists for any finite abelian group.

For a fixed group G and bi-character e we have the following definitions from Fourier analysis following the exposition from [8]. The *Fourier transform* $\widehat{f}(\xi)$ of a function $f(x)$ on G is defined by:

$$\widehat{f}(\xi) = \frac{1}{|G|}\sum_{x\in G} f(x)\overline{e(x, \xi)} .$$

With this definition we have the *Fourier inversion formula*:

$$f(x) = \sum_{\xi\in G} \widehat{f}(\xi)e(x, \xi) .$$

For functions $f(x)$ and $g(x)$ on G we define the convolution $(f * g)(x)$ to be:

$$(f * g)(x) = \frac{1}{|G|}\sum_{a,b\in G, ab=x} f(a)g(b) .$$

We note that $\widehat{(f * g)}(\xi) = \widehat{f}(\xi)\widehat{g}(\xi)$. *Plancherel's formula* is:

$$\sum_{x\in G} |f(x)|^2 = |G| \sum_{\xi\in G} \left|\widehat{f}(\xi)\right|^2 .$$

III. Heights

The main theorem we prove in this section is:

Theorem 3. *Let $A \subset \mathbb{Q}\setminus\{0\}$ be a set of n non-zero rational numbers and let k be an integer. The number of pairs $(a, a') \in A^2$ such that $H\left(\frac{a}{a'}\right) \leqslant k$ is at most $nk \cdot 2^{O\left(\sqrt{(\log n)\cdot(\log\log 3k)}\right)}$.*

We prove Theorem 3 by counting the number of solutions to $ab = a'b'$ where $a, a' \in A$ and b, b' are small numbers. An upper bound on this number is given in Lemma 1.

Lemma 1. *Let $A \subset \mathbb{N}$ be a set of n positive integers and $B = \{1, 2, \ldots, k\}$. Let t be the number of solutions to $ab =$*

$a'b'$ *where* $a, a' \in A, b, b' \in B$. *Then for any positive integer* r*:*

$$t \leqslant n^{1+1/r} \left(\sum_{i=1}^{k^r} (d_r(i))^2 \right)^{1/r} .$$

Before we prove Lemma 1 we will show how Theorem 3 follows from it.

Proof of Theorem 3: First we note that by multiplying all the numbers of A with a sufficiently large integer we can wlog assume that $A \subset \mathbb{Z} \setminus \{0\}$. Let $A' = \{|a| \mid a \in A\}$. We will use Lemma 1 with A' and $B = \{1, 2, \ldots, k\}$. We note that for every pair $(a, a') \in A$ such that $H\left(\frac{a}{a'}\right) \leqslant k$ there exist $b, b' \in B$ such that $|a|\, b = |a'|\, b'$. Furthermore each such equation can correspond to at most 4 pairs $(a, a') \in A$ where $H\left(\frac{a}{a'}\right) \leqslant k$. So $4t$ where t is from Lemma 1 is an upper bound on the number of pairs $(a, a') \in A$ where the height of the ratio is $\leqslant k$.

We will need the following theorem proved in [9].

Theorem 4 ([9])**.** *Let* $\varepsilon > 0$ *be a constant and let* $x \geqslant 1, s \geqslant 1 + \varepsilon$*, and let* $z \geqslant 2$ *be an integer. Then* $\sum_{n \leqslant x} (d_z(n))^s$ *is bounded by:*

$$x \frac{(\log x + z^s \log(3z^s))^{z^s - 1}}{(z^s \cdot \log(3z^s) e^{\gamma - 1})^{z^s}} \cdot e^{O(z^s/(\log(3z^s)))} .$$

where $\gamma \approx 0.577\ldots$ *is Euler's constant and the constant in the* O*-notation depends on* ε*.*

Letting $\varepsilon = 1, z = r, x = k^r, s = 2$ in Theorem 4 we get:

$$\sum_{n \leqslant k^r} (d_r(n))^2 = (k \cdot (O(\log 3k))^r)^r .$$

By combining this with Lemma 1 we get:

$$4t \leqslant n^{1+1/r} k \, (O(\log 3k))^r .$$

Letting $r \approx \sqrt{\frac{\log n}{\log \log 3k}}$ gives the desired. ∎

Now we turn to proving Lemma 1. The idea is to count the number of solutions using Fourier analysis. Subsequently, we give an upper bound on the corresponding sum using Hlder's inequality. The result now comes from using the Fourier inversion formula on the upper bound.

Proof of Lemma 1: Let r be a positive integer. Let q be a prime greater than $\max_{a \in A} \{a\} \cdot k^r$, and let $G = \mathbb{Z}_q^*$. We note that A and B can be considered to be subsets of G and we will do so. Let $\chi_A, \chi_B : G \to \{0, 1\}$ be the characteristic functions of A and B, respectively. For any positive integer n less than q we identify n with its residue $\bmod q$, i.e. we consider it to be an element of G. We note that $|G| \cdot (\chi_A * \chi_B)(n)$ is the number of pairs $(a, b) \in A \times B$ such that $ab = n$ since $ab < q$ for each $a \in A$, $b \in B$, and hence t is equal to

$$\begin{aligned} t &= |G|^2 \sum_{x \in G} ((\chi_A * \chi_B)(x))^2 = |G|^3 \sum_{\xi \in G} \left| \widehat{(\chi_A * \chi_B)}(\xi) \right|^2 \\ &= |G|^3 \sum_{\xi \in G} \left|\widehat{\chi_A}(\xi)\right|^2 \left|\widehat{\chi_B}(\xi)\right|^2 , \end{aligned} \tag{1}$$

where we use Plancherel's formula. By Hölder's inequality we have:

$$\begin{aligned} &\sum_{\xi \in G} \left|\widehat{\chi_A}(\xi)\right|^2 \left|\widehat{\chi_B}(\xi)\right|^2 \\ &\leqslant \left(\sum_{\xi \in G} \left|\widehat{\chi_A}(\xi)\right|^{2r/(r-1)} \right)^{(r-1)/r} \left(\sum_{\xi \in G} \left|\widehat{\chi_B}(\xi)\right|^{2r} \right)^{1/r} \\ &= S_1^{(r-1)/r} S_2^{1/r} , \end{aligned} \tag{2}$$

where S_1, S_2 are defined in the obvious way. We will bound S_1 and S_2 in turn starting with S_1.

Note that by the definition of $\widehat{\chi_A}$ we have $|\widehat{\chi_A}(\xi)| \leqslant \frac{|A|}{|G|}$ for all $\xi \in G$. This implies together with Plancherel's formula that

$$S_1 \leqslant \frac{|A|^{2/(r-1)}}{|G|^{2/(r-1)}} \sum_{\xi \in G} \left|\widehat{\chi_A}(\xi)\right|^2 = \left(\frac{|A|}{|G|}\right)^{(r+1)/(r-1)} . \tag{3}$$

We will now bound S_2. We note that $\left|\widehat{\chi_B}(\xi)\right|^{2r} = \left|\widehat{(\chi_B * \ldots * \chi_B)}(\xi)\right|^2$ where there are r terms in the convolution. Hence we have

$$\begin{aligned} S_2 &= \sum_{\xi \in G} \left|\widehat{\chi_B}(\xi)\right|^{2r} = \sum_{\xi \in G} \left|\widehat{\chi_B * \ldots * \chi_B}(\xi)\right|^2 \\ &= \frac{1}{|G|} \sum_{x \in G} \left|(\chi_B * \ldots * \chi_B)(x)\right|^2 . \end{aligned} \tag{4}$$

Since $q > k^r$ for any positive integer n less than q we have that $|G|^{r-1} (\chi_B * \ldots * \chi_B)(n)$ is equal to the number of ways n can be written as a product of r integers where none is larger than k. This is equal to 0 for any $n > k^r$ and for any $n \leqslant k^r$ it is bounded by $d_r(n)$. Combining this with (4) gives:

$$\begin{aligned} S_2 &= \frac{1}{|G|^{2r-1}} \sum_{x \in G} \left| |G|^{r-1} \cdot (\chi_B * \ldots * \chi_B)(x) \right|^2 \\ &\leqslant \frac{1}{|G|^{2r-1}} \sum_{x=1}^{k^r} (d_r(x))^2 . \end{aligned} \tag{5}$$

Combing (1), (2), (3), and (5) gives

$$\begin{aligned} t &\leqslant |G|^3 \cdot \left(\frac{|A|}{|G|}\right)^{(r+1)/r} \left(\frac{1}{|G|^{2r-1}} \sum_{n=1}^{k^r} (d_r(n))^2 \right)^{1/r} \\ &= |A|^{1+1/r} \left(\sum_{n=1}^{k^r} (d_r(n))^2 \right)^{1/r} , \end{aligned}$$

as desired. ∎

In the reductions in Section IV we will be interested in the sum of the reciprocals of the heights studied in Theorem 3. Corollary 4 gives an upper bound on this quantity.

Corollary 4. *Let* $A \subset \mathbb{Q}^+$ *be a set of* n *positive rational numbers. Then:*

$$\sum_{a, a' \in A} \frac{1}{H\left(\frac{a}{a'}\right)} \leqslant n 2^{O(\sqrt{(\log n) \cdot (\log \log 3n)})} .$$

Proof: For k a positive integer let V_k be the subset of A^2 containing pairs where the height of the ratio is a most k, i.e.:

$$V_k = \left\{(a,a') \in A \mid H\left(\frac{a}{a'}\right) \leqslant k\right\}.$$

For convenience let $V_0 = \varnothing$. Then we know that:

$$\begin{aligned}\sum_{a,a'\in A} \frac{1}{H\left(\frac{a}{a'}\right)} &= \sum_{k=1}^{\infty} \frac{1}{k} \cdot |V_k \backslash V_{k-1}| \\ &= \sum_{k=1}^{\infty} |V_k| \cdot \left(\frac{1}{k} - \frac{1}{k+1}\right). \end{aligned} \tag{6}$$

Firstly we note that:

$$\begin{aligned}&\sum_{k=n+1}^{\infty} |V_k| \cdot \left(\frac{1}{k} - \frac{1}{k+1}\right) \\ &\leqslant \sum_{k=n+1}^{\infty} n^2 \cdot \left(\frac{1}{k} - \frac{1}{k+1}\right) = \frac{n^2}{n+1} < n. \end{aligned} \tag{7}$$

For any $k \leqslant n$ we see by Theorem 3 that $|V_k| \leqslant nk2^{O(\sqrt{(\log n)\cdot(\log\log 3n)})}$. Hence:

$$\begin{aligned}&\sum_{k=1}^{n} |V_k| \cdot \left(\frac{1}{k} - \frac{1}{k+1}\right) \leqslant \sum_{k=1}^{n} \frac{|V_k|}{k^2} \\ &\leqslant n2^{O(\sqrt{(\log n)\cdot(\log\log 3n)})}. \end{aligned} \tag{8}$$

Combining (6), (7), and (8) yields the desired conclusion. ∎

IV. Reduction

The goal of this section is to prove Theorem 1.

For $x \in \mathbb{Z}_p$, $r \in \mathbb{N}$ we let $I(x,r)$ denote the set $\left\{x, x+[1]_p, \ldots, x+[r-1]_p\right\}$. A subset $A \subset \mathbb{Z}_p$ is called an *interval* if it is on the form $I(x,r)$. A subset $B \subset \mathbb{Z}_p$ is called a *generalized interval* if $B = xA$ for some interval A and $x \in \mathbb{Z}_p^*$. For $x \in \mathbb{Z}_p$ we let $|x|$ be defined as

$$|x| = \min\{\iota(x), \iota(-x)\}$$

For $x \in \mathbb{Z}_p^*$ we let $H_p(x)$ be the *restricted height* defined by:

$$H_p(x) = \min\left\{\max\{|a|,|b|\} \mid a,b \in \mathbb{Z}_p^*,\ x = ab^{-1}\right\}$$

We note that $H_p(x) = H_p\left(x^{-1}\right)$. In Lemma 2 it is proved that $H_p(x) \leqslant \sqrt{p}$. The following lemma also shows that if x, y is much smaller than p, then if the height of $\frac{x}{y}$ is large then the restricted height is large as well.

Lemma 2. *Let $x, y \in \mathbb{Z}$ be integers not divisible by p, then:*

$$H_p\left([x]_p [y]_p^{-1}\right) \geqslant \min\left\{\frac{p}{|x|+|y|}, H\left(\frac{x}{y}\right)\right\}, \tag{9}$$

$$H_p(x) < \sqrt{p}. \tag{10}$$

Proof: First we prove (9). Fix integers $a, b \in \mathbb{Z}$ that satisfy $|a|, |b| \leqslant H_p\left([x]_p [y]_p^{-1}\right)$ and $[x]_p [y]_p^{-1} = [a]_p [b]_p^{-1}$. Then $p \mid xb - ay$. Either $|xb-ay| < p$ or $|xb-ay| \geqslant p$. If $|xb-ay| < p$ then since $p \mid xb-ay$ we have $xb = ay$ and $\frac{x}{y} = \frac{a}{b}$. This implies that $H_p\left([x]_p [y]_p^{-1}\right) \geqslant H\left(\frac{x}{y}\right)$, and so (9) holds. Otherwise $|xb-ay| \geqslant p$. By the triangle inequality $p \leqslant |xb-ay| \leqslant H_p\left([x]_p [y]_p^{-1}\right) \cdot (|x|+|y|)$ and therefore (9) also holds in this case.

Now we turn to proving (10). Let $z = \left\lfloor\sqrt{p}\right\rfloor$ and $S = \left\{[0]_p x, [1]_p x, \ldots, [z]_p x\right\}$, then S contains $z+1$ distinct elements. Write S as $S = \{s_0, s_1, \ldots, s_z\}$ where the elements of S are ordered such that $\iota(s_0) < \ldots < \iota(s_z)$. Let $s_{z+1} = s_0$, then we have $\sum_{i=0}^{z} \iota(s_{i+1}) - \iota(s_i) = p$. Hence there must exist an index i such that $s_{i+1} - s_i \leqslant \frac{p}{z+1} < \sqrt{p}$. By the definition of S we have that $s_{i+1} - s_i = tx$ for some $t \in \mathbb{Z}_p$ with $|t| \leqslant z < \sqrt{p}$. So we can write $x = t(s_{i+1}-s_i)^{-1}$ with $|t|, |s_{i+1}-s_i| < \sqrt{p}$ and hence $H_p(x) < \sqrt{p}$ which proves (10). ∎

We will need the following technical lemma. Lemma 3 uses the fact that that the numbers $[i]_m x^{-1}$ for $i = 0, 1, \ldots, \iota(x)-1$ are essentially evenly spaced in $\mathbb{Z}_m$ for any $x \in \mathbb{Z}_m$. This implies that given an interval I only a fraction of about $\frac{|I|}{m}$ of these numbers can be contained in I.

Lemma 3. *Let m be a positive integer and et $x, y \in \mathbb{Z}_m^*$ be elements satisfying $\sqrt{m} \geqslant |x| \geqslant |y|$ and $\gcd(|x|,|y|) = 1$ and let $I \subset \mathbb{Z}_m$ be an interval. The number of elements $r \in [|x|]$ such that $[r]_m yx^{-1} \in I$ is at most $\frac{|I||x|}{m} + 2$.*

Proof: If $\iota(x) \neq |x|$ we can replace x and y by $-x$ and $-y$, so wlog assume that $\iota(x) = |x|$.

Let $A = \left\{[i]_m x^{-1} \mid i \in \{0,1,\ldots,\iota(x)-1\}\right\}$ and $B = \left\{\left[\left\lfloor\frac{im}{|x|}\right\rfloor\right]_m \mid i \in [|x|]\right\}$. First we prove that $A = B$. Since A and B are finite sets of the same cardinality it suffices to prove that $A \subset B$. Let $i \in \{0,1,\ldots,\iota(x)-1\}$ and $j \in [|x|]$ be the unique integer satisfying $|x| \mid mj + \iota(i)$. Then $\left[\frac{mj+\iota(i)}{|x|}\right]_m = \iota(i)x^{-1}$ and $\frac{mj+\iota(i)}{|x|} = \left\lceil\frac{mj}{|x|}\right\rceil$ and hence $\iota(i)x^{-1} \in B$, so we conclude that $A \subset B$.

Say that there are $k \geqslant 2$ elements $r \in [|x|]$ such that $ryx^{-1} \in I$. (If $k \leqslant 1$ there is nothing to prove.) If $|y| = \iota(y)$ we let $I' = I - \{0,1,\ldots,|y|-1\}$ and otherwise we let $I' = I + \{0,1,\ldots,|y|-1\}$. We note that $|I'| < |I| + |y|$. For each $r \in [|x|]$ that satisfies $[r]_m yx^{-1} \in I$ there is an element $a \in A$ such that $a \in I'$. Since I' is an interval there exists $r \in \mathbb{Z}_m, s \in \mathbb{Z}$ such that $I' = \{r, r+[1]_m, \ldots, r+[s-1]_m\}$. Since I' contains k elements from B the interval $[r, r+s)$ contains k elements on the form $\frac{ip}{|x|}, i \in \mathbb{Z}$, and hence $s \geqslant (k-1)\frac{p}{|x|}$. Using that $s = |I'| < |I| + |y|$ we get that:

$$k \leqslant |I'|\frac{|x|}{m} + 1 < \frac{|I|\,|x|}{m} + \frac{|x|\,|y|}{m} + 1 \leqslant \frac{|I|\,|x|}{m} + 2.$$

∎

Lemma 4 below show how we can connect the concept of heights to linear hash functions. Given elements $x, y \in \mathbb{Z}_p^*$ it gives an upper bound on the probability that $(ax, ay) \in I^2$, when $a \in \mathbb{Z}_p$ is chosen uniformly at random. If the random variables ax and ay had been independent (they are clearly not!) then the probability would have been exactly $(|I|/p)^2$. Lemma 4 shows that when the restricted height of xy^{-1} is large the probability is close to $(|I|/p)^2$.

Lemma 4. *Let $x, y \in \mathbb{Z}_p^*$ and $I \subset \mathbb{Z}_p$ be a generalized interval. Then:*

$$P((ax, ay) \in I^2) \leqslant (P(ax \in I))^2 + O\left(\frac{P(ax \in I)}{H_p(xy^{-1})} + \frac{1}{p}\right),$$

where $a \in \mathbb{Z}_p$ is chosen uniformly at random.

The idea of the proof is the following (assuming that I is an interval for simplicity). By multiplying x, y with an appropriate factor we can wlog assume $|y| \leqslant \iota(x) \leqslant H_p(xy^{-1}) < \sqrt{p}$. Instead of directly calculating the probability that $(ax, ay) \in I^2$ we consider the set $S = \left\{a, [1]_p x^{-1}, \ldots, [\iota(x) - 1]_p x^{-1}\right\}$ and upper bound the expected number of elements $s \in S$ satisfy $(sx, sy) \in I^2$. By linearity of expectation this will allow us to upper bound the desired probability. This turns out to be easier for the following reason. The set Sx is an interval and hence, for most values of a, it is either contained in I or disjoint from I. When it is disjoint from I there are no elements $s \in S$ such that $(sx, sy) \in I^2$. When Sx is contained in I we just need to calculate $|Sy \cap I|$, which we do by using Lemma 3. The details are given below.

Proof: Firstly, we see that $I = zJ$ for some interval J and $z \in \mathbb{Z}_p^*$. Since $(ax, ay) \in I \times I$ iff $(axz, ayz) \in J \times J$ we can replace x, y, I with xz, yz, J respectively and wlog assume that I is an interval in the following.

For any $\ell \in \mathbb{Z}_p^*$ the probability that $(ax, ay) \in I \times I$ does not change if we exchange x and y with $x\ell$ and $y\ell$ respectively. Let $H_p(xy^{-1}) = t$, then there exists ℓ such that $|x\ell|, |y\ell| \leqslant t$. Since we may also swap x and y we can wlog assume that $|y| \leqslant \iota(x) = t < \sqrt{p}$.

We let T be the set of elements $s \in \mathbb{Z}_p$ such that $(sx, sy) \in I^2$, i.e. $T = x^{-1}I \cap y^{-1}I$. Let $a \in \mathbb{Z}_p$ be chosen uniformly at random and let P_0 be the probability that $(ax, ay) \in I^2$,

$$P_0 = P((ax, ay) \in I^2).$$

Let $S = \left\{a, [1]_p x^{-1}, \ldots, [\iota(x) - 1]_p x^{-1}\right\}$. By linearity of expectation we have:

$$tP_0 = \iota(x)P_0 = E\left(\sum_{s \in S} \left[(sx, sy) \in I^2\right]\right),$$

Clearly the number of elements $s \in S$ such that $(sx, sy) \in I^2$ is bounded by $|Sx \cap I|$ and $|Sy \in \cap I|$. Inserting this gives:

$$tP_0 \leqslant E\left(\min\left\{|Sx \cap I|, |Sy \cap I|\right\}\right).$$

By Lemma 3 we have that $|Sy \cap I|$ is bounded by $\frac{|I|t}{p} + 2$, so we get

$$tP_0 \leqslant \left(\frac{|I|\,t}{p} + 2\right) \cdot P\left(Sx \cap I \neq \varnothing\right).$$

Since Sx is an interval consisting of t elements, we have that the probability that $Sx \cap I \neq \varnothing$ less than $\frac{|I|+t}{p}$. Inserting this gives

$$\begin{aligned} P_0 &\leqslant \left(\frac{|I|}{p} + \frac{2}{t}\right) \cdot \left(\frac{|I|}{p} + \frac{t}{p}\right) \\ &< \left(\frac{|I|}{p}\right)^2 + 3\frac{|I|}{pt} + \frac{2}{p}, \end{aligned} \tag{11}$$

which finishes the proof as $P(ax \in I) = \frac{|I|}{p}$. ∎

Instead of proving Theorem 1 we will prove a slightly more general theorem from which Theorem 1 follows.

Theorem 5. *Let $U = [u]$ be a universe of keys and $X \subset U$ a set of n keys such that $p > un$. Let $x \in X$, $v \in [m]$ be a hash value and let $I \subset [m]$ be an interval when considered as a subset of $\mathbb{Z}_m$. Let $y, z \in X \backslash \{x\}$ be chosen independently and uniformly at random, then:*

$$\begin{aligned} &P\left((\tilde{h}(y), \tilde{h}(z)) \in I^2 \mid \tilde{h}(x) = v\right) \\ &\leqslant \left(\frac{|I|}{m}\right)^2 + n^{-1+o(1)} \cdot \frac{|I|}{m} + O(p^{-1}). \end{aligned} \tag{12}$$

If I consists of a single element the statement also holds if we replace $\tilde{h}$ with $\overline{h}$.

Proof: Fix $y, z \in X \backslash \{x\}$. Later in the proof we will unfix y, z and consider them to be random variables. Let $J = \tilde{h}^{-1}(I)$, then J is an interval. (If I is a singleton, then $\overline{h}^{-1}(I)$ is a generalized interval, showing that the statement also holds if we replace $\tilde{h}$ with $\overline{h}$ in this case.) Let $u \in \tilde{h}^{-1}(v)$. We will show that (12) holds when we condition on $h(x) = u$ instead of $\tilde{h}(x) = v$. Since u is arbitrarily chosen this will be sufficient. If $h(x) = u$ and $(\tilde{h}(y), \tilde{h}(z)) \in I^2$ then $(h(y) - h(x), h(z) - h(x)) \in (J - u)^2$, i.e. $(a(y - x), a(z - x)) \in (J - u)^2$. Since h is 2-independent $h(x)$ and $h(y) - h(x) = a(y - x)$ are independent, showing that $h(x)$ and a are independent. So we conclude that:

$$\begin{aligned} &P\left((\tilde{h}(y), \tilde{h}(z)) \in I^2 \mid h(x) = u\right) \\ =&P\left((a(y - x), a(z - x)) \in (J - u)^2 \mid h(x) = u\right) \\ =&P\left((a(y - x), a(z - x)) \in (J - u)^2\right). \end{aligned} \tag{13}$$

By Lemma 4 we get:

$$\begin{aligned} &P\left((a(y - x), a(z - x)) \in (J - u)^2\right) \\ &\leqslant \left(\frac{|J - u|}{p}\right)^2 + O\left(\frac{1}{H_p\left(\frac{y-x}{z-x}\right)} \cdot \frac{|J - u|}{p} + \frac{1}{p}\right). \end{aligned} \tag{14}$$

Since $|J-u|\,p \leqslant \frac{|I|}{m}\cdot\left(1+\frac{1}{p}\right)$ we can combine (13) and (14) to conclude that:

$$P\left((\tilde{h}(y),\tilde{h}(z))\in I^2 \mid \tilde{h}(x)=v\right) \leqslant \left(\frac{|I|}{m}\right)^2 + O\left(\frac{1}{H_p\left(\frac{y-x}{z-x}\right)}\cdot\frac{|I|}{m}+\frac{1}{p}\right). \quad (15)$$

Now we unfix y,z and consider $y,z\in X\backslash\{x\}$ to be chosen uniformly at random. By averaging over (15) we get:

$$P\left((\tilde{h}(y),\tilde{h}(z))\in I^2 \mid \tilde{h}(x)=v\right) \leqslant \left(\frac{|I|}{m}\right)^2 + O\left(\frac{1}{(n-1)^2}\sum_{\substack{y',z'\in\\ X\backslash\{x\}}}\frac{|I|}{H_p\left(\frac{y'-x}{z'-x}\right)m}+\frac{1}{p}\right). \quad (16)$$

By Lemma 2 we see that:

$$\frac{1}{H_p\left(\frac{y'-x}{z'-x}\right)} \leqslant \frac{1}{\min\left\{\Omega(n), H\left(\frac{y'-x}{z'-x}\right)\right\}} \leqslant O\left(\frac{1}{n}\right)+\frac{1}{H\left(\frac{y'-x}{z'-x}\right)}. \quad (17)$$

We now see that Corollary 4 applied to (17) in combination with (16) gives the desired upper bound. ∎

V. Applications

In this section we apply the results from Section IV to show performance guarantees when using $\overline{h}$ and $\tilde{h}$ for hash tables with chaining, for min-wise hashing and for linear probing.

We first prove that Corollary 1 is a corollary of Theorem 1. The idea in the proof is that if there exists a large bucket, then there must be a high probability that three random keys have the same hash values.

Corollary 1. *Let $U=[u]$ be a universe of keys and $X\subset U$ a set of n keys such that $p>un$. Assume that $m\geqslant n$ and let M be the number of keys that hash to the most popular hash value, i.e. $M=\max_{v\in[m]}\left|\left\{x\in X \mid \overline{h}(x)=v\right\}\right|$. Then $E(M)\leqslant n^{1/3+o(1)}$. The same result holds when $\overline{h}$ is replaced with $\tilde{h}$ in the definition of M.*

Proof: Fix $\overline{h}$. Choosing $x,y,z\in X$ randomly the probability that all of the keys are in the same largest bucket is $\left(\frac{M}{n}\right)^3$. But the average probability of this over all choices of $\overline{h}$ is $O(m^{-2}+n^{-2+o(1)})$. Hence:

$$E\left(\left(\frac{M}{n}\right)^3\right)=O(m^{-2}+n^{-2+o(1)})=n^{-2+o(1)}. \quad (18)$$

Applying Jensen's inequality to the convex function $x\to x^3$ now gives $E(M)=n^{1/3+o(1)}$ as desired. ∎

The following restatement of Theorem 2 shows that even if we fix one hash value we can still use second order moment bounds. If $\tilde{h}$ had been 3-independent the upper bound in (20) could have been replaced with $O\left(\delta^{-2}\right)$. In most cases the $n^{o(1)}$ term in (20) will dominate the $\frac{n^2}{p\mu}$ term, and we will be able to prove bounds that are only worse by a factor of $n^{o(1)}$ than the corresponding bounds for 3-independent hash functions.

Theorem 6. *Let $U=[u]$ be a universe of keys and $X\subset U$ a set of n keys such that $p>un$. Let $v\in[m]$, $I\subset[m]$ be an interval and $x_0\in U\backslash X$. Let $A=\sum_{x\in X}[\tilde{h}(x)\in I]$. Then $\mu=E(A)$ satisfies*

$$\mu=E(A\mid\tilde{h}(x_0)=v)=\Theta\left(n\cdot\frac{|I|}{m}\right), \quad (19)$$

and for every $\delta>0$ it holds that:

$$P\left(|A-\mu|\geqslant\delta\sqrt{\mu}\mid\tilde{h}(x_0)=v\right) \leqslant \frac{1}{\delta^2}\cdot\left(n^{o(1)}+O\left(\frac{n^2}{p\mu}\right)\right). \quad (20)$$

Proof: (19) follows from the 2-independence of $\tilde{h}$. By Theorem 5:

$$E\left(A^2\mid\tilde{h}(x_0)=v\right) = n^2\cdot\left(\frac{|J|}{p}\right)^2+n^{1+o(1)}\cdot\frac{|J|}{p}+O\left(n^2p^{-1}\right). \quad (21)$$

Letting $\sigma^2=V(A\mid\tilde{h}(x_0)=v)$, we get:

$$\sigma^2=E\left(A^2\mid\tilde{h}(x_0)=v\right)-\mu^2 = \mu\cdot\left(n^{o(1)}+O\left(\frac{n^2}{p\mu}\right)\right).$$

By an application of Chebyshev's Inequaliy we now get Equation (20). ∎

Consider linear probing. In [4] it is proved that when using a 3-independent hash function the worst case expected query time for any element is $O(\log n)$. This is done by fixing the hash value of the element we query for, and then using second order moment bounds. By using Theorem 6 instead of the usual second order moment bound for 3-independent hash functions we get an upper bound that is $n^{o(1)}\cdot O(\log n)=n^{o(1)}$ instead. This proves Corollary 3.

Similarly, we can prove that $\tilde{h}$ performs almost as well as 3-independent hash functions when using it for min-wise hashing. This is done in Corollary 2 below. We should compare the upper bound of $n^{-1+o(1)}$ to the upper bound of $O\left(\frac{\log n}{n}\right)$ that holds for 3-independent hash functions [5].

Corollary 2. *Let $U=[u]$ be a universe of keys and $X\subset U$ a set of n keys such that $p>un$. Let $x_0\in X$. Then:*

$$P\left(\tilde{h}(x_0)<\min_{x\in X\backslash\{x_0\}}\left\{\tilde{h}(x)\right\}\right)\leqslant n^{-1+o(1)}.$$

Proof: Let $X' = X \setminus \{x_0\}$. For $v \in [m]$ let $I_v = [v+1]$ and let A_v be the random variable defined by:

$$A_v = \sum_{x \in X'} [\tilde{h}(x) \leqslant v] = \sum_{x \in X'} [\tilde{h}(x) \in I_v] .$$

Now we note that $\tilde{h}(x_0) < \min_{x \in X'} \left\{ \tilde{h}(x) \right\}$ iff there exists $v \in [m]$ such that $\tilde{h}(x_0) = v$ and $A_v = 0$. So we get:

$$\begin{aligned} & P\left(\tilde{h}(x_0) < \min_{x \in X'} \left\{ \tilde{h}(x) \right\} \right) \\ & = \sum_{v \in [m]} P\left(A_v = 0 \mid \tilde{h}(x_0) = v \right) \cdot P\left(\tilde{h}(x_0) = v \right) . \end{aligned}$$

Since $P\left(\tilde{h}(x_0) = v \right) = O\left(m^{-1} \right)$ Theorem 6 with $\delta = \sqrt{E(A_v)}$ gives:

$$\begin{aligned} & P\left(\tilde{h}(x_0) < \min_{x \in X'} \left\{ \tilde{h}(x) \right\} \right) \\ \leqslant & O\left(\frac{1}{m} \right) \cdot \sum_{v \in [m]} \min \left\{ 1, \frac{n^{-1+o(1)} m}{v+1} + O\left(\frac{m^2}{p(v+1)^2} \right) \right\} \\ = & n^{-1+o(1)} + O\left(\frac{1}{\sqrt{p}} \right) = n^{-1+o(1)} . \end{aligned}$$

■

ACKNOWLEDGMENT

The author would like to thank the anonymous referees for the most thorough review he has ever received. The author also thanks Rikke Langhede for introducing him to the concept of heights, and Mikkel Thorup for helpful comments.

REFERENCES

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (3. ed.)*. MIT Press, 2009.

[2] L. Carter and M. N. Wegman, "Universal classes of hash functions," *J. Comput. Syst. Sci.*, vol. 18, no. 2, pp. 143–154, 1979.

[3] I. Baran, E. D. Demaine, and M. Pătraşcu, "Subquadratic algorithms for 3SUM," *Algorithmica*, vol. 50, no. 4, pp. 584–596, 2008.

[4] M. Pătraşcu and M. Thorup, "On the *k*-independence required by linear probing and minwise independence," *ACM Trans. Algorithms*, vol. 12, no. 1, p. 8, 2016.

[5] M. B. T. Knudsen and M. Stöckel, "Quicksort, largest bucket, and min-wise hashing with limited independence," in *Algorithms - ESA 2015 - 23rd Annual European Symposium, Patras, Greece, September 14-16, 2015, Proceedings*, 2015, pp. 828–839.

[6] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos, "Linear hash functions," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 667–683, 1999.

[7] E. Bombieri and W. Gubler, *Heights in Diophantine geometry*. Cambridge University Press, 2007, vol. 4.

[8] T. Tao, "Lecture notes 2 for 254a." [Online]. Available: https://www.math.cmu.edu/ af1p/Teaching/AdditiveCombinatorics/Tao.pdf

[9] K. K. Norton, "Upper bounds for sums of powers of divisor functions," *Journal of Number Theory*, vol. 40, no. 1, pp. 60–85, 1992.