

Honors Algorithms  
G22.3520-001 Fall 2007

Lecture 2

## Recall: Universal Hashing

$\mathcal{K}$  – a finite, non-empty set of **hash keys**

$\mathcal{H} = \{h_k\}_{k \in \mathcal{K}}$  – a **family** of hash functions

$h_k : \mathcal{U} \rightarrow \{0, \dots, m-1\}$ , indexed by  $\mathcal{K}$

**Def'n:**  $\mathcal{H}$  is called **universal** if for all  $a, b \in \mathcal{U}$  with  $a \neq b$ ,

$$|\{k \in \mathcal{K} : h_k(a) = h_k(b)\}| \leq \frac{|\mathcal{K}|}{m}.$$

**Probabilistic interpretation:** if  $R$  is a random variable, uniformly distributed over  $\mathcal{K}$ , then

$$\Pr[h_R(a) = h_R(b)] \leq \frac{1}{m}.$$

# Constructing Universal Families

Arithmetic modulo  $n$ :

- $\mathbb{Z}_n$  — residue classes mod  $n$ :
  - $\mathbb{Z}_n = \{[0], [1], \dots, [n-1]\}$
  - $[a] + [b] := [(a + b) \bmod n]$
  - $[a] \cdot [b] := [(a \cdot b) \bmod n]$
- $\mathbb{Z}_n$  is a *ring*:
  - $+$ ,  $\cdot$  are commutative, associative
  - $\cdot$  distributes over  $+$
  - $[0]$  is the additive identity
  - $[1]$  is the multiplicative identity

## Arithmetic modulo a prime $p$ :

- $\mathbb{Z}_p$  is a *field*:
  - Every non-zero  $\alpha \in \mathbb{Z}_p$  has a multiplicative inverse  $\beta \in \mathbb{Z}_p$  (i.e,  $\alpha\beta = 1$ )
  - More generally: if  $\alpha, \beta \in \mathbb{Z}_p$  with  $\alpha \neq 0$ , then the equation

$$\alpha x = \beta$$

has a unique solution  $x$

## A universal family

Let  $m$  be a prime, and  $t$  a positive integer

Define  $\mathcal{U} := \mathbb{Z}_m^{t+1}$ ,  $\mathcal{K} := \mathbb{Z}_m^t$

For  $k = (k_1, \dots, k_t) \in \mathcal{K}$ ,  $a = (a_0, a_1, \dots, a_t) \in \mathcal{U}$ ,  
define

$$h_k(a) := a_0 + \sum_{i=1}^t a_i k_i$$

Define

$$\mathcal{H} := \{h_k\}_{k \in \mathcal{K}}$$

**Theorem:**  $\mathcal{H}$  is universal

*Proof of theorem.*

Suppose  $(a_0, \dots, a_t) \neq (b_0, \dots, b_t)$

We want to count the number  $N$  of solutions  $(k_1, \dots, k_t)$  to the equation

$$a_0 + \sum_i a_i k_i = b_0 + \sum_i b_i k_i.$$

Re-write this as

$$c_0 + \sum_i c_i k_i = 0$$

where  $c_i := a_i - b_i$

By assumption, not all  $c_i$ 's are zero

Want to show:  $N \leq |\mathcal{K}|/m = m^{t-1}$

*Proof (cont'd).*

**Case 1:**  $c_i = 0$  for all  $i = 1, \dots, t$

- $N = 0$

**Case 2:**  $c_j \neq 0$ , for some  $j = 1, \dots, t$

- For every choice of  $k_1, \dots, k_{j-1}, k_{j+1}, \dots, k_t$ , there is a unique  $k_j$  such that  $(k_1, \dots, k_t)$  is a solution (since  $\mathbb{Z}_m$  is a field)
- Therefore,  $N = m^{t-1}$

QED

## Practical considerations

Input space:

- View data items as bit strings of some fixed length  $\ell$
- Break up into “chunks” of length  $w$ , where  $2^w \leq m < 2^{w+1}$ :



- View each  $s_i$  as a number between 0 and  $2^w - 1$ , and each such number as an element of  $\mathbb{Z}_m$
- This map from  $\{0, 1\}^\ell$  to  $\mathbb{Z}_m^{t+1}$  is injective
- Variable length inputs (padding)

## Practical considerations (cont'd)

Output space, choice of prime  $m$

**Bertrand's Postulate:** There is always a prime between  $x$  and  $2x$  for all integers  $x \geq 1$

## Another universal family

Let  $p$  be a prime, and  $m$  a positive integer

Define  $\mathcal{U} := \{0, \dots, p-1\}$ ,

$$\mathcal{K} := \{1, \dots, p-1\} \times \{0, \dots, p-1\}$$

For  $k = (k_1, k_2) \in \mathcal{K}$ ,  $a \in \mathcal{U}$ , define

$$h_k(a) := \left( (k_1 a + k_2) \bmod p \right) \bmod m$$

**Theorem:**  $\mathcal{H} := \{h_k\}_{k \in \mathcal{K}}$  is universal (see text)

Pros: free choice of  $m$

Cons: multiplication of large numbers

**General problem:** large key space — almost as large as the input space

**Solution:** weaker (but still useful) hashing requirements

## $\epsilon$ -universal Hashing

$\mathcal{K}$  – a finite, non-empty set of **hash keys**

$\mathcal{H} = \{h_k\}_{k \in \mathcal{K}}$  – a **family** of hash functions

$h_k : \mathcal{U} \rightarrow \{0, \dots, m - 1\}$ , indexed by  $\mathcal{K}$

**Def'n:** Let  $0 \leq \epsilon \leq 1$ .  $\mathcal{H}$  is called  **$\epsilon$ -universal** if for all  $a, b \in \mathcal{U}$  with  $a \neq b$ ,

$$|\{k \in \mathcal{K} : h_k(a) = h_k(b)\}| \leq \epsilon \cdot |\mathcal{K}|.$$

**Probabilistic interpretation:** if  $R$  is a random variable, uniformly distributed over  $\mathcal{K}$ , then

$$\Pr[h_R(a) = h_R(b)] \leq \epsilon$$

universal =  $(1/m)$ -universal

## Using $\epsilon$ -universal hash families

As long as  $\epsilon$  is not too big, many of the results we proved have useful analogs.

E.g., in a table with at most  $n$  data items

- expected cost of each dictionary operation in a table containing  $n$  items is  $\leq 1 + \epsilon n$ .
- expected value of maximum load is  $\leq \sqrt{\epsilon n^2 + n}$

## Building $\epsilon$ -universal hash families

We can get by with much shorter hash keys

## More about modular arithmetic

Let  $p$  be a prime

Let  $f = c_d X^d + c_{d-1} X^{d-1} + \dots + c_1 X + c_0$  be a polynomial with coefficients in  $\mathbb{Z}_p$ , with  $c_d \neq 0$  (the *degree* of  $f$  is  $d$ )

Then  $f$  has at most  $d$  roots in  $\mathbb{Z}_p$ :

$$\left| \{u \in \mathbb{Z}_p : \sum_{i=0}^d c_i u^i = 0\} \right| \leq d.$$

This is a general fact that holds for any *field* (e.g., the reals, the complex numbers,  $\mathbb{Z}_p$ ), but not for arbitrary rings (e.g.,  $\mathbb{Z}_n$ ).

## An $\epsilon$ -universal family

Let  $m$  be a prime, and  $t$  a positive integer

Define  $\mathcal{U} := \mathbb{Z}_m^{t+1}$ ,  $\mathcal{K} := \mathbb{Z}_m$

For  $k \in \mathcal{K}$ ,  $a = (a_0, a_1, \dots, a_t) \in \mathcal{U}$ , define

$$h_k(a) := \sum_{i=0}^t a_i k^i$$

Define

$$\mathcal{H} := \{h_k\}_{k \in \mathcal{K}}$$

**Theorem:**  $\mathcal{H}$  is  $(t/m)$ -universal

*Proof.* Suppose  $(a_0, \dots, a_t) \neq (b_0, \dots, b_t)$

We want to count the number  $N$  of solutions  $k$  to the equation

$$\sum_i a_i k^i = \sum_i b_i k^i.$$

Re-write this as

$$\sum_i c_i k^i = 0$$

where  $c_i := a_i - b_i$

$N = \#$  roots of  $\sum_i c_i X^i$ , which is a non-zero polynomial of degree at most  $t$

$\therefore N \leq t = (t/m) \cdot m.$      QED

## Pairwise Independent Hashing: a stronger notion

$\mathcal{K}$  – a finite, non-empty set of **hash keys**

$\mathcal{H} = \{h_k\}_{k \in \mathcal{K}}$  – a **family** of hash functions  
 $h_k : \mathcal{U} \rightarrow \{0, \dots, m-1\}$ , indexed by  $\mathcal{K}$

**Def'n:**  $\mathcal{H}$  is called **pairwise independent** if for all  $a, b \in \mathcal{U}$  with  $a \neq b$ , and for all  $r, s \in \{0, \dots, m-1\}$ , we have

$$|\{k \in \mathcal{K} : h_k(a) = r \text{ and } h_k(b) = s\}| = \frac{|\mathcal{K}|}{m^2}.$$

## Probabilistic interpretation

Let  $R$  be a random variable, uniformly distributed over  $\mathcal{K}$

For each  $a \in \mathcal{U}$ , define the random variable  
 $V_a := h_R(a)$

**Claim:** if  $|\mathcal{U}| > 1$ , then each  $V_a$  is uniformly distributed over  $\{0, \dots, m-1\}$

*Proof.* Let  $a \in \mathcal{U}$ ,  $r \in \{0, \dots, m-1\}$

Let  $b \in \mathcal{U}$ ,  $b \neq a$ . Then:

$$\begin{aligned}\Pr[V_a = r] &= \sum_s \Pr[V_a = r \text{ and } V_b = s] \\ &= \sum_s 1/m^2 = 1/m.\end{aligned}$$

Recall:  $X$  and  $Y$  are **independent** if for all possible  $x$  and  $y$ ,

$$\Pr[X = x \text{ and } Y = y] = \Pr[X = x] \Pr[Y = y],$$

or equivalently

$$\Pr[X = x] = \Pr[X = x \mid Y = y].$$

A family of random variables  $\{X_i\}_{i \in \mathcal{I}}$  is called **pairwise independent** if  $X_i$  and  $X_j$  are independent for all  $i \neq j$

The family of random variables  $\{V_a\}_{a \in \mathcal{U}}$  is pairwise independent:

$$\begin{aligned}\Pr[V_a = r \text{ and } V_b = s] \\ = \frac{1}{m^2} = \Pr[V_a = r] \cdot \Pr[V_b = s]\end{aligned}$$

pairwise independent  $\implies$  universal:

$$\begin{aligned}\Pr[V_a = V_b] &= \sum_s \Pr[V_a = s \text{ and } V_b = s] \\ &= \sum_s \frac{1}{m^2} = \frac{1}{m}\end{aligned}$$

(Homework) pairwise indep. families are easily constructed, but *must* have large key spaces

## Application: message authentication

Alice and Bob share a random key  $R$

Later, Alice sends a message  $a$  to Bob, together with a hash code  $C := h_R(a)$

An adversary can try to fool Bob, by sending him a different message with a correct hash code, that is, a message  $B$ , and a hash code  $D$ , such that  $B \neq a$  and  $h_R(B) = D$

Here,  $B$  and  $D$  are functions of  $C$

(Homework) Pairwise independent hashing implies

$$\Pr[B \neq a \text{ and } h_R(B) = D] \leq \frac{1}{m}$$