Universal hashing
Ben Langmead

Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).
Randomness & independence

<table>
<thead>
<tr>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_2$</th>
<th>$h_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>73735</td>
<td>45963</td>
<td>78134</td>
<td>63873</td>
</tr>
<tr>
<td>02965</td>
<td>58303</td>
<td>90706</td>
<td>20025</td>
</tr>
<tr>
<td>98859</td>
<td>23851</td>
<td>27965</td>
<td>62394</td>
</tr>
<tr>
<td>33666</td>
<td>62570</td>
<td>64775</td>
<td>78428</td>
</tr>
<tr>
<td>81666</td>
<td>26440</td>
<td>20422</td>
<td>05720</td>
</tr>
<tr>
<td>15838</td>
<td>47174</td>
<td>76866</td>
<td>14330</td>
</tr>
<tr>
<td>89793</td>
<td>34378</td>
<td>08730</td>
<td>56522</td>
</tr>
<tr>
<td>78155</td>
<td>22466</td>
<td>81978</td>
<td>57323</td>
</tr>
<tr>
<td>16381</td>
<td>66207</td>
<td>11698</td>
<td>99314</td>
</tr>
<tr>
<td>75002</td>
<td>80827</td>
<td>53867</td>
<td>37797</td>
</tr>
<tr>
<td>99982</td>
<td>27601</td>
<td>62686</td>
<td>44711</td>
</tr>
<tr>
<td>84543</td>
<td>87442</td>
<td>50033</td>
<td>14021</td>
</tr>
<tr>
<td>77757</td>
<td>54043</td>
<td>46176</td>
<td>42391</td>
</tr>
<tr>
<td>80871</td>
<td>32792</td>
<td>87989</td>
<td>72245</td>
</tr>
<tr>
<td>30500</td>
<td>28220</td>
<td>12444</td>
<td>71840</td>
</tr>
</tbody>
</table>
Universal hashing

A family of hash functions \( H \) from universe \( U \) with \( |U| \geq n \) to range \( \{0, 1, \ldots, n - 1\} \) is **2-universal** if for distinct elements \( x_1, x_2 \) and for function \( h \) drawn uniformly from \( H \):

\[
\Pr \left( h(x_1) = h(x_2) \right) \leq \frac{1}{n}
\]

Let's prove a useful expectation for hash tables…
Universal hashing

A set $S$ of $m$ items have been hashed to an $n$-bucket hash table using $h$ from a 2-universal family.

For given element $x$ let r.v. $X$ be the number of items in bucket $h(x)$. We want to show:

$$E[X] \leq \begin{cases} 
\frac{m}{n} & \text{if } x \notin S \\
1 + \frac{(m - 1)}{n} & \text{if } x \in S 
\end{cases}$$

Not-in-table case

$1$ if $m = n$

In-table case

$< 2$ if $m = n$
Universal hashing

\[ E[X] \leq \begin{cases} 
  m/n & \text{if } x \notin S \\
  1 + (m - 1)/n & \text{if } x \in S 
\end{cases} \]

Let \( X_i \) be a r.v. \( X_i = 1 \) when the \( i^{th} \) element of \( S \) is in same bucket as \( x \). \( X_i = 0 \) otherwise

\[ \Pr(X_i = 1) \leq \frac{1}{n} \quad \text{By 2-universality!} \]
Universal hashing

\[ x \not\in S \text{ case} \]

\[
\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbb{E}[X_i] \leq \frac{m}{n}
\]

**Linearity**

**2-universality**

+ expectation of indicator

\[
\mathbb{E}[X_i] = \Pr(X_i = 1) \leq \frac{1}{n}
\]
Universal hashing

$$E[X] \leq \begin{cases} 
\frac{m}{n} & \text{if } x \notin S \\
1 + \frac{(m - 1)}{n} & \text{if } x \in S 
\end{cases}$$

Let $X_i$ be a r.v. $X_i = 1$ when the $i^{th}$ element of $S$ is in same bucket as $x$. $X_i = 0$ otherwise

Without loss of generality, use $i = 1$ for item $x$

$$\Pr(X_i = 1) \leq \frac{1}{n} \quad \text{for } i > 1$$
Universal hashing

\[ x \in S \text{ case} \]

\[
\mathbb{E}[X] = \mathbb{E} \left[ \sum_{i=1}^{m} X_i \right] = 1 + \sum_{i=2}^{m} \mathbb{E}[X_i] \leq 1 + \frac{m - 1}{n}
\]

2-universality

+ expectation of indicator
Universal hashing

Proving a key property; with 2-universal hashing, expected query time is $O(1)$ when $m \leq n$

$$
E[X] \leq \begin{cases} 
m/n & \text{if } x \notin S \\
1 + (m - 1)/n & \text{if } x \in S
\end{cases}
$$

Not-in-table case

1 if $m = n$

In-table case

$\sim 2$ if $m = n$
Universal hashing

What kind of family has this property?

Are functions easy to draw from the family?

Are functions easy to store and compute with?
Universal hashing

Universe $U : \{0, 1, 2, \ldots, m - 1\}$

Range $V : \{0, 1, 2, \ldots, n - 1\}$ with $n \leq m$

Prime $p \geq m$
Universal hashing

Example of a 2-universal family from $U$ to $V$:

$$H = \{h_{a,b} \mid 1 \leq a \leq p - 1, \ 0 \leq b \leq p - 1\}$$

$$h_{a,b}(x) = ((ax + b) \mod p) \mod n$$
A prime field $\mathbb{F}_p$ is a number system consisting of integers modulo a prime $p$, and rules for plus & times

Plus & times have many of our favorite properties
Prime field

Fields are special for having *multiplicative inverses*

Each number (except 0) has another it multiplies with to get 1

\[
\begin{align*}
2 \cdot 3 &= 3 \cdot 2 = 1 \mod 5 \\
4 \cdot 4 &= 1 \mod 5 \\
1 \cdot 1 &= 1 \mod 5
\end{align*}
\]

\[
\begin{array}{c|cccccc}
\times & 0 & 1 & 2 & 3 & 4 \\\n\hline
0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 2 & 3 & 4 \\
2 & 0 & 2 & 4 & 1 & 3 \\
3 & 0 & 3 & 1 & 4 & 2 \\
4 & 0 & 4 & 3 & 2 & 1 \\
\end{array}
\]
Prime field

Does modulo a non-prime work?

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

F\text{?}6

Signs of trouble. 1) We sometimes get 0s when multiplying non-0s
Prime field

Does modulo a non-prime work?

![Table](image)

F? 6

Signs of trouble. 1) We sometimes get 0s when multiplying non-0s

2) Some rows don't have 1; no multiplicative inverse
Prime field

\[ \mathbb{F}_7 \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Universal hashing

Choose distinct $x_1, x_2 \in U$. Can they collide in $p$?

$$ax_1 + b \stackrel{?}{=} ax_2 + b \mod p$$

$$ax_1 + b = ax_2 + b \mod p$$

$$ax_1 = ax_2 \mod p$$

$$a(x_1 - x_2) = 0 \mod p$$

We said $a \geq 1$ and $x_1 \neq x_2$

Left side is product of two numbers and neither is $0 \mod p$. 
Prime field

\[
\mathbb{F}_7
\]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

No 0s
Universal hashing

Can \( ac = zp \), where \( p \) is a prime, \( z \) is some integer multiple, and \( a \& c \) are not \( 0 \mod p \)?

\[
ac = zp
\]

Consider prime factorizations of \( a \) and \( c \)

For equality to hold, \( p \) must be a prime factor of \( a \) or \( c \), contradicting "\( a \& c \) are not \( 0 \mod p \)"

\[
ac \neq zp
\]
Universal hashing

Fact 1: Distinct items from $U$ won't collide in prime field

$$(ax + b) \mod p$$
Universal hashing

\[(ax + b) \mod 5\]

Each column is a permutation of integers mod 5

Copied from \(F_5 \times \text{table}\)
Universal hashing

| b | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| a | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| x | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

\[(ax + b) \mod 5\]

Same as block to left but + 1 mod 5
Universal hashing

\[(ax + b) \mod 5\]

<table>
<thead>
<tr>
<th>$b$</th>
<th>0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4</td>
</tr>
<tr>
<td>$x$</td>
<td>0 0 0 0 1 1 1 1 2 2 2 2 3 3 4 1 3 0 3 1 4</td>
</tr>
<tr>
<td></td>
<td>1 1 2 3 4 2 3 4 0 3 4 0 1</td>
</tr>
<tr>
<td></td>
<td>2 2 4 1 3 3 0 2 4 4 1 3 0</td>
</tr>
<tr>
<td></td>
<td>3 3 1 4 2 4 2 0 3 0 3 1 4</td>
</tr>
<tr>
<td></td>
<td>4 4 3 2 1 0 4 3 2 1 0 4 3</td>
</tr>
</tbody>
</table>
Universal hashing

Every column is a permutation of integers mod 5. Therefore: no collisions. Distinct $x$s get distinct answers.
### Universal hashing

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

\[(ax + b) \mod 4\]

Is every column necessarily a permutation of another column?
Universal hashing

Given $x_1, x_2, u, v$, what is the chance that $h_{a,b}(x_1) = u$ and $h_{a,b}(x_2) = v$?
Universal hashing

\[(a \cdot x_1 + b) = u \mod p\]

\[(a \cdot x_2 + b) = v \mod p\]

\[a = \frac{v - u}{x_2 - x_1} \mod p\]

\[b = u - ax_1 \mod p\]

Fact 2: Single choice of \(a, b\) satisfies the equations. 
\[0 \leq b \leq p - 1\] and \[1 \leq a \leq p - 1\], so chance is \[\frac{1}{p(p-1)}\]

\(u, v\) pairs are equally likely
Universal hashing

$((ax + b) \mod p) \mod n$

Last concern: collisions from $\text{final mod } n$
Universal hashing

Taking a number \( u \) in the prime field, the others \( \pm zn \) are its colliders w/r/t \( V \)

For \( p = 11 \) & \( n = 4 \), 20 out of 110 \( u, v \) pairs collide (red squares)
Universal hashing

For given $u$, number of possible $v$'s ($u \neq v$) is $p - 1$, all equally likely.

At most $\lceil p/n \rceil - 1$ choices are collisions.

\[
\Pr(h_{a,b}(x_1) = h_{a,b}(x_2)) \leq \frac{\lceil p/n \rceil - 1}{p - 1} \leq \frac{(p - 1)/n}{p - 1} = \frac{1}{n}
\]

2-universality ✓