DS-563 / CS-543 2025-02-06

Logistics:

- HW1 (due yesterday, submit today for late deadline)
- HW2 (out soon, programming, will involve implementing Count-Min Sketch)
- final project ideas

Today: k-wise independent hash functions

Reminder: independent variables (discrete distributions)

$X_1, X_2, X_3, \ldots, X_n$ - random variables taking on values in $U_1, U_2, \ldots, U_n$, respectively

$X_1, X_2, \ldots, X_n$ are independent if for all $u_1 \in U_1, u_2 \in U_2, \ldots, u_n \in U_n$,

$$\Pr[X_1 = u_1 \wedge X_2 = u_2 \wedge \ldots \wedge X_n = u_n] = \prod_{i=1}^{n} \Pr[X_i = u_i]$$

$k \in \mathbb{Z} \wedge k \geq 2$

$X_1, X_2, \ldots, X_n$ are k-wise independent if every subset of at most k of the variables are independent

---

Example: 2-wise independence (aka. pairwise independence) but no 3-wise independence

random variables $X, Y, Z \in \{0,1\}$

relationship: $X \otimes Y = Z$ ($\otimes$ = xor (= exclusive or))

each row's probability = $\frac{1}{4}$

0	0	0
1	0	1
0	1	1
1	1	0

← Boolean table

Easy to verify:

- each of X, Y, Z is uniform on $\{0,1\}$
- X & Y independent
- X & Z independent
- Y & Z independent
- X, Y, Z not independent:

$$\Pr[X=Y=Z=0] = \frac{1}{4} \neq \frac{1}{8} = \Pr[X=0]\cdot\Pr[Y=0]\cdot\Pr[Z=0]$$

k-wise independent hash functions

$h: X \to H$ selected from some distribution on functions from X to H is k-wise independent if the set of its values $\{h(x) : x \in X\}$ is k-wise independent

Note: h is a random variable

H - set of hash values, e.g., $[m]$

random variable

(Note: the uniform distribution of each $h(x)$ is often a part of the definition, but here we will state this independently)

Construction of uniform and k-wise independent family of hash functions from $\{0, 1, \ldots, p-1\}$ to $\{0, 1, \ldots, p-1\}$ for prime p:

important

We use modular arithmetic of $\mathbb{F}_p$:

- $\mathbb{F}_p$ is $\{0, \ldots, p-1\}$ with operations + & × modulo p

Example: $p = 7$, $3 + 5 = 1$

$3 + 5 = 8 = 7 + 1$

$3 \times 5 = 1$

$3 \times 5 = 15 = 2 \cdot 7 + 1$

Popular notation that identifies integers i & j ($\in \mathbb{Z}$) such that p divides $(i-j)$:

$$i \equiv j \pmod{p}$$

In this notation:

$$3 \times 5 \equiv 15 \equiv 1 \pmod{7}$$

- $\mathbb{F}_p$ is a field, so in particular:

  Fact: For any $x \in \mathbb{F}_p \setminus \{0\}$, there is a unique element $y \in \mathbb{F}_p$ s.t.

  $$x \cdot y = 1$$

Proof: Consider any $x \in \mathbb{F}_p \setminus \{0\}$.
Let $M \overset{\text{def}}{=} \{x \cdot z : z \in \mathbb{F}_p\}$ be the set of all multiples of x.
We claim that $|M| = p$.
If this were not the case, there would be two multiples of x, $x \cdot z_1$ and $x \cdot z_2$, such that $x \cdot z_1 = x \cdot z_2$ and $z_1 \neq z_2$, where $z_1, z_2 \in \mathbb{F}_p$.
So we would have $x(z_1 - z_2) = 0$

but neither x nor $z_1 - z_2$ are divisible by p. So their product over $\mathbb{Z}$ is not divisible by p and $x(z_1 - z_2) \neq 0$ over $\mathbb{F}_p$. So it is impossible for two multiples of x to be identical and there is a unique y in $\mathbb{F}_p$ such that $x \cdot y = 1$ ∎

- Notation: For any $x \in \mathbb{F} \setminus \{0\}$, we write x^{-1} to denote the unique y such that $xy = 1$ over $\mathbb{F}_p$.
- Polynomials over $\mathbb{F}_p$: no surprises here
  - Example over $\mathbb{F}_7$:

$$g(x) = 3x^4 + 2x^2 + x + 4$$

degree of $g(x) = 4$

$$g(3) = 3 \cdot 3^4 + 2 \cdot 3^2 + 3 + 4$$
$$= 5 + 4 + 3 + 4$$
$$= 2$$

  - Degree of a polynomial: highest power of x

- Our construction: to get k-wise independence, draw a random polynomial of degree at most k-1 from the uniform distribution:

$$h(x) = \sum_{i=0}^{k-1} c_i x^i$$

each coefficient c_i selected independently at random from the uniform distribution on $\mathbb{F}_p$

- Observation 1: There are p^k such polynomials

- Observation 2: Each $h(x)$ distributed uniformly on $\mathbb{F}_p$.

Proof: Whatever $\sum_{i=1}^{k-1} c_i x^i$ is, adding c_0 selected independently and uniformly from $\mathbb{F}_p$ distributes $h(x) = c_0 + \sum_{i=1}^{k-1} c_i x^i$ uniformly on $\mathbb{F}_p$. ■

Example: Pairwise independence (i.e., $k=2$) over $\mathbb{F}_7$ ($p=7$)

$$h(x) = c_1 x + c_0$$

c_1, c_0: independently selected from the uniform distribution on $\mathbb{F}_7$

uniform distribution on $p^2 = 49$ possible polynomials

Each $h(x)$ is uniformly distributed on $\mathbb{F}_p$. To show that $h(x)$ & $h(y)$ are independent for some $x \neq y$, it suffices to show that each of the $p^2 = 49$ combinations of $h(x)$ & $h(y)$ is possible.

Consider two polynomials:

$$W_1(z) \stackrel{\text{def}}{=} z - x$$
$$W_2(z) \stackrel{\text{def}}{=} z - y$$

} Each of degree 1

Let $v_1 \stackrel{\text{def}}{=} W_1(y)$ & $v_2 \stackrel{\text{def}}{=} W_2(x)$.

Modify W_1 & W_2:

non-zero values

$$P_1(z) = (v_1^{-1}) W_1(z)$$
$$P_2(z) = (v_2^{-1}) W_2(z)$$

} Each of degree 1

We have

	$P_1(z)$	$P_2(z)$
$z=x$	0	1
$z=y$	1	0

Hence we can get each $h(x)=\alpha$ and $h(y)=\beta$ by setting

$$h(z) = \underbrace{\alpha P_1(z) + \beta P_2(z)}_{\text{degree at most one}}$$

General case: arbitrary k

Consider k different $x_1, \ldots, x_k \in \mathbb{Z}_p$

We have p^k different polynomials of degree at most $k-1$ from which we select uniformly at random.

Can we obtain all possible p^k configurations of values of h on x_i's?

YES!

This will show that $h(x_i)$'s are independent.

So why is it true?

Consider the following polynomials

$$W_i(z) \stackrel{\text{def}}{=} \prod_{j \in [k] \setminus \{i\}} (z - x_j)$$

← degree $k-1$

and values

$$V_i \stackrel{\text{def}}{=} W_i(x_i)$$

← non-zero value in $\mathbb{Z}_p$

so V_i^{-1} exists

Now normalize the outputs of W_i's on the corresponding x_i's

$$P_i(z) \stackrel{\text{def}}{=} (V_i^{-1}) \cdot W_i(z)$$

↑ still degree $k-1$ polynomial

For each P_i, we have

$$P_i(x_i) = 1$$

and for each $j \neq i$,

$$P_i(x_j) = 0$$

Hence we can get any configuration $h(x_i) = \alpha_i$ by setting

$$h(z) = \underbrace{\sum_{i=1}^{k} \alpha_i P_i(z)}_{\text{degree at most } k-1}$$

■

What about hashing to $[m]$ where m is not prime?

- Select prime $p \gg m$. Select hash function h over $\mathbb{F}_p$ as above.
- Use $h'(x) = (h(x) \bmod m) + 1$
- Still k-wise independent
- Slightly non-uniform, but if p is much bigger than m, impact limited.

Described in Homework 2:

How to get a reasonable pairwise independent hash function on strings without converting them to huge integers.