Universal Classes of Hash Functions

Nitin Verma
mathsanew.com

February 25, 2021

In this article, we will find out how a certain proportion of hash functions in a *Universal Class* must have some properties.

Consider a set $\mathcal{H}$ of hash-functions, each of which map the set of all possible keys U to $\{0, 1, 2, \ldots, m-1\}$, where m is a positive integer. $\mathcal{H}$ is called a *Universal Class* if any two distinct keys $x, y \in U$ map to the same value by at most $|\mathcal{H}|/m$ functions in $\mathcal{H}$. In other words, the set of functions where x and y collide:

$$\{h \in \mathcal{H} \mid h(x) = h(y)\}$$

has at most $|\mathcal{H}|/m$ functions.

We will denote $|\mathcal{H}|$ by H. For any non-negative integer i, the set $\{0, 1, 2, \ldots, i-1\}$ will be denoted as $\mathbb{Z}_i$, and the set $\{1, 2, \ldots, i-1\}$ as $\mathbb{Z}_i^*$.

Some set of hash-functions may demonstrate the above property for a proportion of functions which is not necessarily $1/m$. Say this proportion is ϵ, for some real number ϵ in $(0, 1)$. We may call such set as ϵ-*Universal Class*.

The *Universal Class* (without ϵ specified) defined above has proportion $\epsilon = 1/m$, and so we will refer them as "$1/m$-Universal". In our derivation of properties, we will consider the more general ϵ-Universal classes so that the results are more useful.

Below are two examples of ϵ-Universal Classes. The set of all possible keys is: $U = \{0, 1, 2, \ldots, u-1\}$, for some positive integer u. p is a prime, $p \geq u$.

1. For each $a \in \mathbb{Z}_p^*$, $b \in \mathbb{Z}_p$, define:

$$h_{a,b}(x) = ((ax + b) \bmod p) \bmod m$$
$$\mathcal{H} = \{h_{a,b} \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p\}$$

   $\mathcal{H}$ is a $1/m$-Universal class.

2. For each $a \in \mathbb{Z}_p^*$, define:

$$h_a(x) = ((ax) \bmod p) \bmod m$$
$$\mathcal{H} = \{h_a \mid a \in \mathbb{Z}_p^*\}$$

   $\mathcal{H}$ is a $2/m$-Universal class.

Given any distinct keys x and y, these universal classes, due to their definition, carry an upper-bound on the number of their functions where x, y collide. It is this upper bound, ϵH, which we will utilize below to count the functions in the class with certain properties.

The kinds of relations as discussed here, were originally derived and utilized in a method of *Perfect Hashing* called *FKS Method*, by Fredman, Komlós and Szemerédi in [1]. The relations and proofs from [1] are also elaborated nicely in [2]. In these, the class of hash-functions considered is the one from example-2 above. In book [3], chapter "Hash Tables", we find derivation of similar relations for $1/m$-Universal Classes. The relations and proofs presented in this article are adaptation from these sources, and are for general ϵ-Universal classes.

Counting Collision-Free Functions

Consider a ϵ-Universal class $\mathcal{H}$ of H functions, mapping U to $\mathbb{Z}_m$. Say we are given a set S of n keys from U. Are there any functions in $\mathcal{H}$ which give no collision for all keys in S? We will refer to such functions as "Collision-Free" for S. So, these functions map the n distinct keys to n distinct values of $\mathbb{Z}_m$. Such a function would exist only if $|\mathbb{Z}_m| \geq |S|$, i.e. $m \geq n$.

The definition of ϵ-Universal class provides an upper bound of ϵH on the number of functions where any key pair would collide. We can form $\binom{n}{2}$ key pairs from S. Thus, the maximum number of functions where any of these $\binom{n}{2}$ pairs can collide is:

$$\binom{n}{2}\epsilon H \tag{1}$$

Note that this is only an upper-bound and can even exceed the available functions count of H. The subset of functions where a pair collides may overlap with the subset of functions where some other pair collides. Since above we simply added maximum size of each such subset (i.e. ϵH), once for each pair, this upper bound may be loose.

Based on (1), the minimum number of functions in $\mathcal{H}$ which must be collision-free for S, if $\binom{n}{2}\epsilon H \leq H$:

$$H - \binom{n}{2}\epsilon H$$

So, the minimum proportion of such functions in $\mathcal{H}$:

$$\alpha = \frac{\left(H - \binom{n}{2}\epsilon H\right)}{H} = 1 - \binom{n}{2}\epsilon \tag{2}$$

Although (2) need not provide a tight lower bound on the proportion of functions which are collision-free for a given set of n keys, it can still be useful. It relates the lower-bound α to ϵ, and it may be possible to adjust the parameters of the universal class to achieve certain ϵ. For example, for a $1/m$-Universal class, $\epsilon = 1/m$ can be modified by simply modifying the table-size m.

Suppose we want to ensure the lower bound α is at least $1/2$; so at least half of the functions in $\mathcal{H}$ are collision-free for a given set of n keys. Then (2) gives:

$$1 - \binom{n}{2}\epsilon \geq \frac{1}{2} \quad \Leftrightarrow \quad \epsilon \leq \frac{1}{n(n-1)} \tag{3}$$

For any $1/m$-Universal class, (3) becomes:

$$\frac{1}{m} \leq \frac{1}{n(n-1)} \quad \Leftrightarrow \quad m \geq n(n-1)$$

And for $2/m$-Universal class, (3) becomes:

$$\frac{2}{m} \leq \frac{1}{n(n-1)} \quad \Leftrightarrow \quad m \geq 2n(n-1)$$

To make α at least $f \in (0,1)$, we need:

$$1 - \binom{n}{2}\epsilon \geq f \quad \Leftrightarrow \quad \epsilon \leq \frac{2(1-f)}{n(n-1)}$$

Theorem 1. *Given ϵ-Universal class $\mathcal{H}$ and any set of n keys, the proportion of functions in $\mathcal{H}$ which are collision-free for those keys is at least $f \in (0,1)$ if:*

$$\epsilon \leq \frac{2(1-f)}{n(n-1)}.$$

Corollary 2. *Given ϵ-Universal class $\mathcal{H}$ with $\epsilon = 1/m$, and any set of n keys, at least half of the functions in $\mathcal{H}$ are collision-free for those keys if: $m \geq n(n-1)$. For $\epsilon = 2/m$, this condition is: $m \geq 2n(n-1)$.*

Counting Colliding-Pairs

Now we perform another counting, for the key-pairs which collide. For any distinct keys $x, y \in S$ and $h \in \mathcal{H}$, we will call the pair (x, y) a "colliding-pair under h" if $h(x) = h(y)$. We will not distinguish between pairs (x, y) and (y, x). Among the total $\binom{n}{2}$ pairs in S, can we say anything about the number of colliding-pairs under h? Let us denote the set of all $\binom{n}{2}$ pairs as P, and the number of colliding-pairs under h as C_h.

The definition of ϵ-Universal class only tells us about the maximum number of functions under which a pair in P becomes a colliding-pair (ϵH). For a particular function h, we don't know how many pairs from P will become a colliding-pair. But if we count the number of colliding-pairs for each $h \in \mathcal{H}$ and add them together, we can progress as below:

$$\begin{aligned}
\sum_{h \in \mathcal{H}} C_h &= \sum_{h \in \mathcal{H}} (\text{Number of } p \in P \text{ such that } p \text{ is colliding-pair under } h) \\
&= \sum_{h \in \mathcal{H}} |\{p \in P \mid p \text{ is a colliding-pair under h}\}| \\
&\quad \{\text{interestingly, this count can also be performed as below}\} \\
&= \sum_{p \in P} (\text{Number of } h \in \mathcal{H} \text{ such that } p \text{ is colliding-pair under } h) \\
&= \sum_{p \in P} |\{h \in \mathcal{H} \mid p \text{ is a colliding-pair under h}\}| \\
&\quad \{\text{any pair collides under maximum } \epsilon H \text{ functions}\} \\
&\leq \sum_{p \in P} \epsilon H \\
&= \binom{n}{2} \epsilon H
\end{aligned}$$

Thus,

$$\frac{\sum_{h\in\mathcal{H}} C_h}{H} \leq \binom{n}{2}\epsilon \tag{4}$$

In words, the number of colliding-pairs for a function (C_h), averaged over all functions in $\mathcal{H}$, is maximum $\binom{n}{2}\epsilon$.

Note that for all $h \in \mathcal{H}$, $0 \leq C_h \leq \binom{n}{2}$. It is easy to prove that among any H non-negative numbers, with average A, at least $\lceil H/2 \rceil$ numbers must be less than $2A$. So, we can conclude:

Theorem 3. *Given ϵ-Universal class $\mathcal{H}$ and any set of n keys, at least half of the functions in class $\mathcal{H}$ must have number of colliding-pairs $C_h < 2\binom{n}{2}\epsilon = n(n-1)\epsilon$.*

This provides an upper-bound on C_h attained by at least half of the functions. So, to make sure that at least half of the functions are collision-free for a given set of n keys, i.e. have $C_h = 0$, we can restrict this upper-bound to be 1 (C_h are integers):

$$n(n-1)\epsilon \leq 1 \quad \Leftrightarrow \quad \epsilon \leq \frac{1}{n(n-1)}$$

and adjust our universal-class to achieve such ϵ. Note that this relation is same as equation (3) obtained in the last section.

Similarly, we know that at least one function must have C_h not exceeding the average of all C_h. So, to make sure that at least one function is collision-free ($C_h = 0$), we can restrict this average to be less than 1:

$$\binom{n}{2}\epsilon < 1 \quad \Leftrightarrow \quad \epsilon < \frac{2}{n(n-1)}.$$

From C_h to s_i^2

For a function $h \in \mathcal{H}$, let us denote by S_i the set of keys from S which are mapped to $i \in \mathbb{Z}_m$ by h. Say $s_i = |S_i|$. So, the number of colliding-pairs in slot i is $\binom{s_i}{2}$. Hence,

$$\begin{aligned} C_h &= \sum_{i \in \mathbb{Z}_m} \binom{s_i}{2} \\ &= \sum_{i \in \mathbb{Z}_m} \frac{s_i^2 - s_i}{2} \\ &= \frac{1}{2} \left(\sum_{i \in \mathbb{Z}_m} s_i^2 - \sum_{i \in \mathbb{Z}_m} s_i \right) \\ &= \frac{1}{2} \left(\sum_{i \in \mathbb{Z}_m} s_i^2 - n \right) \end{aligned}$$

So the following inequality from theorem 3 can be rewritten:

$$\begin{aligned} & C_h < n(n-1)\epsilon \\ \Leftrightarrow \quad & \frac{1}{2} \left(\sum_{i \in \mathbb{Z}_m} s_i^2 - n \right) < n(n-1)\epsilon \\ \Leftrightarrow \quad & \sum_{i \in \mathbb{Z}_m} s_i^2 < 2n(n-1)\epsilon + n \end{aligned}$$

Corollary 4. *Given ϵ-Universal class $\mathcal{H}$ and any set of n keys, if s_i is the number of keys mapped to slot i by a function $h \in \mathcal{H}$, at least half of the functions in class $\mathcal{H}$ must have s_i such that:*

$$\sum_{i \in \mathbb{Z}_m} s_i^2 < 2n(n-1)\epsilon + n.$$

This upper-bound on the sum of s_i^2 finds its use for optimizing space in the FKS Method of Perfect Hashing [1].

■

References

[1] M. L. Fredman, J. Komlós, E. Szemerédi. *Storing a Sparse Table with O(1) Worst Case Access Time*. J. ACM, Vol 31 (3) (1984), 538–544.

[2] Z. J. Czech, G. Havas, B. S. Majewski. *Fundamental Study: Perfect Hashing*. Theoretical Computer Science, Vol 182 (1–2) (1997), 1–143.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein. *Introduction to Algorithms*, Third Edition. The MIT Press, 2009.