c-UNIVERSAL FAMILIES OF HASH FUNCTIONS

Definition. Let c be any positive constant.

A family $\mathcal{H}$ of hash functions $h : U \to \{0, \ldots, m-1\}$ is *c-universal* if for all $x, y \in U$ $(x \neq y)$

$$|\{h \in \mathcal{H} : h(x) = h(y)\}| \leq c \cdot |\mathcal{H}|/m.$$

c-universal families of hash functions share most of the important properties of universal families of hash functions. In fact, if we set $c = 1$ and replace the $\leq$ by an $=$ in the definition above we have the definition for a universal family of hash functions. In general the smaller c, the fewer collisions we expect.

Remark. The last example given in class (with $h_a(x) = ((ax \mod p) \mod m)$) contained a mistake in the proof (find it!). The family $\mathcal{H} = \{h_a : 1 \leq a < p\}$ is in general not universal, but one can show that this family is 2-universal (by fixing the proof). As an example look at the simple case when $p = 5$, $m = 2$ – this example is not universal, but it is 2-universal – as a matter of fact it is even 1-universal!

Here is a better example: Let $U = \{0, 1, \ldots, p-1\}$, for p a prime and $m \leq p$. Define $h_{a,b}(x) = ((ax + b \mod p) \mod m)$ and $\mathcal{H}_1 = \{h_{a,b} : a, b \in U, a \neq 0\}$.

Theorem 1. *$\mathcal{H}_1$ is a 1-universal family of hash functions.*

Proof. Consider distinct elements $x, y \in U$. We must show that

$$|\{h \in \mathcal{H}_1 : h(x) = h(y)\}| \leq |\mathcal{H}_1|/m = p(p-1)/m.$$

If $h_{a,b}(x) = h_{a,b}(y)$, then there must be integers $q \in \{0, 1, \ldots, m-1\}$ and $r \neq s$ (with $0 \leq q + rm < p$ and $0 \leq q + sm < p$) such that

$$ax + b \equiv q + rm \mod p$$
$$ay + b \equiv q + sm \mod p.$$

By subtracting the two congruences we conclude that $a(x-y) \equiv (r-s)m \mod p$, so that $a = ((x-y)^{-1}(r-s)m \mod p)$ and thus $b = (q + rm - ax \mod p)$. So for every choice of q, r, s there is at most one choice for a and b. But altogether there are at most p possible choices for (q, r) – and for every choice of q and r there are at most $(p-1)/m$ choices for s: the integers in $[0, (p-1-q)/m]$, except r. $\square$

It can be seen from the proof that $\mathcal{H}_1$ is not universal, but the following family, which you may have seen in tutorial, is universal. It is mentioned here since one of your homework problems refers to it.

Let $U = \mathbb{Z}_2^k$ and $m = 2^\ell$ for some $\ell \leq k$. If B is a binary $\ell \times k$ matrix, then the linear transformation $h_B : U \to \mathbb{Z}_2^\ell$, $h_B(x) = Bx$, can be viewed as a hash function, since $\mathbb{Z}_2^\ell$ has exactly 2^ℓ elements. Let $\mathcal{H}_2 = \{h_B : B \text{ is a binary } \ell \times k \text{ matrix}\}$.

Theorem 2. *$\mathcal{H}_2$ is a universal family of hash functions.*

Proof. Let $x, y \in U$ be distinct keys, and $h_B \in \mathcal{H}_2$. Now

$$h_B(x) = h_B(y) \iff 0 = h_B(y) - h_B(x) = h_B(y - x) \qquad (*)$$

where the last step follows from linearity. Furthermore, since $x \neq y$ we can assume that $x_r \neq y_r$ for some $1 \leq r \leq k$. We will assume that $r = 1$, since for all other r the proof is similar. Thus $y_1 - x_1 = 1$ in $\mathbb{Z}_2$.

Next let the entries of B be denoted by b_{ij}, where $1 \leq i \leq \ell$ and $1 \leq j \leq k$. The ith row of expression $(*)$ is exactly $0 = \sum_{1 \leq j \leq k} b_{ij}(y_j - x_j)$, or equivalently $-\sum_{2 \leq j \leq k} b_{ij}(y_j - x_j) = b_{i1}(y_1 - x_1) = b_{i1}$.

This condition being fulfilled for all $1 \leq i \leq \ell$ is equivalent to $h_B(x) = h_B(y)$. So for every possible choice for b_{ij} ($2 \leq j \leq k$ and $1 \leq i \leq \ell$) there is a unique choice for b_{i1} ($1 \leq i \leq \ell$) such that x and y collide. We conclude that

$$|\{h \in \mathcal{H}_2 : h(x) = h(y)\}| = 2^{\ell(k-1)} = 2^{\ell k}/2^\ell = |\mathcal{H}_2|/m,$$

so that $\mathcal{H}_2$ is indeed universal. $\square$