Lecture 5: Hashing

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

rafael.oliveira.teaching@gmail.com

September 21, 2021

Overview

- Introduction
  - Hash Functions
  - Why is hashing?
  - How to hash?
- Succinctness of Hash Functions
  - Coping with randomness
  - Universal Hashing
  - Hashing using 2-universal families
  - Perfect Hashing
- Acknowledgements

Computational Model

Before we talk about hash functions, we need to state our model of computation:

Definition (Word RAM model)

In the word RAM[a] model:

- all elements are integers that fit in a machine word of w bits
- Basic operations (comparison, arithmetic, bitwise) on such words take $\Theta(1)$ time
- We can also access *any* position in the array in $\Theta(1)$ time

[a] RAM stands for Random Access Model

Computational Model

Before we talk about hash functions, we need to state our model of computation:

Definition (Word RAM model)

In the word RAM[a] model:

- all elements are integers that fit in a machine word of w bits
- Basic operations (comparison, arithmetic, bitwise) on such words take $\Theta(1)$ time
- We can also access *any* position in the array in $\Theta(1)$ time

[a] RAM stands for Random Access Model

Wait, but aren't we working on *asymptotic analysis* of algorithms?
Yes, but this model is still relevant for problems of good enough size (so asymptotics can kick in) but not super huge that words don't fit in a machine word.

What is hashing?

Store $O(n)$ elements (keys) from the set $U = \{0, 1, \ldots, m-1\}$,[1] where $m >> n$, in a data structure that supports *insertions, deletions, search* "as efficiently as possible."

[1] Here we assume that m smaller than our memory. So $\log m \leq w$.

What is hashing?

Store $O(n)$ elements (keys) from the set $U = \{0, 1, \ldots, m-1\}$,[1] where $m >> n$, in a data structure that supports *insertions, deletions, search* "as efficiently as possible."

Naive approach: use an array A of m elements, initially $A[i] = 0$ for all i, and when a key is inserted, set $A[i] = 1$.

---

[1] Here we assume that m smaller than our memory. So $\log m \leqslant w$.

What is hashing?

Store $O(n)$ elements (keys) from the set $U = \{0, 1, \ldots, m-1\}$,[1] where $m >> n$, in a data structure that supports *insertions, deletions, search* "as efficiently as possible."

Naive approach: use an array A of m elements, initially $A[i] = 0$ for all i, and when a key is inserted, set $A[i] = 1$.

- Insertion: $O(1)$, Deletion: $O(1)$, Search: $O(1)$ wonderful
- Memory: $O(m)$ *(this is very bad!)*

[1] Here we assume that m smaller than our memory. So $\log m \leq w$.

What is hashing?

Store $O(n)$ elements (keys) from the set $U = \{0, 1, \ldots, m-1\}$,[1] where $m >> n$, in a data structure that supports *insertions, deletions, search* "as efficiently as possible."

Naive approach: use an array A of m elements, initially $A[i] = 0$ for all i, and when a key is inserted, set $A[i] = 1$.

- Insertion: $O(1)$, Deletion: $O(1)$, Search: $O(1)$
- Memory: $O(m)$ *(this is very bad!)*

Want to also achieve optimal memory $O(n)$. For this we will use a technique called *hashing*.

- A *hash function* is a function $h : U \to [0, n-1]$, where $|U| = m >> n$.
- A *hash table* is a data structure that consists of:
  - a table T with n cells $[0, n-1]$, each cell storing a word
  - a hash function $h : U \to [0, n-1]$

From now on, we will define memory as *# of cells*.

[1] Here we assume that m smaller than our memory. So $\log m \leq w$.

Why is hashing useful?

- Designing efficient data structures (dictionaries) for searching
- Data streaming algorithms
- Derandomization
- Cryptography
- Complexity Theory
- many more

Challenges in Hashing

Setup:

- Universe $U = \{0, \ldots, m-1\}$ of size $m >> n$ where n is the size of the range of our hash function $h : U \rightarrow [0, n-1]$
- Store $O(n)$ elements of U (keys) in hash table T (which has n cells)

Challenges in Hashing

Setup:

- Universe $U = \{0, \ldots, m-1\}$ of size $m >> n$ where n is the size of the range of our hash function $h : U \to [0, n-1]$
- Store $O(n)$ elements of U (keys) in hash table T (which has n cells)

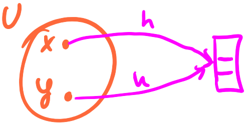
Ideally, want hash function to map *different keys* into *different locations*.

Definition (Collision)

We say that a *collision* happens for hash function h with inputs $x, y \in U$ if $x \neq y$ and $h(x) = h(y)$.

By pigeonhole principle, impossible to achieve no collisions without knowing keys in advance ($|U| >> n$).

collision:



Challenges in Hashing

Setup:

- Universe $U = \{0, \ldots, m-1\}$ of size $m >> n$ where n is the size of the range of our hash function $h : U \to [0, n-1]$
- Store $O(n)$ elements of U (keys) in hash table T (which has n cells)

Ideally, want hash function to map *different keys* into *different locations*.

Definition (Collision)

We say that a *collision* happens for hash function h with inputs $x, y \in U$ if $x \neq y$ and $h(x) = h(y)$.

By pigeonhole principle, impossible to achieve no collisions without knowing keys in advance ($|U| >> n$).

Will settle for: # collisions *small with high probability*.

Our solution: family of hash functions

Construct *family* of hash functions $\mathcal{H}$ such that the *number of collisions* is **small** with **high probability**, when we pick hash function uniformly at random from the family $\mathcal{H}$.

Our solution: family of hash functions

Construct *family* of hash functions $\mathcal{H}$ such that the *number of collisions* is **small** with **high probability**, when we pick hash function uniformly at random from the family $\mathcal{H}$.

Simplest version to keep in mind:

$$\Pr_{h \in_R \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{\text{poly}(n)} \qquad \forall x \neq y \in U$$

Prob. when picking h unif. at random from $\mathcal{H}$

they collide

any pair of keys

Our solution: family of hash functions

Construct *family* of hash functions $\mathcal{H}$ such that the *number of collisions* is **small** with **high probability**, when we pick hash function uniformly at random from the family $\mathcal{H}$.

Simplest version to keep in mind:

$$\Pr_{h \in_R \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{\text{poly}(n)} \qquad \forall x \neq y \in U$$

Assumptions:

- keys are independent from hash function we choose.
- we **do not** know keys in advance (even if we did, nontrivial problem!)

Question

Still could have collisions. How do we handle them?

Random Hash Functions?

Natural to consider following approach:

From all functions $h : U \to [0, n-1]$, just pick one uniformly at random.

$:= \{0, 1, 2, \ldots, n-1\}$

Random Hash Functions?

Natural to consider following approach:

From all functions $h : U \to [0, n-1]$, just pick one uniformly at random.

This setting is same as our balls-and-bins setting!

U set of balls

$[0, n-1]$ set of bins

Random Hash Functions?

Natural to consider following approach:

From all functions $h : U \to [0, n-1]$, just pick one uniformly at random.

This setting is same as our balls-and-bins setting!
So, if we have to store n keys:

- Expected number of keys in a location: 1
- maximum number of collisions (max load) in one particular location: $O(\log n / \log \log n)$ keys

typical

Random Hash Functions?

Natural to consider following approach:

From all functions $h : U \to [0, n-1]$, just pick one uniformly at random.

This setting is same as our balls-and-bins setting!
So, if we have to store n keys:

- Expected number of keys in a location: 1
- maximum number of collisions (max load) in one particular location: $O(\log n / \log \log n)$ keys

Solving collisions: store all keys hashed into location i by a linked list.

Known as *chain hashing*.

Insert $O(1)$ search / deletion $O(\log n / \log \log n)$ whp

Random Hash Functions?

Natural to consider following approach:

From all functions $h : U \to [0, n-1]$, just pick one uniformly at random.

This setting is same as our balls-and-bins setting!
So, if we have to store n keys:

- Expected number of keys in a location: 1
- maximum number of collisions (max load) in one particular location: $O(\log n / \log \log n)$ keys

Solving collisions: store all keys hashed into location i by a linked list.

Known as *chain hashing*.

Could also pick *two* random hash functions and use *power of two choices*. Collision bound becomes $O(\log \log n)$.

Random Hash Functions?

Random hash functions look very good. However, we haven't discussed the following:

Random Hash Functions?

Random hash functions look very good. However, we haven't discussed the following:

Question

How much resource (time & space) does it take to compute random hash functions?

Random Hash Functions?

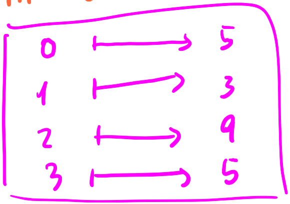
Random hash functions look very good. However, we haven't discussed the following:

Question

How much resource (time & space) does it take to compute random hash functions?

- Storing entire function $h : U \to [0, n-1]$ require $O(m)$ cells (way too much space!)
- Even if we only stored the elements we saw, would require $O(n)$ time to evaluate $h(x)$ (need to decide if we had already computed it!)

$m = 10000$ $\quad$ $n = 10$



Random Hash Functions?

Random hash functions look very good. However, we haven't discussed the following:

Question

How much resource (time & space) does it take to compute random hash functions?

- Storing entire function $h : U \to [0, n-1]$ require $O(m)$ cells (way too much space!)
- Even if we only stored the elements we saw, would require $O(n)$ time to evaluate $h(x)$ (need to decide if we had already computed it!)

Remark

Thus, for random function all operations (insert, delete, search) take $O(n)$ time (at best!)

Random Hash Functions?

Random hash functions look very good. However, we haven't discussed the following:

Question

How much resource (time & space) does it take to compute random hash functions?

- Storing entire function $h : U \to [0, n-1]$ require $O(m)$ cells (way too much space!)
- Even if we only stored the elements we saw, would require $O(n)$ time to evaluate $h(x)$ (need to decide if we had already computed it!)

Remark

Thus, for random function all operations (insert, delete, search) take $O(n)$ time (at best!)

How do we cope with the computational problem that arose with randomness?

- Introduction
  - Hash Functions
  - Why is hashing?
  - How to hash?
- Succinctness of Hash Functions
  - Coping with randomness
  - Universal Hashing
  - Hashing using 2-universal families
  - Perfect Hashing
- Acknowledgements

How to cope with "hardness" of randomness?

We want something that is *random-like* (few collisions w.h.p.) but *easy to compute/represent*.

[2] Reminder that we are in the word RAM model - in general we would have $O(\log m)$ bits

How to cope with "hardness" of randomness?

We want something that is *random-like* (few collisions w.h.p.) but *easy to compute/represent*.

Ideally something that takes $O(1)$ time to compute (as this is the size of our input).[2]

[2] Reminder that we are in the word RAM model - in general we would have $O(\log m)$ bits

How to cope with "hardness" of randomness?

We want something that is *random-like* (few collisions w.h.p.) but *easy to compute/represent*.

Ideally something that takes $O(1)$ time to compute (as this is the size of our input).[2]

$O(\log m)$ time (bit complexity)

Question

How many hash functions can we have with the property above?

function better have description ("code length")

$O(\log m)$ bits

[2] Reminder that we are in the word RAM model - in general we would have $O(\log m)$ bits

How to cope with “hardness” of randomness?

We want something that is *random-like* (few collisions w.h.p.) but *easy to compute/represent*.

Ideally something that takes $O(1)$ time to compute (as this is the size of our input).[2]

Question

How many hash functions can we have with the property above?

poly(m) functions, as each function takes at most $O(\log m)$ bits to describe. Thus these are *succinct functions* (easy to describe and compute) which have *random-like* properties!

$= 2^{O(\log m)} = 2^{c \cdot \log m} = m^c$

[2] Reminder that we are in the word RAM model - in general we would have $O(\log m)$ bits

How to cope with "hardness" of randomness?

We want something that is *random-like* (few collisions w.h.p.) but *easy to compute/represent*.

Ideally something that takes $O(1)$ time to compute (as this is the size of our input).[2]

Question

How many hash functions can we have with the property above?

poly(m) functions, as each function takes at most $O(\log m)$ bits to describe. Thus these are *succinct functions* (easy to describe and compute) which have *random-like* properties!

Part of *derandomization/pseudorandomness*: huge subfield in TCS!

[2] Reminder that we are in the word RAM model - in general we would have $O(\log m)$ bits

k-wise independence

Weaker notion of independence.

k-wise independence

Weaker notion of independence.

Definition (Full Independence)

A set of random variables $X_1, \ldots, X_n$ are said to be (fully) independent if for any subset $J \subseteq [n]$ they satisfy

$$\Pr\left[\bigcap_{i \in J} X_i = a_i\right] = \prod_{i \in J} \Pr[X_i = a_i]$$

any values a_i $i \in J$

k-wise independence

Weaker notion of independence.

Definition (Full Independence)

A set of random variables $X_1, \ldots, X_n$ are said to be (fully) independent if for any subset $J \subseteq [n]$ they satisfy

$$\Pr\left[\bigcap_{i \in J} X_i = a_i\right] = \prod_{i \in J} \Pr[X_i = a_i]$$

Definition (k-wise Independence)

A set of random variables $X_1, \ldots, X_n$ are said to be k-wise independent if for any set $J \subseteq [n]$ such that $|J| \leq k$ they satisfy

small

$$\Pr\left[\bigcap_{i \in J} X_i = a_i\right] = \prod_{i \in J} \Pr[X_i = a_i]$$

Pairwise independence

When $k = 2$, k-wise independence is called *pairwise independence*.

independent

Example (XOR pairwise independence)

Given b uniformly random bits $Y_1, \ldots, Y_b$, we can generate $2^b - 1$ uniformly distributed pairwise independent random variables as follows:

$$X_S := \bigoplus_{i \in S} Y_i \qquad S \subseteq [b], \quad S \neq \emptyset$$

$X_{\{1,2,3\}} = Y_1 \oplus Y_2 \oplus Y_3$

$X_{\{1,3\}} = Y_1 \oplus Y_3$

Pairwise independence

When $k = 2$, k-wise independence is called *pairwise independence*.

Example (XOR pairwise independence)

Given b uniformly random bits $Y_1, \ldots, Y_b$, we can generate $2^b - 1$ uniformly distributed pairwise independent random variables as follows:

$$X_S := \bigoplus_{i \in S} Y_i \qquad S \subseteq [b], \quad S \neq \emptyset$$

Pairwise independence

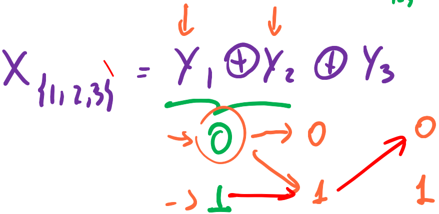
When $k = 2$, k-wise independence is called *pairwise independence*.

Example (XOR pairwise independence)

Given b uniformly random bits $Y_1, \ldots, Y_b$, we can generate $2^b - 1$ uniformly distributed pairwise independent random variables as follows:

$$X_S := \bigoplus_{i \in S} Y_i \qquad S \subseteq [b], \quad S \neq \emptyset$$

- Why are they even random?



$X_{123} = 0 \leftarrow \Pr_{Y_1,Y_2,Y_3}[Y_3 = Y_1 \oplus Y_2] = 1/2$

$\Pr[Y_3 \neq Y_1 \oplus Y_2] = 1/2 \rightarrow X_{123} = 1$

Pairwise independence

When $k = 2$, k-wise independence is called *pairwise independence*.

Example (XOR pairwise independence)

Given b uniformly random bits $Y_1, \ldots, Y_b$, we can generate $2^b - 1$ uniformly distributed pairwise independent random variables as follows:

$$X_S := \bigoplus_{i \in S} Y_i \qquad S \subseteq [b], \quad S \neq \emptyset$$

- Why are they even random?
- Why are they pairwise independent?

$S \neq T \quad \exists\, i \in S \setminus T$

Y_i independent of $Y_j \quad j \in T$

X_S X_T

$$\Pr[X_S = b_1 \text{ and } X_T = b_2] = \Pr[X_T = b_2] \cdot \Pr[X_S = b_1 \mid X_T = b_2] \to \Pr[X_S = b_1]$$

$Y_i \qquad Y_j = a_j$

Pairwise independence

When $k = 2$, k-wise independence is called *pairwise independence*.

Example (XOR pairwise independence)

Given b uniformly random bits $Y_1, \ldots, Y_b$, we can generate $2^b - 1$ uniformly distributed pairwise independent random variables as follows:

$$X_S := \bigoplus_{i \in S} Y_i \qquad S \subseteq [b], \quad S \neq \emptyset$$

- Why are they even random?
- Why are they pairwise independent?
- Are they also 3-wise independent?

NO : $X_{\{1\}}$ $X_{\{2\}}$ $X_{\{1,2\}}$

1 1 0

Pairwise independence II

Example (Pairwise independence in $\mathbb{F}_p$)

set integers mod p

Let p be a prime number. Given 2 uniformly random variables $Y_1, Y_2 \sim [0, \ldots, p-1]$, generate p pairwise independent random variables as follows:

$\mathbb{F}_p$

$$X_i := Y_1 + i \cdot Y_2 \mod p \qquad i \in [0, p-1]$$

Pairwise independence II

Example (Pairwise independence in $\mathbb{F}_p$)

Let p be a prime number. Given 2 uniformly random variables $Y_1, Y_2 \sim [0, \ldots, p-1]$, generate p pairwise independent random variables as follows:

$$X_i := Y_1 + i \cdot Y_2 \mod p \qquad i \in [0, p-1]$$

- Why are they even random?

fixing Y_2 Y_1 still uniform

Pairwise independence II

Example (Pairwise independence in $\mathbb{F}_p$)

Let p be a prime number. Given 2 uniformly random variables $Y_1, Y_2 \sim [0, \ldots, p-1]$, generate p pairwise independent random variables as follows:

$$X_i := Y_1 + i \cdot Y_2 \mod p \qquad i \in [0, p-1]$$

- Why are they even random? $\Pr[X_i = a] = 1/p$
- Why are they pairwise independent?

$X_i = a$, $X_j = b$

$$\left.\begin{aligned} Y_1 + i\,Y_2 &= a \\ Y_1 + j\,Y_2 &= b \end{aligned}\right\} \text{ exactly 1 solution } (Y_1, Y_2)$$

$$\Pr[X_i = a, X_j = b] = \frac{1}{p^2} = \Pr[X_i = a] \cdot \Pr[X_j = b]$$

Pairwise independence II

Example (Pairwise independence in $\mathbb{F}_p$)

Let p be a prime number. Given 2 uniformly random variables $Y_1, Y_2 \sim [0, \ldots, p-1]$, generate p pairwise independent random variables as follows:

$$X_i := Y_1 + i \cdot Y_2 \mod p \qquad i \in [0, p-1]$$

- Why are they even random?
- Why are they pairwise independent?
- Are they also 3-wise independent?

No : X_0 X_1 X_2

Pairwise independence II

Example (Pairwise independence in $\mathbb{F}_p$)

Let p be a prime number. Given 2 uniformly random variables $Y_1, Y_2 \sim [0, \ldots, p-1]$, generate p pairwise independent random variables as follows:

$$X_i := Y_1 + i \cdot Y_2 \mod p \qquad i \in [0, p-1]$$

- Why are they even random?
- Why are they pairwise independent?
- Are they also 3-wise independent?

Can think of these random variables as picking a random line over a finite field. If we only know one point of the line, the second point is still uniformly random. However two points determine the line.

Universal Hash Functions

We want hash functions. Why are we talking about random variables?

Universal Hash Functions

We want hash functions. Why are we talking about random variables?

Definition (Universal Hash Functions)

Let U be a universe with $|U| \geq n$. A family of hash functions $\mathcal{H} = \{h : U \to [0, n-1]\}$ is *k-universal* if, *for any* **distinct** elements $u_1, \ldots, u_k \in U$, we have

$$\Pr_{h \in_R \mathcal{H}}[h(u_1) = h(u_2) = \ldots = h(u_k)] \leq 1/n^{k-1}$$

Universal Hash Functions

We want hash functions. Why are we talking about random variables?

Definition (Universal Hash Functions)

Let U be a universe with $|U| \geq n$. A family of hash functions $\mathcal{H} = \{h : U \to [0, n-1]\}$ is *k-universal* if, *for any* **distinct** elements $u_1, \ldots, u_k \in U$, we have

$$\Pr_{h \in_R \mathcal{H}} [h(u_1) = h(u_2) = \ldots = h(u_k)] \leq 1/n^{k-1}$$

Definition (Strongly Universal Hash Functions)

$\mathcal{H} = \{h : U \to [0, n-1]\}$ is *strongly k-universal* if, *for any* **distinct** elements $u_1, \ldots, u_k \in U$ and *for any* values $y_1, \ldots, y_k \in [0, n-1]$, we have

$$\Pr_{h \in_R \mathcal{H}} [h(u_1) = y_1, \ldots, h(u_k) = y_k] = 1/n^k$$

Relation to k-wise independent random variables

What do the previous definitions have to do with random variables?

Relation to k-wise independent random variables

What do the previous definitions have to do with random variables?

Family $\mathcal{H}$ is *strongly k-universal* if the random variables $h(0), \ldots, h(|U| - 1)$ are *k-wise independent*.

Relation to k-wise independent random variables

What do the previous definitions have to do with random variables?

Family $\mathcal{H}$ is *strongly k-universal* if the random variables $h(0), \ldots, h(|U| - 1)$ are *k-wise independent*.

Can use random variables to construct universal hash functions!

Strongly 2-universal families of hash functions

Let p be a prime number, $U = [0, p-1]$.

Proposition

$$\mathcal{H} = \{h_{a,b}(x) := a \cdot x + b \mod p \quad | \quad a, b \in [0, p-1]\}$$

is strongly 2-universal.

$h : U \to U$

Strongly 2-universal families of hash functions

Let p be a prime number, $U = [0, p-1]$.

Proposition

$$\mathcal{H} = \{h_{a,b}(x) := a \cdot x + b \mod p \quad | \quad a, b \in [0, p-1]\}$$

is strongly 2-universal.

How do we make the domain U much larger than image of the maps? (as usually in hashing size of universe much larger than size of table)

Strongly 2-universal families of hash functions

Let p be a prime number, $U = [0, p-1]$.

Proposition

$$\mathcal{H} = \{h_{a,b}(x) := a \cdot x + b \mod p \mid a, b \in [0, p-1]\}$$

is strongly 2-universal.

How do we make the domain U much larger than image of the maps? (as usually in hashing size of universe much larger than size of table)

$\mathbb{F}_{p^k} \equiv \mathbb{F}_p^k$ $n = p$ $[0, p-1]$

Proposition

Let $U = [0, p^k - 1] \equiv [0, p-1]^k$ *and* $\vec{a} = (a_0, \ldots a_{k-1})$

$$\mathcal{H} = \{h_{\vec{a},b}(\vec{x}) := \vec{a} \cdot \vec{x} + b \mod p \mid \vec{a} \in U, b \in [0, p-1]\}$$

is strongly 2-universal.

Strongly 2-universal families of hash functions

Proposition

Let $U = [0, p^k - 1] \equiv [0, p-1]^k$ *and* $\vec{a} = (a_0, \ldots a_{k-1})$

$$\mathcal{H} = \{h_{\vec{a},b}(\vec{x}) := \vec{a} \cdot \vec{x} + b \mod p \quad | \quad \vec{a} \in U, b \in [0, p-1]\}$$

is strongly 2-universal.

2-universal families of hash functions

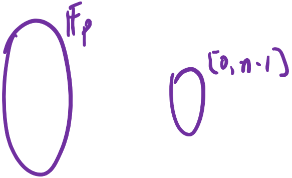
What if my hash table size is not a prime?

Proposition

$$\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \mid a, b \in [0, p-1], a \neq 0\}$$

is 2-universal (but not strongly 2-universal).

Practice problem: prove the proposition above.



k-universal families of hash functions

Can we construct k-universal families of hash functions like this?

k-universal families of hash functions

Can we construct k-universal families of hash functions like this?

- YES! Instead of constructing random lines (degree 1 polynomials), can construct random univariate polynomials of degree $k - 1$

$$y_1 + t y_2 + t^2 y_3$$

3-wise independent

k-universal families of hash functions

Can we construct k-universal families of hash functions like this?

- YES! Instead of constructing random lines (degree 1 polynomials), can construct random univariate polynomials of degree $k - 1$
- Two points determine a line. Similarly, k points determine a univariate polynomial of degree $k - 1$

k-universal families of hash functions

Can we construct k-universal families of hash functions like this?

- YES! Instead of constructing random lines (degree 1 polynomials), can construct random univariate polynomials of degree $k-1$
- Two points determine a line. Similarly, k points determine a univariate polynomial of degree $k-1$
- Random degree $k-1$ polynomials are k-wise independent!
- Practice problem: prove this!

Efficiency

How did pairwise independent improve the problems we were having with random functions?

Efficiency

How did pairwise independent improve the problems we were having with random functions?

Remark

For random function all operations (insert, delete, search) take $O(n)$ time (at best!)

Efficiency

How did pairwise independent improve the problems we were having with random functions?

Remark

For random function all operations (insert, delete, search) take $O(n)$ time (at best!)

Remark

- In XOR example, our function takes $O(b)$ storage space, and $O(b)$ time to compute.[a]
- In $\mathbb{F}_p$ examples, our function takes $O(1)$ storage space and $O(1)$ time to compute![b]

[a] Reminder that we assume that $b < w$.
[b] We assume that $p < 2^w$.

$O(\log n)$ input $S \subset [b]$ $n = 2^b - 1$

$h_{a,b}(x) = ax + b \mod p$

store a, b

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$

$a \neq 0$

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$
- Only need to choose $a, b \in [0, p-1]$ to store a function from $\mathcal{H}$.

$a \neq 0$

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$
- Only need to choose $a, b \in [0, p-1]$ to store a function from $\mathcal{H}$.
- Computation time of $h_{a,b}$ is also $O(1)$

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$
- Only need to choose $a, b \in [0, p-1]$ to store a function from $\mathcal{H}$.
- Computation time of $h_{a,b}$ is also $O(1)$
- Can this hash function match chain hashing parameters? ($O(\log \log n)$ ~~search time~~)
  max load

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$
- Only need to choose $a, b \in [0, p-1]$ to store a function from $\mathcal{H}$.
- Computation time of $h_{a,b}$ is also $O(1)$
- Can this hash function match chain hashing parameters? ($O(\log \log n)$ search time)

Do not have same expected search time as chain hashing.

Hashing with 2-universal families

- Let $U = [0, m-1]$, and p be a prime number such that $m \leq p < 2m$ (exists by Bertrand's postulate)
- $\mathcal{H} = \{h_{a,b}(x) := (a \cdot x + b \mod p) \mod n \quad | \quad a, b \in [0, p-1]\}$
- Only need to choose $a, b \in [0, p-1]$ to store a function from $\mathcal{H}$.
- Computation time of $h_{a,b}$ is also $O(1)$
- Can this hash function match chain hashing parameters? ($O(\log \log n)$ search time)

Do not have same expected search time as chain hashing.

Lemma (Maximum number of collisions)

The expected number of collisions when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \ell^2/2n$$

Hashing with 2-universal families

Lemma (Maximum number of collision)

The expected number of collisions when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \ell^2/2n$$

$$X_{ij} = \begin{cases} 1 & \text{if keys } i, j \text{ map to same cell} \\ 0 & \text{otherwise} \end{cases} \qquad \Pr[X_{ij}=1] \leq \frac{1}{n}$$

$$X = \sum_{i<j} X_{ij} \quad \# \text{ collisions}$$

$$\mathbb{E}[X] = \sum_{i<j} \mathbb{E}[X_{ij}] \leq \sum_{i<j} \frac{1}{n} \leq \frac{\ell^2}{2n}$$

By Markov $\Pr\left[X > \ell^2/n\right] < \frac{1}{2}$

many collisions

Hashing with 2-universal families

Lemma (Maximum number of collision)

The expected number of collisions when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \ell^2/2n$$

Thus, by Markov's inequality, we have

Lemma (Maximum load of entry of hash table)

With probability $\geq 1/2$ the maximum load when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \sqrt{\frac{2\ell^2}{n}}.$$

When $\ell \approx n$ (as is usually assumed in hashing), we expect $\sqrt{2n}$.

Number of collisions

Lemma (Maximum load of entry of hash table)

With probability $\geq 1/2$ the maximum load when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \sqrt{\frac{2\ell^2}{n}}.$$

When $\ell \approx n$ (as is usually assumed in hashing), we expect $\sqrt{2n}$.

Number of collisions

Lemma (Maximum load of entry of hash table)

With probability $\geq 1/2$ the maximum load when inserting ℓ elements in a table of size n using a 2-universal hash family is

$$\leq \sqrt{\frac{2\ell^2}{n}}.$$

When $\ell \approx n$ (as is usually assumed in hashing), we expect $\sqrt{2n}$.

- Let C be the ~~number of collisions in~~ maximum load a cell

$$\binom{C}{2} \leq X \Rightarrow \Pr\left[C \geq \sqrt{\frac{2\ell^2}{n}}\right] \leq 1/2$$

$\binom{C}{2}$: # collisions coming from that cell

X: total # collisions

$\Pr\left[\frac{C^2}{2} > \frac{\ell^2}{n}\right] < \frac{1}{2}$ ($C > \sqrt{\frac{2\ell^2}{n}}$)

$\Pr\left[\frac{C^2}{2} > t\right] < \Pr[X > t]$

Markov: $\Pr\left[X > \frac{\ell^2}{n}\right] < 1/2$

Perfect Hashing

Setup: the set of keys is *static* (i.e., we know them in advance).
How to build a hash table with $O(1)$ search time and $O(n)$ memory? Can we still do it with a 2-universal family of hash functions?

Perfect Hashing

Setup: the set of keys is *static* (i.e., we know them in advance).
How to build a hash table with $O(1)$ search time and $O(n)$ memory? Can we still do it with a 2-universal family of hash functions?

Corollary

If $h \in \mathcal{H}$ is a random hash function from a 2-universal family of hash functions, then for any set $S \subseteq U$ of size $\ell \leq \sqrt{n}$, the probability of h being perfect for S is at least $1/2$.

Proof: There is no collision with probability $\geq 1/2$.

$$\Pr\left[\text{max load} \geq \sqrt{\frac{2\ell^2}{n}}\right] \leq 1/2$$

$$\ell \leq \sqrt{n} \Rightarrow \sqrt{\frac{2\ell^2}{n}} \leq \sqrt{2} < 2$$

Perfect Hashing

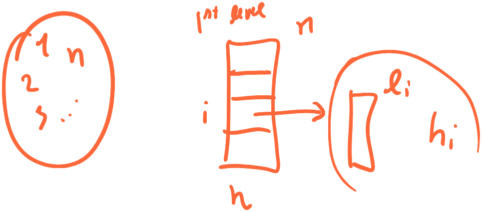
Setup: the set of keys is *static* (i.e., we know them in advance).
How to build a hash table with $O(1)$ search time and $O(n)$ memory? Can we still do it with a 2-universal family of hash functions?

Corollary

If $h \in \mathcal{H}$ is a random hash function from a 2-universal family of hash functions, then for any set $S \subseteq U$ of size $\ell \leq \sqrt{n}$, the probability of h being perfect for S is at least $1/2$.

Proof: There is no collision with probability $\geq 1/2$.

New idea: build a *two-level* hash table!



Perfect Hashing

Setup: the set of keys is *static* (i.e., we know them in advance).
How to build a hash table with $O(1)$ search time and $O(n)$ memory? Can we still do it with a 2-universal family of hash functions?

Corollary

If $h \in \mathcal{H}$ is a random hash function from a 2-universal family of hash functions, then for any set $S \subseteq U$ of size $\ell \leq \sqrt{n}$, the probability of h being perfect for S is at least $1/2$.

Proof: There is no collision with probability $\geq 1/2$.

New idea: build a *two-level* hash table!

Theorem

The two-level approach gives perfect hashing scheme.

have $O(n)$ hash tables memory $O(n)$

Proof of Theorem

S set of keys

Approach: pick first layer hash function h uniformly at random. Test h on S.

With probability $\geq \frac{1}{2}$ max load in a cell is $\leq \sqrt{n}$. We know that get good h in constant # tries.

Assume that max load (h, S) is $\leq \sqrt{n}$

$l_i \leftarrow$ load at i^{th} cell from (h, S)

Proof of Theorem

know: $l_i \leq \sqrt{n}$ (because h is good)

also know $\sum_{i=1}^{n} l_i = n = |S|$

Corollary if take h_i random from 2-universal family $h_i : S \to l_i^2$ h_i is perfect for the l_i elements mapping to ith cell. (w.h.p.)

expected # collisions $= 0$ this scheme is perfect

Bound on memory: $n + \sum_{i=0}^{n-1} l_i^2$

Proof of Theorem

$$\sum_{i=0}^{n-1} l_i^2 \leq 2(\#\text{ collisions in hash function } h)$$

collisions for 1^{st} hash function is w.h.p.

$$\leq \frac{n^2}{2n} = \frac{n}{2}$$

$$\sum_{i=0}^{n-1} l_i^2 = O(n) \implies \text{memory is } O(n)$$

Proof of Theorem

Acknowledgement

- Lecture based largely on Lap Chi's notes.
- See Lap Chi's notes at
  https://cs.uwaterloo.ca/~lapchi/cs466/notes/L05.pdf

References I

Motwani, Rajeev and Raghavan, Prabhakar (2007)
Randomized Algorithms

Mitzenmacher, Michael, and Eli Upfal (2017)
Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.
Cambridge university press, 2017.