COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 3

Logistics

- Problem Set 1 has been posted on the course website and is due **Friday 9/23 at 11:59pm**.
- I have to end my office hours at 3pm today. I will add office hours from **11am-12pm next Tuesday 9/20** to compensate.
- We generally don't give extensions on quizzes, since we discuss solutions in class on Tuesday. To make up for this, we drop the lowest quiz grade at the end of the semester.
- On the quiz feedback question, several people asked for more practice questions/examples. Check out the MIT and Khan academy material posted on the Schedule. We will also keep having probability practice questions on the first few quizzes.
- It is common to not catch everything in lecture. I strongly encourage going back to the slides to review/check your understanding after class. Also come to office hours for more in-depth discussion/examples.

Content Overview

Last Class:

- Linearity of variance. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$ if X, Y independent
- Markov's inequality: the most fundamental concentration bound. $\Pr(X \geq t \cdot \mathbb{E}[X]) \leq 1/t$.
- Algorithmic applications of Markov's inequality, linearity of expectation, and indicator random variables:
  - Counting collisions to estimate CAPTCHA database size.
  - Start on analyzing hash tables with random hash functions.

Content Overview

Last Class:

- Linearity of variance.
- Markov's inequality: the most fundamental concentration bound. $\Pr(X \geq t \cdot \mathbb{E}[X]) \leq 1/t$.
- Algorithmic applications of Markov's inequality, linearity of expectation, and indicator random variables:
  - Counting collisions to estimate CAPTCHA database size.
  - Start on analyzing hash tables with random hash functions.

Today:

- Finish up random hash functions and hash tables.
- 2-level hashing, 2-universal and pairwise independent hash functions.

Quiz Questions

Question **4**

Not complete

Points out of 1.00

Flag question

Edit question

The expected number of inches of rain on Saturday is 4 and the expected number of inches on Sunday is 2. ~~There is a 50% chance of rain on Saturday. If it rains on Saturday, there is a 75% chance of rain on Sunday. If it does not rain on Saturday, there is only a 25% chance of rain on Sunday.~~ ~~What is the expected number of inches of rainfall~~ total over the ~~wee~~kend?

Answer:

Check

linearity of expectation

$\mathbb{E}[X_{sat}] = 4$

$\mathbb{E}[X_{sun}] = 2$

$\mathbb{E}[X_{sat} + X_{sun}] = 4 + 2 = 6$

Quiz Questions

$$\frac{\#\text{items}}{\#\text{buckets}} = \frac{1}{2}$$

m items

n buckets

You store 1000 items in a hash table with 2000 buckets. You use a fully random hash function to implement the table. What is the expected number of items stored in bucket i? Enter your answer to two decimal places.

Answer:

Check

Question 14

Not complete

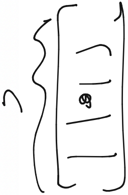
Points out of 1.00

Flag question

Edit question

$$\mathbb{E}X_{ij} = \underbrace{Pr(X_{ij}=1)}_{1/n}\cdot 1 + Pr(X_{ij}=0)\cdot 0$$

$$= 1/n$$



$$\mathbb{E}[X_j] = \mathbb{E}\left[\sum_{i=1}^{m} X_{ij}\right]$$

Bernoulli

$$\mathbb{E}[X_j] = \sum_{i=1}^{m} \underbrace{\mathbb{E}[X_{ij}]}_{1/n} = \frac{m}{n} = \frac{1000}{2000} = \frac{1}{2}$$

binomial

$X_{ij} = 1$ if item i is hashed to bucket j.

$X_{ij} = 0$ o.w.

X_j = # number of items in bucket j

Quiz Questions

As above, you store 1000 items in a hash table, this time with 10000 buckets. You use a fully random hash function to implement the table. What is the total expected number of pairwise collisions across the hash table? Enter your answer to two decimal places.

Answer:

Check

Question 15

Not complete

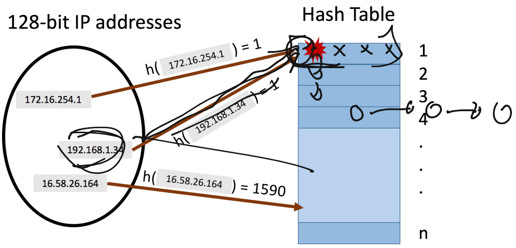
Points out of 1.00

Flag question

Edit question

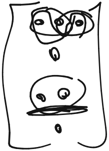
Hash Tables

We store m items from a large universe in a hash table with n positions.



- Want to show that when $\mathbf{h} : U \to [n]$ is a fully random hash function, query time is $O(1)$ with good probability.
- Equivalently: want to show that there are few collisions between hashed items.

Let $\mathsf{C}_{i,j} = 1$ if items i and j collide ($\mathsf{h}(x_i) = \mathsf{h}(x_j)$), and 0 otherwise. The number of pairwise ~~duplicates~~ collisions is:



$$\mathsf{C} = \sum_{i,j \in [m], i<j} \mathsf{C}_{i,j}.$$

x_i, x_j: pair of stored items, m: total number of stored items, n: hash table size, C: total pairwise collisions in table, h: random hash function.

Linearity of Expectation

Let $\mathsf{C}_{i,j} = 1$ if items i and j collide ($\mathsf{h}(x_i) = \mathsf{h}(x_j)$), and 0 otherwise. The number of pairwise duplicates is:

$$\mathbb{E}[\mathsf{C}] = \sum_{i,j \in [m], i<j} \mathbb{E}[\mathsf{C}_{i,j}]. \qquad \text{(linearity of expectation)}$$

$$Pr(C_{ij} = 1) = \frac{1}{n}$$

x_i, x_j: pair of stored items, m: total number of stored items, n: hash table size, C: total pairwise collisions in table, h: random hash function.

Linearity of Expectation

Let $\mathsf{C}_{i,j} = 1$ if items i and j collide ($\mathbf{h}(x_i) = \mathbf{h}(x_j)$), and 0 otherwise. The number of pairwise duplicates is:

$$\mathbb{E}[\mathsf{C}] = \sum_{i,j \in [m], i<j} \mathbb{E}[\mathsf{C}_{i,j}]. \qquad \text{(linearity of expectation)}$$

For any pair i, j, $i < j$:

$$\underline{\mathbb{E}[\mathsf{C}_{i,j}]} = \underline{\Pr[\mathsf{C}_{i,j} = 1]} = \Pr[\underline{\mathbf{h}(x_i) = \mathbf{h}(x_j)}] = \underline{\tfrac{1}{n}}.$$

x_i, x_j: pair of stored items, m: total number of stored items, n: hash table size, C: total pairwise collisions in table, $\mathbf{h}$: random hash function.

Linearity of Expectation

Let $C_{i,j} = 1$ if items i and j collide ($h(x_i) = h(x_j)$), and 0 otherwise. The number of pairwise duplicates is:

$$\mathbb{E}[C] = \sum_{i,j \in [m], i<j} \mathbb{E}[C_{i,j}]. \qquad \text{(linearity of expectation)}$$

For any pair i, j, $i < j$:

$\mathbb{E}[C_{i,j}] = \Pr[C_{i,j} = 1] = \Pr[h(x_i) = h(x_j)] = \frac{1}{n}$.

$$\mathbb{E}[C] = \underbrace{\sum_{i,j \in [m], i<j}} \frac{1}{n} = \frac{\binom{m}{2}}{n} = \boxed{\frac{m(m-1)}{2n}}.$$

x_i, x_j: pair of stored items, m: total number of stored items, n: hash table size, C: total pairwise collisions in table, h: random hash function.

Linearity of Expectation

Let $C_{i,j} = 1$ if items i and j collide ($h(x_i) = h(x_j)$), and 0 otherwise. The number of pairwise duplicates is:

$$\mathbb{E}[C] = \sum_{i,j \in [m], i<j} \mathbb{E}[C_{i,j}]. \qquad \text{(linearity of expectation)}$$

For any pair i, j, $i < j$:

$\mathbb{E}[C_{i,j}] = \Pr[C_{i,j} = 1] = \Pr[h(x_i) = h(x_j)] = \frac{1}{n}$.

$$\mathbb{E}[C] = \sum_{i,j \in [m], i<j} \frac{1}{n} = \frac{\binom{m}{2}}{n} = \frac{m(m-1)}{2n}.$$

Identical to the CAPTCHA analysis!

x_i, x_j: pair of stored items, m: total number of stored items, n: hash table size, C: total pairwise collisions in table, h: random hash function.

Collision Free Hashing

$$\boxed{\mathbb{E}[\mathsf{C}] = \frac{m(m-1)}{2n}.}$$

m: total number of stored items, *n*: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[\mathsf{C}] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[\mathsf{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.

$$\leq \frac{m^2}{2n} = \frac{m^2}{8m^2} \leq \frac{1}{8}$$

m: total number of stored items, *n*: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[C] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[C] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[C \geq 1]$. $\leq$

(assuming $n = 4m^2$)

$$\Pr[C \geq 1] \leq \frac{\mathbb{E}[C]}{1} \leq \frac{1}{8}$$

m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[\mathrm{C} \geq 1]$.

Apply Markov's Inequality:

m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[\mathrm{C} \geq 1]$.

Apply Markov's Inequality: $\Pr[\mathrm{C} \geq 1] \leq \frac{\mathbb{E}[\mathrm{C}]}{1}$

m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[\mathrm{C} \geq 1]$.

Apply Markov's Inequality: $\Pr[\mathrm{C} \geq 1] \leq \frac{\mathbb{E}[\mathrm{C}]}{1} = \frac{1}{8}$.

m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

$$\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[\mathrm{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[\mathrm{C} \geq 1]$.

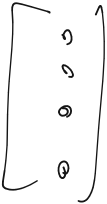
Apply Markov's Inequality: $\underline{\Pr[\mathrm{C} \geq 1]} \leq \frac{\mathbb{E}[\mathrm{C}]}{1} = \underline{\frac{1}{8}}$.

$$\Pr[\mathrm{C} = 0] = 1 - \Pr[\mathrm{C} \geq 1] \quad \geq$$

m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

Collision Free Hashing

$$\mathbb{E}[\mathbf{C}] = \frac{m(m-1)}{2n}.$$



- For $n = 4m^2$ we have: $\mathbb{E}[\mathbf{C}] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[\mathbf{C} \geq 1]$.

Apply Markov's Inequality: $\Pr[\mathbf{C} \geq 1] \leq \frac{\mathbb{E}[\mathbf{C}]}{1} \leq \frac{1}{8}$.

$$\underline{\Pr[\mathbf{C} = 0]} = 1 - \Pr[\mathbf{C} \geq 1] \geq 1 - \frac{1}{8} = \frac{7}{8}$$

> m: total number of stored items, n: hash table size, $\mathbf{C}$: total pairwise collisions in table.

Collision Free Hashing

m = # items

space

$$\mathbb{E}[C] = \frac{m(m-1)}{2n}.$$

- For $n = 4m^2$ we have: $\mathbb{E}[C] = \frac{m(m-1)}{8m^2} \leq \frac{1}{8}$.
- **Think-Pair-Share:** Give an upper bound on the probability that we have at least one collision, i.e., $\Pr[C \geq 1]$.

Apply Markov's Inequality: $\Pr[C \geq 1] \leq \frac{\mathbb{E}[C]}{1} = \frac{1}{8}$.

$$\Pr[C = 0] = 1 - \Pr[C \geq 1] \geq 1 - \frac{1}{8} = \frac{7}{8}$$

Pretty good...but we are using $O(m^2)$ space to store m items...

> m: total number of stored items, n: hash table size, C: total pairwise collisions in table.

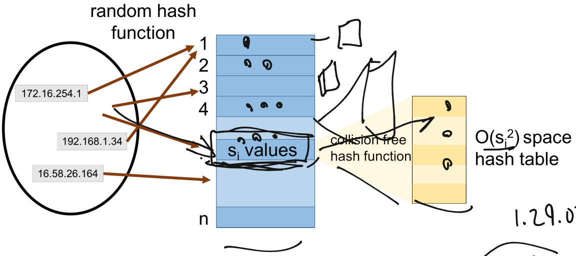
Two Level Hashing

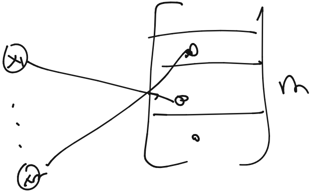
Want to preserve $O(1)$ query time while using $O(m)$ space.

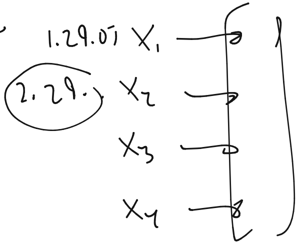
Want to preserve $O(1)$ query time while using $O(m)$ space.

Two-Level Hashing:

S_i = # items in bucket i



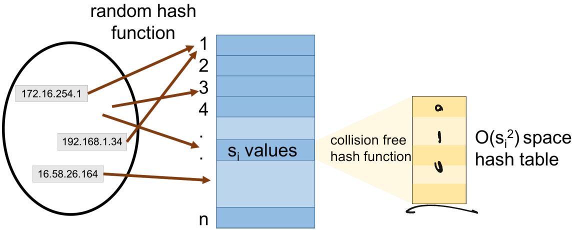




Two Level Hashing

Want to preserve $O(1)$ query time while using $O(m)$ space.

Two-Level Hashing:

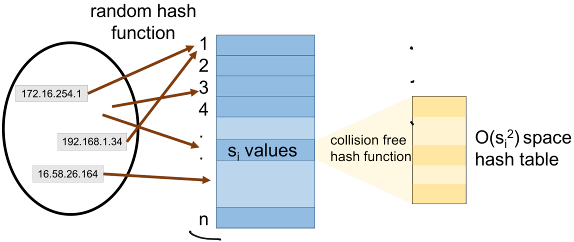


- For each bucket with s_i values, pick a collision free hash function mapping $[s_i] \to [s_i^2]$.

Two Level Hashing

Want to preserve $O(1)$ query time while using $O(m)$ space.

Two-Level Hashing:



- For each bucket with s_i values, pick a collision free hash function mapping $[s_i] \to [s_i^2]$.
- **Just Showed:** A random function is collision free with probability $\geq \frac{7}{8}$ so can just generate a random hash function and check if it is collision free.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions.

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: $\#$ items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

size of i^{th} table at level 2

Up to constants, space used is: $\mathbf{S} = n + \sum_{i=1}^{n} \mathbf{s}_i^2 = O(m)$

n = # buckets in table 1

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\underbrace{\sum_{j=1}^{m} \overset{X_{ij}}{\mathbb{I}_{\mathbf{h}(x_j)=i}}}_{s_i}\right)^2\right]$$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \underbrace{\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}}_{x_j \text{ \& } x_k \text{ both hashed to bucket } i}\right]$$

$$(x_1 + x_2 + \dots x_m)^2$$

$$(x_1 + x_2 + \dots x_m)(x_1 + x_2 \dots)$$

Collisions again!

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$\mathbb{E}\mathbf{s}_i = \frac{m}{n}$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,

> x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,

$$\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\underbrace{\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2}\right] = \mathbb{E}\left[\mathbb{I}_{h(x_j)=1}\right] = \frac{1}{n}$$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,

$$\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i]$$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \underbrace{\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]}_{1/n}.$$

- For $j = k$,

$$\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i] = \frac{1}{n}.$$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,
$\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i] = \frac{1}{n}.$

- For $j \neq k$,

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,
  $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i] = \frac{1}{n}.$
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,
  $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i] = \frac{1}{\underline{n}}.$
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \Pr[\underline{\mathbf{h}(x_j) = i} \cap \underline{\mathbf{h}(x_k) = i}]$

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

Query time for two level hashing is $O(1)$: requires evaluating two hash functions. What is the expected space usage?

Up to constants, space used is: $\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2]$

1 if item j hashes bucket i

$$S_i = \sum_{j=1}^{m} \mathbb{I}_{h(x_j)=i}$$

$$\mathbb{E}[\mathbf{s}_i^2] = \mathbb{E}\left[\left(\sum_{j=1}^{m} \mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j,k\in[m]} \mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right].$$

- For $j = k$,
  $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \mathbb{E}\left[\left(\mathbb{I}_{\mathbf{h}(x_j)=i}\right)^2\right] = \Pr[\mathbf{h}(x_j) = i] = \frac{1}{n}.$
- For $j \neq k$, $\mathbb{E}\left[\overset{1,0}{\mathbb{I}_{\mathbf{h}(x_j)=i}} \cdot \overset{1,0}{\mathbb{I}_{\mathbf{h}(x_k)=i}}\right] = \Pr[\mathbf{h}(x_j) = i \cap \mathbf{h}(x_k) = i] = \frac{1}{n^2}.$

= 1 if items j+k hash to bucket i

= 0 o.w.

x_j, x_k: stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored in hash table at position i.

Space Usage

$$\mathbb{E}[\mathsf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathsf{h}(x_j)=i} \cdot \mathbb{I}_{\mathsf{h}(x_k)=i}\right]$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathsf{h}(x_j)=i} \cdot \mathbb{I}_{\mathsf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathsf{h}(x_j)=i} \cdot \mathbb{I}_{\mathsf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

x_j, x_k: stored items, m: # stored items, n: hash table size, h: random hash function, S: space usage of two level hashing, s_i: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathbf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$
$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

$$(x_1 + x_2 + x_3)(x_1 + x_2 + x_3)$$
$$x_1^2 + x_1 x_2 + \ldots$$

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathbf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathbf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.

$$\mathbb{E}[\mathbf{s}_i^2] = \sum_{j,k\in[m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \sum_{j=k=1}^{m} \underbrace{\mathbb{E}[\dots]}_{1/n} + \sum_{\substack{j\neq k\\ j,k\in[m]}} \underbrace{\mathbb{E}[\dots]}_{1/n^2}$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

$$= \frac{m}{n} + \frac{m(m-1)}{n^2}$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathsf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$
$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$
$$= \frac{m}{n} + \frac{m(m-1)}{n^2} \leq 2 \text{ (If we set } n = m.)$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathsf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

1 m^2

$$= \underbrace{\frac{m}{n} + \frac{m(m-1)}{n^2}}_{} \leq 2 \text{ (If we set } n = m.)$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

Total Expected Space Usage: (if we set $n = m$)

2

$$\mathbb{E}[\mathsf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathsf{s}_i^2]$$

x_j, x_k: stored items, m: # stored items, n: hash table size, **h**: random hash function, **S**: space usage of two level hashing, s_i: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathsf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

$$= \frac{m}{n} + \frac{m(m-1)}{n^2} \leq 2 \text{ (If we set } n = m.)$$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

Total Expected Space Usage: (if we set $n = m$)

$$\mathbb{E}[\mathsf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathsf{s}_i^2] \leq n + n \cdot 2 = 3n = 3m.$$

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, S: space usage of two level hashing, s_i: # items stored at pos i.

Space Usage

$$\mathbb{E}[\mathbf{s}_i^2] = \sum_{j,k \in [m]} \mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right]$$

$$= m \cdot \frac{1}{n} + 2 \cdot \binom{m}{2} \cdot \frac{1}{n^2}$$

$$= \frac{m}{n} + \frac{m(m-1)}{n^2} \leq 2 \text{ (If we set } n = m.\text{)}$$

$(x_1 + x_2)(x_1 + x_2)$

$x_1^2 + \underline{x_2^2 + x_1 x_2} + x_2 x_1$

$1 \cdot 1$

$0 \cdot 0$

- For $j = k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n}$.
- For $j \neq k$, $\mathbb{E}\left[\mathbb{I}_{\mathbf{h}(x_j)=i} \cdot \mathbb{I}_{\mathbf{h}(x_k)=i}\right] = \frac{1}{n^2}$.

Total Expected Space Usage: (if we set $n = m$)

$$\mathbb{E}[\mathbf{S}] = n + \sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i^2] \leq n + n \cdot 2 = 3n = 3m.$$

Near optimal space with $O(1)$ query time!

x_j, x_k: stored items, m: # stored items, n: hash table size, $\mathbf{h}$: random hash function, $\mathbf{S}$: space usage of two level hashing, $\mathbf{s}_i$: # items stored at pos i.