Hash Functions

Data Structures and Algorithms (094224)

Yuval Emek

Winter 2022/23

Stronger assumptions on the keys

- Interested in data structures supporting dictionary operations: search, insert, delete
- All efficient data structures we've seen so far are comparison based
  - No assumptions on the keys other than total order
- Stronger assumptions may yield more efficient data structures
- Example: n objects whose keys are integers in $[n]$
  - Store directly in an array of size n
  - Dictionary operations implemented in $O(1)$ time
- What if #possible keys $\gg n$?

Hashing

Hash table ⟨טבלת גיבוב⟩

- Assumption: keys belong to some arbitrarily large universe U
- Backbone of hash table: array $T[0, 1, \ldots, m-1]$
  - Typically $m \ll |U|$
- Assignment of objects to array entries is determined by *hash function* ⟨פונקציית גיבוב⟩

$$h : U \to \{0, 1, \ldots, m-1\}$$

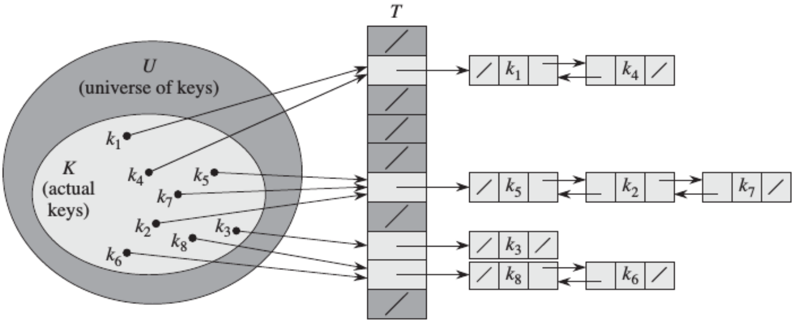
  - Object with key $k \in U$ is assigned to entry $h(k)$ of T
- Requirement: $h(k)$ can be evaluated in $O(1)$ time for any key $k \in U$

Collision resolution

- When $|U| > m$, there must exist $k \neq k' \in U$ such that $h(k) = h(k')$
  - Why?
  - Referred to as collision
- How do we store multiple objects whose keys collide?
- Need collision resolution scheme
  - In this lecture: chaining
  - Other schemes: open addressing, cuckoo hashing, and more

Resolve collisions by chaining

- Each entry in T is a (doubly) linked list
- Insert new object x to head of list $T[h(x.key)]$
  - Run-time: $O(1)$
- Delete object x by deleting it from its list
  - Run-time: $O(1)$
- Search key k by searching list $T[h(k)]$
  - Run-time: $O(\text{length of list } T[h(k)])$



Aim for short lists

- How long can lists be?
- Worst-case: $\Omega(n)$
  - As bad as storing all objects in one linked list
- Advantage(s) of hash tables:
  - Good behavior on average
  - Good behavior on expectation
- Good behavior: lists not much longer than load factor $\alpha = n/m$
  - As good as it gets
- Example:
  - Keys are 9-digit Israeli IDs, $m = 100$
  - Hash according to 2 least significant digits
  - Expected list length $\approx \alpha$ if objects (and keys) are picked randomly
    - Good behavior for average instance
  - Can we hash according to 2 most significant digits?
- What if cannot assume average instance?
  - Keys chosen by adversary aiming for the worst case

The perks of using randomness

- Conventions:
  - $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$
  - Universe $U = \mathbb{Z}_u$ for large integer u
  - Identify T with set of keys stored in table T
    - $|T| = n$
  - Identify $T[i]$ with set of keys stored in list $T[i]$
- Keys chosen by adversary
  - Wishes to maximize $|T[h(k)]|$ when searching for key k
  - Knows algorithm, but oblivious to its random coin tosses
    - A.k.a. oblivious adversary
- If $h : \mathbb{Z}_u \rightarrow \mathbb{Z}_m$ is random, then $\mathbb{E}(|T[i]|) = \alpha$ for any $i \in \mathbb{Z}_m$
  - Optimal
- Can we use "totally" random function h as hash function?
  - Requires huge space to represent h
    - Better off storing objects directly in array of size u
- Aim for class $\mathcal{H}$ of functions such that
  - $h \in \mathcal{H}$ can be represented succinctly
  - $h \in \mathcal{H}$ can be evaluated in $O(1)$ time
  - Functions in $\mathcal{H}$ appear random enough to fool adversary

1. Pairwise independence

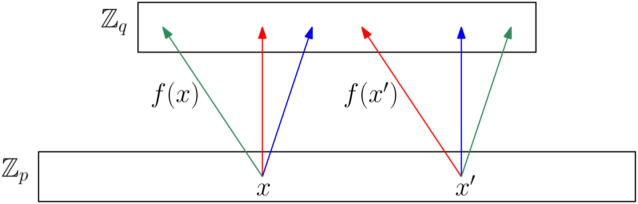
2. Constructing a pairwise independent family of hash functions
   - Pairwise independent family of functions $\mathbb{Z}_p \to \mathbb{Z}_p$
   - Smaller range

Pairwise independent function family

- Family $\mathcal{F}$ of functions $\mathbb{Z}_p \to \mathbb{Z}_q$ is *pairwise independent* if

$$\mathbb{P}_{f \in_R \mathcal{F}} \left(f(x) = y \wedge f(x') = y' \right) = \frac{1}{q^2}$$

  for any $x, x' \in \mathbb{Z}_p$, $x \neq x'$, and $y, y' \in \mathbb{Z}_q$
  - $f \in_R \mathcal{F}$ means that f is picked uniformly at random (u.a.r.) from $\mathcal{F}$
- When f is picked u.a.r. from $\mathcal{F}$, $f(x)$ and $f(x')$ are random variables



Pairwise independent function family — cont.

- Consider pairwise independent family $\mathcal{F}$ of functions $\mathbb{Z}_p \to \mathbb{Z}_q$
- Let $f \in_R \mathcal{F}$
- Observation: random variable $f(x)$ is uniform for every $x \in \mathbb{Z}_p$
  - Show that $\mathbb{P}(f(x) = y) = \frac{1}{q}$ for any $y \in \mathbb{Z}_q$
  - Fix some $x' \in \mathbb{Z}_p$, $x' \neq x$
  - $\mathbb{P}(f(x) = y) = \sum_{y' \in \mathbb{Z}_q} \mathbb{P}(f(x) = y \wedge f(x') = y') = q \cdot \frac{1}{q^2} = \frac{1}{q}$ ■
- Observation: random variables $f(x)$ and $f(x')$ are independent for every $x, x' \in \mathbb{Z}_p$, $x \neq x'$
  - $\mathbb{P}(f(x) = y \wedge f(x') = y') = \frac{1}{q^2} = \mathbb{P}(f(x) = y) \cdot \mathbb{P}(f(x') = y')$ ■
- Pairwise independence does not imply mutual ("total") independence
  - $f(x)$, $f(x')$, and $f(x'')$ are not necessarily independent
  - $\mathbb{P}(f(x) = y \wedge f(x') = y' \wedge f(x'') = y'')$ is not necessarily $\frac{1}{q^3}$

Approximate pairwise independence

- Given parameter $\delta \geq 1$, family $\mathcal{F}$ of functions $\mathbb{Z}_p \to \mathbb{Z}_q$ is *δ-approximately pairwise independent* if

$$\frac{1}{\delta} \cdot \frac{1}{q^2} \leq \mathbb{P}_{f \in_R \mathcal{F}} \left(f(x) = y \wedge f(x') = y' \right) \leq \delta \cdot \frac{1}{q^2}$$

  for any $x, x' \in \mathbb{Z}_p$, $x \neq x'$, and $y, y' \in \mathbb{Z}_q$
- Pairwise independence = 1-approx. pairwise independence

Pairwise independent hash functions

Theorem

Let $\mathcal{H}$ be a δ-approx. pairwise independent family of hash functions $\mathbb{Z}_u \to \mathbb{Z}_m$ and let T be a hash table whose hash function h is picked u.a.r. from $\mathcal{H}$. Following any sequence of insert/delete operations, we have

$$\mathbb{E}\left(|T[h(k)]|\right) \leq \begin{cases} \delta\alpha\,, & \text{if } k \notin T \\ 1 + \delta\alpha & \text{if } k \in T \end{cases}$$

for any $k \in \mathbb{Z}_u$.

- For $\ell \in \mathbb{Z}_u - \{k\}$, define random variable

$$X_\ell = 1\{h(k) = h(\ell)\}$$

- Since $\mathcal{H}$ is δ-approx. pairwise independent, for every $\ell \in \mathbb{Z}_u - \{k\}$,

$$\mathbb{P}\left(X_\ell = 1\right) = \sum_{i \in \mathbb{Z}_m} \mathbb{P}\left(h(k) = i \wedge h(\ell) = i\right) \leq m \cdot \frac{\delta}{m^2} = \frac{\delta}{m}$$

The proof continues

- Define random variable
$$Y = \sum_{\ell \in T - \{k\}} X_\ell$$
- Develop
$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}\left(\sum_{\ell \in T - \{k\}} X_\ell\right) = \sum_{\ell \in T - \{k\}} \mathbb{E}(X_\ell) \\ &\leq \sum_{\ell \in T - \{k\}} \frac{\delta}{m} = |T - \{k\}| \cdot \frac{\delta}{m} \end{aligned}$$
- If $k \notin T$, then $|T[h(k)]| = Y$ and $|T - \{k\}| = n$, hence
$$\mathbb{E}(|T[h(k)]|) = \mathbb{E}(Y) \leq n \cdot \frac{\delta}{m} = \delta\alpha$$
- If $k \in T$, then $|T[h(k)]| = 1 + Y$ and $|T - \{k\}| = n - 1$, hence
$$\mathbb{E}(|T[h(k)]|) = 1 + \mathbb{E}(Y) \leq 1 + (n-1) \cdot \frac{\delta}{m} \leq 1 + \delta\alpha \;\blacksquare$$

The quest for a small pairwise independent family

- By definition, family of all functions $\mathbb{Z}_u \to \mathbb{Z}_m$ is pairwise independent
  - Does not help
- Is there much smaller (approx.) pairwise independent family $\mathcal{H}$?
  - Functions in $\mathcal{H}$ can be represented succinctly
  - Functions in $\mathcal{H}$ can be evaluated in $O(1)$ time
- CLRS: related notion of universal hashing

1. Pairwise independence

2. Constructing a pairwise independent family of hash functions
   - Pairwise independent family of functions $\mathbb{Z}_p \to \mathbb{Z}_p$
   - Smaller range

The plan

- Goal: Construct δ-approx. pairwise independent family $\mathcal{H}$ of hash functions $\mathbb{Z}_u \to \mathbb{Z}_m$
  - Functions $h \in \mathcal{H}$ can be represented succinctly
  - Functions $h \in \mathcal{H}$ can be evaluated in $O(1)$ time
- Construction works in two stages
  1. Construct pairwise independent family $\mathcal{F}$ of functions $\mathbb{Z}_p \to \mathbb{Z}_p$
     - Parameter $p \geq u$ ($\gg m$) can be made arbitrarily large
  2. Generate $\mathcal{H}$ from $\mathcal{F}$ while introducing approx. error $\delta \to 1$ as $\frac{p}{m} \to \infty$

1. Pairwise independence

2. Constructing a pairwise independent family of hash functions
   - Pairwise independent family of functions $\mathbb{Z}_p \to \mathbb{Z}_p$
   - Smaller range

A reminder from Algebra 101

- A field ⟨שדה⟩ is defined over set F with two operations:
  - Addition $+ : F \times F \to F$
  - Multiplication $\cdot : F \times F \to F$
- Satisfies the following axioms for every $a, b, c \in F$:
  - Associativity: $a + (b + c) = (a + b) + c$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$
  - Commutativity: $a + b = b + a$ and $a \cdot b = b \cdot a$
  - Identity: there exist designated $0, 1 \in F$ s.t. $a + 0 = a$ and $a \cdot 1 = a$
  - Additive inverse: there exists $-a \in F$ s.t. $a + (-a) = 0$
  - Multiplicative inverse: $a \neq 0 \implies$ there exists $a^{-1} \in F$ s.t. $a \cdot a^{-1} = 1$
  - Distributivity: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$
- Examples: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$
  - Not fields: $\mathbb{N}$, $\mathbb{Z}$

Theorem

$\mathbb{Z}_p$ with addition and multiplication modulo p is a field for every prime p.

Linear functions

- Take p to be sufficiently large prime
- Given parameters $a, b \in \mathbb{Z}_p$, define $f_{a,b} : \mathbb{Z}_p \to \mathbb{Z}_p$ by setting

$$f_{a,b}(x) = ax + b \bmod p$$

for every $x \in \mathbb{Z}_p$

  - Usual conventions: omit '$\cdot$', operator precedence
  - Linear function over finite field $\mathbb{Z}_p$

- Define function family

$$\mathcal{F} = \{f_{a,b} \mid a, b \in \mathbb{Z}_p\}$$

  - Representation of $f_{a,b} \in \mathcal{F}$ requires only a and b
  - Evaluation of $f_{a,b} \in \mathcal{F}$ is done in $O(1)$ time

Theorem

$\mathcal{F}$ is pairwise independent.

Proving the theorem

- Throughout this proof, all arithmetic is modulo p
- Consider some $x, x' \in \mathbb{Z}_p$, $x \neq x'$
- For $a, b \in \mathbb{Z}_p$ and $y, y' \in \mathbb{Z}_p$, we have

$$f_{a,b}(x) = y \iff ax + b = y \quad \text{and} \quad f_{a,b}(x') = y' \iff ax' + b = y'$$

- Subtract one equation from the other:

$$a(x - x') \;=\; ax - ax' \;=\; y - y'$$

- $x \neq x'$, thus $x - x' \neq 0$ and there exists $(x - x')^{-1} \in \mathbb{Z}_p$, so

$$\begin{aligned} a &= (y - y')(x - x')^{-1} \\ b &= y - (y - y')(x - x')^{-1} x \end{aligned}$$

- Mapping from $(a, b) \in \mathbb{Z}_p^2$ to $(y, y') \in \mathbb{Z}_p^2$ is invertible $\Longrightarrow$ bijection
- (a, b) chosen u.a.r. from $\mathbb{Z}_p^2$ when picking $f \in_R \mathcal{F}$, hence

$$\mathbb{P}_{f \in_R \mathcal{F}} \left(f(x) = y \wedge f(x') = y' \right) \;=\; \frac{1}{p^2} \;\blacksquare$$

1. Pairwise independence

2. Constructing a pairwise independent family of hash functions
   - Pairwise independent family of functions $\mathbb{Z}_p \to \mathbb{Z}_p$
   - Smaller range

From functions $\mathbb{Z}_p \to \mathbb{Z}_p$ to functions $\mathbb{Z}_u \to \mathbb{Z}_m$

- Recall $p \geq u$ ($\gg m$)
  - Require $p \geq 2m$
- Given function $f : \mathbb{Z}_p \to \mathbb{Z}_p$, define the function $h_f^{\langle u,m \rangle} : \mathbb{Z}_u \to \mathbb{Z}_m$ by setting

$$h_f^{\langle u,m \rangle}(k) = f(k) \bmod m$$

  for every $k \in \mathbb{Z}_u$

Lemma

If $\mathcal{F}$ is pairwise independent, then the family $\mathcal{H} = \left\{ h_f^{\langle u,m \rangle} \mid f \in \mathcal{F} \right\}$ is δ-approx. pairwise independent for $\delta = \left(1 + \frac{2m}{p}\right)^2$.

- Indeed, $\delta \to 1$ as $\frac{p}{m} \to \infty$

Proving the lemma

- Consider some $k, k' \in \mathbb{Z}_u$, $k \neq k'$
- Pick function f u.a.r. from $\mathcal{F}$
  - Let $h = h_f^{\langle u,m \rangle}$
- Given $j \in \mathbb{Z}_m$, let $M_j = \{w \in \mathbb{Z}_p \mid w = j \mod m\}$
- Observe:

$$\frac{p-m}{m} < \left\lfloor \frac{p}{m} \right\rfloor \leq |M_j| \leq \left\lceil \frac{p}{m} \right\rceil < \frac{p+m}{m}$$

- For $i, i' \in \mathbb{Z}_m$,

$$h(k) = i \iff f(k) \in M_i \quad \text{and} \quad h(k') = i' \iff f(k') \in M_{i'}$$

- Develop

$$\mathbb{P}\left(f(k) \in M_i \wedge f(k') \in M_{i'}\right) = \sum_{y \in M_i, y' \in M_{i'}} \mathbb{P}\left(f(k) = y \wedge f(k') = y'\right)$$

$$= |M_i| \cdot |M_{i'}| \cdot \frac{1}{p^2}$$

The proof continues

- Upper bound:

$$|M_i| \cdot |M_{i'}| \cdot \frac{1}{p^2} < \left(\frac{p+m}{m}\right)^2 \cdot \frac{1}{p^2} = \left(\frac{p+m}{p}\right)^2 \cdot \frac{1}{m^2}$$
$$= \left(1+\frac{m}{p}\right)^2 \cdot \frac{1}{m^2} < \left(1+\frac{2m}{p}\right)^2 \cdot \frac{1}{m^2} = \delta \cdot \frac{1}{m^2}$$

- Lower bound:

$$|M_i| \cdot |M_{i'}| \cdot \frac{1}{p^2} > \left(\frac{p-m}{m}\right)^2 \cdot \frac{1}{p^2} = \left(\frac{p-m}{p}\right)^2 \cdot \frac{1}{m^2}$$
$$= \left(1-\frac{m}{p}\right)^2 \cdot \frac{1}{m^2} \geq \frac{1}{\left(1+\frac{2m}{p}\right)^2} \cdot \frac{1}{m^2} = \frac{1}{\delta} \cdot \frac{1}{m^2}$$

  - Penultimate transition: $(1-z) \geq \frac{1}{(1+2z)}$ whenever $0 \leq z \leq \frac{1}{2}$ ∎