Linear Probing

Outline for Today

- ***Count Sketches***
  - We didn’t get there last time, and there’s lots of generalizable ideas here. Let’s go exploring!
- ***Linear Probing***
  - A simple and lightning fast hash table implementation.
- ***Analyzing Linear Probing***
  - Why the degree of independence matters.
- ***Fourth Moment Bounds***
  - Another approach for estimating frequencies.

Recap from Last Time

Distribution Property: Each element should have an equal probability of being placed in each slot.

For any $x \in \mathscr{U}$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

Independence Property: Where one element is placed shouldn't impact where a second goes.

For any distinct $x, y \in \mathscr{U}$ and random $h \in \mathcal{H}$, $h(x)$ and $h(y)$ are independent random variables.

A family of hash functions $\mathscr{H}$ is called ***2-independent*** (or ***pairwise independent***) if it satisfies the distribution and independence properties.

Suppose there are two tunable values

$$\varepsilon \in (0, 1]$$
$$\delta \in (0, 1]$$

where ε represents ***accuracy*** and δ represents ***confidence***.

Goal: Make an estimator $\hat{A}$ for some quantity A where

With probability at least 1 – δ, ***Probably***

$|\hat{A} - A| \leq \varepsilon \cdot size(input)$ ***Approximately Correct***

for some measure of the size of the input.

What does it mean for an approximation to be "good"?

How to Build an Estimator

1. Design a simple data structure that, intuitively, gives you a good estimate.
2. Use a ***sum of indicator variables*** and ***linearity of expectation*** to prove that, on expectation, the data structure is pretty close to correct.
3. Use a ***concentration inequality*** to show that the data structure's output is close to its expectation.
4. Run multiple copies of the data structure in parallel to amplify the success probability.

New Stuff!

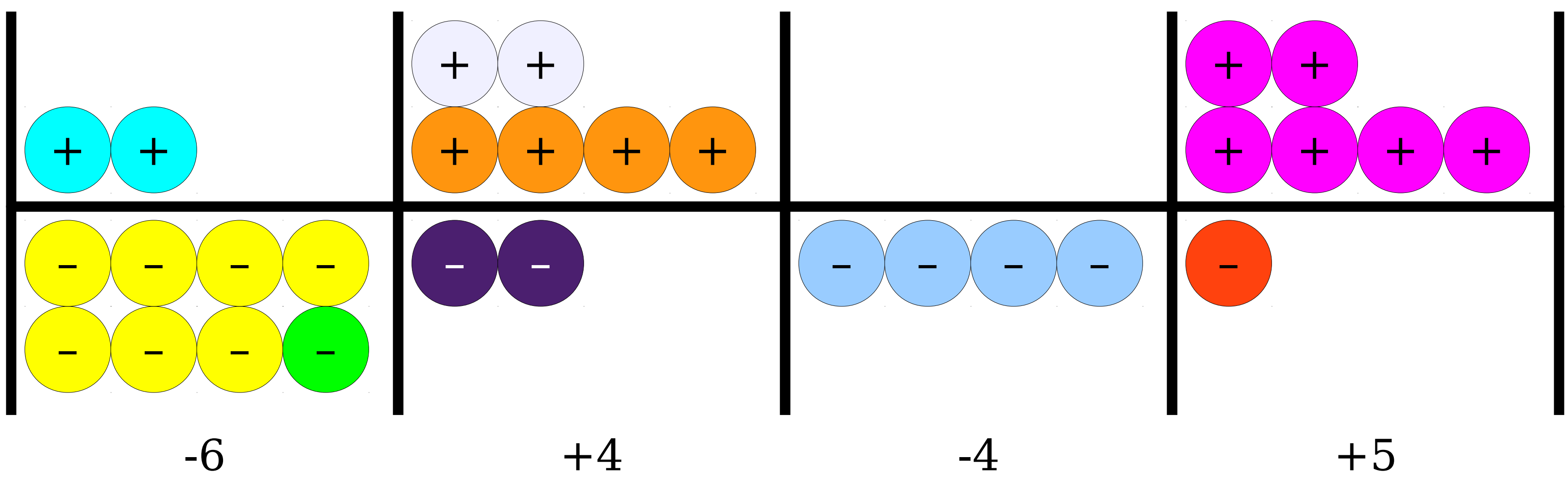
The Count Sketch

Frequency Estimation

- ***Recall:*** A frequency estimator is a data structure that supports
  - ***increment***(x), which increments the number of times that we've seen x, and
  - ***estimate***(x), which returns an estimate of how many times we've seen x.
- ***Notation:*** Assume that the elements we're processing are $x_1, \ldots, x_n$, and that the true frequency of element x_i is $\boldsymbol{a}_i$.
- Remember that the frequencies are not random variables – we're assuming that they're not under our control. Any randomness comes from hash functions.

The Setup

- As before, for some parameter w, we'll create an array **`count`** of length w.
- As before, choose a hash function $h : \mathscr{U} \rightarrow [w]$ from a family $\mathscr{H}$.
- For each $x_i \in \mathscr{U}$, assign x_i either +1 or -1.
- To ***increment***(x), go to **`count`**$[h(x)]$ and add ±1 as appropriate.
- To ***estimate***(x), return **`count`**$[h(x)]$, multiplied by ±1 as appropriate.

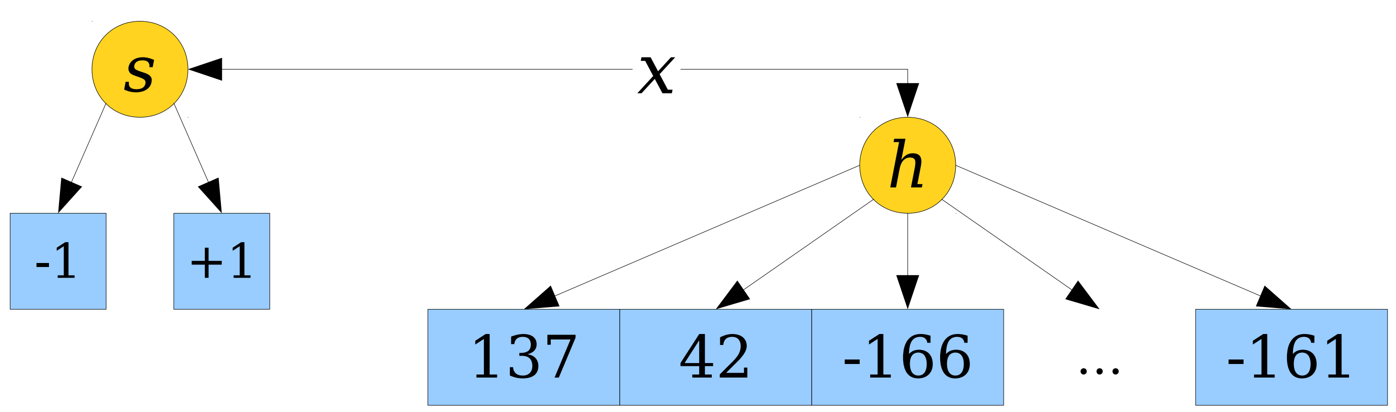


The Intuition

- Think about what introducing the ± 1 term does when collisions occur.
- If an element x collides with a frequent element y, we're not going to get a good estimate for x (but we wouldn't have gotten one anyway).
- If x collides with multiple infrequent elements, the collisions between those elements will partially offset one another and leave a better estimate for x.

More Formally

- Let's have $h \in \mathscr{H}$ chosen uniformly at random from a 2-independent family of hash functions from $\mathscr{U}$. to w.
- Choose $s \in \mathscr{U}$ uniformly randomly and independently of h from a 2-independent family from $\mathscr{U}$ to {-1, +1}.
- To ***increment***(x), add $s(x)$ to **count**[$h(x)$].
- To ***estimate***(x), return $s(x) \cdot$ **count**[$h(x)$].



Formalizing the Intuition

- Define $\hat{\boldsymbol{a}}_i$ to be our estimate of $\boldsymbol{a}_i$.
- As before, $\hat{\boldsymbol{a}}_i$ will depend on how the other elements are distributed. Unlike before, it now also depends on signs given to the elements by s.
- Specifically, for each other x_j that collides with x_i, the estimate $\hat{\boldsymbol{a}}_i$ includes an error term of

$$s(x_i) \cdot s(x_j) \cdot \boldsymbol{a}_j$$

- Why?
  - The counter for x_i will have $s(x_j)\ \boldsymbol{a}_j$ added in.
  - We multiply the counter by $s(x_i)$ before returning it.

Formalizing the Intuition

- Define $\hat{\boldsymbol{a}}_i$ to be our estimate of $\boldsymbol{a}_i$.
- As before, $\hat{\boldsymbol{a}}_i$ will depend on how the other elements are distributed. Unlike before, it now also depends on signs given to the elements by s.
- Specifically, for each other x_j that collides with x_i, the estimate $\hat{\boldsymbol{a}}_i$ includes an error term of

$$s(x_i) \cdot s(x_j) \cdot \boldsymbol{a}_j$$

- Why?
  - If $s(x_i)$ and $s(x_j)$ point in the same direction, the terms add to the total.
  - If $s(x_i)$ and $s(x_j)$ point in different directions, the terms subtract from the total.

Formalizing the Intuition

- In our quest to learn more about $\hat{\boldsymbol{a}}_i$, let's have X_j be a random variable indicating whether x_i and x_j collided with one another:

$$X_j = \begin{cases} 1 & \text{if } h(x_i) = h(x_j) \\ 0 & \text{if } h(x_i) \neq h(x_j) \end{cases}$$

- We can then express $\hat{\boldsymbol{a}}_i$ in terms of the signed contributions from the items x_i collides with:

$$\hat{\boldsymbol{a}}_i = \sum_j \boldsymbol{a}_j s(x_i) s(x_j) X_j = \boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j$$

This is how much the collision impacts our estimate.

We only care about items we collided with.

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}_i] &= \mathrm{E}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{E}[\boldsymbol{a}_i] + \mathrm{E}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j]
\end{aligned}
$$

Hey, it's linearity of expectation!

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}_i] &= \mathrm{E}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{E}[\boldsymbol{a}_i] + \mathrm{E}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j s(x_i) s(x_j) X_j]
\end{aligned}
$$

Remember that $\boldsymbol{a}_i$ and the like aren't random variables.

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}_i] &= \mathrm{E}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{E}[\boldsymbol{a}_i] + \mathrm{E}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[s(x_i) s(x_j)] \mathrm{E}[\boldsymbol{a}_j X_j]
\end{aligned}
$$

We chose the hash functions h and s independently of one another.

$$
X_j = \begin{cases} 1 & \text{if } h(x_i) = h(x_j) \\ 0 & \text{if } h(x_i) \neq h(x_j) \end{cases}
$$

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}_i] &= \mathrm{E}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{E}[\boldsymbol{a}_i] + \mathrm{E}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[s(x_i) s(x_j)] \mathrm{E}[\boldsymbol{a}_j X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[s(x_i)] \mathrm{E}[s(x_j)] \mathrm{E}[\boldsymbol{a}_j X_j]
\end{aligned}
$$

Since s is drawn from a 2-independent family of hash functions, we know $s(x_i)$ and $s(x_j)$ are independent random variables.

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}_i] &= \mathrm{E}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{E}[\boldsymbol{a}_i] + \mathrm{E}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[s(x_i) s(x_j)] \mathrm{E}[\boldsymbol{a}_j X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} \mathrm{E}[s(x_i)] \mathrm{E}[s(x_j)] \mathrm{E}[\boldsymbol{a}_j X_j] \\
&= \boldsymbol{a}_i + \sum_{j \neq i} 0 \\
&= \boldsymbol{a}_i
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[s(x_i)] &= \tfrac{1}{2} \cdot (-1) + \tfrac{1}{2} \cdot (+1) \\
&= 0
\end{aligned}
$$

s is drawn from a 2-independent family of hash functions.

$s(x_i)$ is uniform over $\{-1, +1\}$

$\Pr[s(x_i) = -1] = \tfrac{1}{2}$ $\Pr[s(x_i) = +1] = \tfrac{1}{2}$

A Hitch

- In the count-min sketch, we used Markov's inequality to bound the probability that we get a bad estimate.
- This worked because we had a ***one-sided error***: the distance $\hat{a}_i - a_i$ from the true answer was nonnegative.
- With the count sketch, we have a ***two-sided error***: $\hat{a}_i - a_i$ can be negative in the count sketch because collisions can *decrease* the estimate $\hat{a}_i$ below the true value a_i.
- We'll need to use a different technique to bound the error.

Chebyshev to the Rescue

- ***Chebyshev's inequality*** states that for any random variable X with finite variance, given any $c > 0$, we have

$$\Pr\left[\ |X - \mathrm{E}[X]| > c\ \right] < \frac{\mathrm{Var}[X]}{c^2}.$$

- If we can get the variance of $\hat{a}_i$, we can bound the probability that we get a bad estimate with our data structure.

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j]
\end{aligned}
$$

$$\boxed{\mathrm{Var}[a + X] = \mathrm{Var}[X]}$$

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j \neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j]
\end{aligned}
$$

In general, Var is *not* a linear operator.

However, if the terms in the sum are ***pairwise uncorrelated***, then Var is linear.

Lemma: The terms in this sum are uncorrelated. *(Prove this!)*

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j \neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&\leq \sum_{j \neq i} \mathrm{E}[(\boldsymbol{a}_j s(x_i) s(x_j) X_j)^2]
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
\mathrm{Var}[Z] &= \mathrm{E}[Z^2] - \mathrm{E}[Z]^2 \\
&\leq \mathrm{E}[Z^2]
\end{aligned}
}
$$

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j \neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&\leq \sum_{j \neq i} \mathrm{E}[(\boldsymbol{a}_j s(x_i) s(x_j) X_j)^2] \\
&= \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j^2 s(x_i)^2 s(x_j)^2 X_j^2] \\
&= \sum_{j \neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j^2]
\end{aligned}$$

$s(x) = \pm 1$,

so

$s(x)^2 = 1$

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j\neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j\neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j\neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&\leq \sum_{j\neq i} \mathrm{E}[(\boldsymbol{a}_j s(x_i) s(x_j) X_j)^2] \\
&= \sum_{j\neq i} \mathrm{E}[\boldsymbol{a}_j^2 s(x_i)^2 s(x_j)^2 X_j^2] \\
&= \sum_{j\neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j^2]
\end{aligned}
$$

Useful Fact: If X is an indicator, then $X^2 = X$.

$$
X_j^2 = \begin{cases} 1 & \text{if } h(x_i) = h(x_j) \\ 0 & \text{if } h(x_i) \neq h(x_j) \end{cases}
$$

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j \neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&\leq \sum_{j \neq i} \mathrm{E}[(\boldsymbol{a}_j s(x_i) s(x_j) X_j)^2] \\
&= \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j^2 s(x_i)^2 s(x_j)^2 X_j^2] \\
&= \sum_{j \neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j^2] \\
&= \sum_{j \neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j] \\
&= \frac{1}{w} \sum_{j \neq i} \boldsymbol{a}_j^2
\end{aligned}
$$

$$
\boxed{X_j = \begin{cases} 1 & \text{if } h(x_i) = h(x_j) \\ 0 & \text{if } h(x_i) \neq h(x_j) \end{cases}}
$$

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{a}}_i] &= \mathrm{Var}[\boldsymbol{a}_i + \sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \mathrm{Var}[\sum_{j \neq i} \boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&= \sum_{j \neq i} \mathrm{Var}[\boldsymbol{a}_j s(x_i) s(x_j) X_j] \\
&\leq \sum_{j \neq i} \mathrm{E}[(\boldsymbol{a}_j s(x_i) s(x_j) X_j)^2] \\
&= \sum_{j \neq i} \mathrm{E}[\boldsymbol{a}_j^2 s(x_i)^2 s(x_j)^2 X_j^2] \\
&= \sum_{j \neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j^2] \\
&= \sum_{j \neq i} \boldsymbol{a}_j^2 \mathrm{E}[X_j] \\
&= \frac{1}{w} \sum_{j \neq i} \boldsymbol{a}_j^2
\end{aligned}
$$

I know this might look really dense, but many of these substeps end up being really useful techniques. These ideas generalize, I promise.

Think of [$\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, …] as a vector.

What does the following quantity represent?

$$\sum_j \boldsymbol{a}_j^2$$

This is the square of the magnitude of the vector!

The magnitude of a vector is called its ***L_2 norm*** and is denoted $\|\boldsymbol{a}\|_2$.

$$\|\boldsymbol{a}\|_2 = \sqrt{\sum_j \boldsymbol{a}_j^2}$$

Therefore, our above sum is $\|\boldsymbol{a}\|_2^2$.

$$\mathrm{Var}[\hat{\boldsymbol{a}}_i] \leq \frac{1}{w}\sum_{j \neq i} \boldsymbol{a}_j^2 \leq \frac{\|\boldsymbol{a}\|_2^2}{w}$$

Think of [$\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, …] as a vector.

What does the following quantity represent?

$$\sum_j \boldsymbol{a}_j^2$$

This is the square of the mag

The magnitude of a vector is c
is denoted $\|$

Great exercise: Prove that the L_2 norm of a vector is never greater than the L_1 norm.

$$\|\boldsymbol{a}\|_2 = \sqrt{\sum_j \boldsymbol{a}_j^2}$$

Therefore, our above sum is $\|\boldsymbol{a}\|_2^2$.

$$\mathrm{Var}[\hat{\boldsymbol{a}}_i] \leq \frac{1}{w}\sum_{j \neq i} \boldsymbol{a}_j^2 \leq \frac{\|\boldsymbol{a}\|_2^2}{w}$$

Goal: Make an estimator $\hat{\boldsymbol{a}}$ for some quantity $\boldsymbol{a}$ where

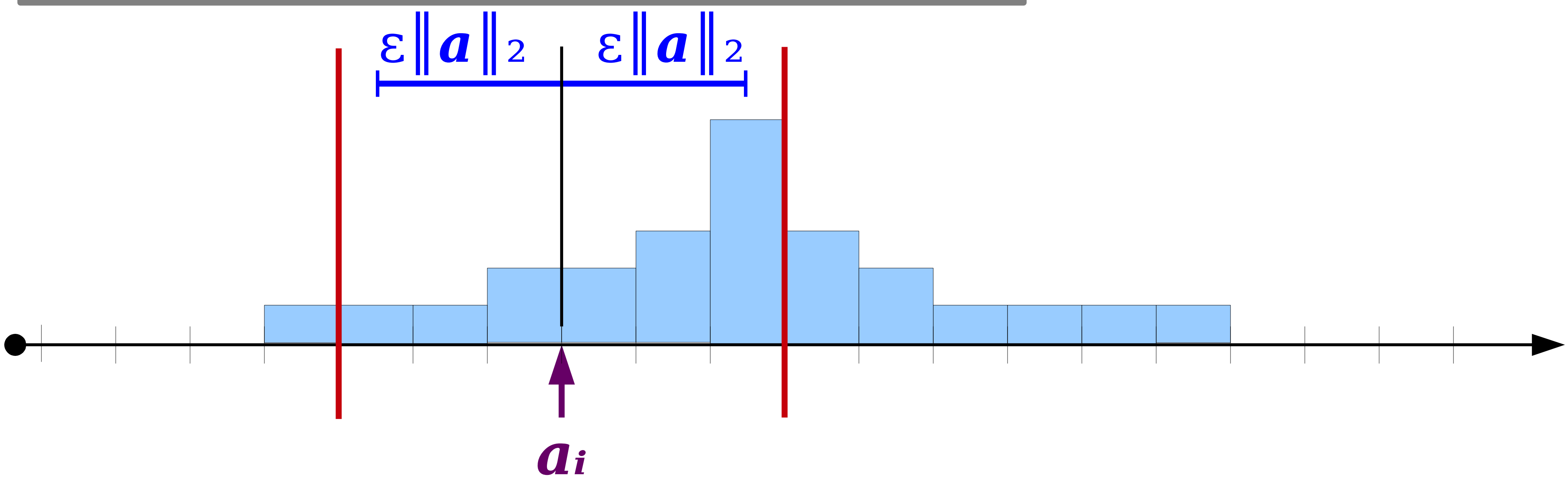
With probability at least $1 - \delta$,

$|\hat{\boldsymbol{a}} - \boldsymbol{a}| \leq \varepsilon \cdot size(input)$

for some measure of the size of the input.

Probably

Approximately Correct



$$\mathrm{Var}[\hat{\boldsymbol{a}}_i] \leq \frac{\|\boldsymbol{a}\|_2^2}{w}$$

$$\Pr\left[\left|\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i\right| > \varepsilon \|\boldsymbol{a}\|_2\right]$$

$$< \quad \frac{\mathrm{Var}[\hat{\boldsymbol{a}}_i]}{\left(\varepsilon \|\boldsymbol{a}\|_2\right)^2}$$

Chebyshev's inequality says that

$$\Pr\left[\ |X - \mathrm{E}[X]| > c\ \right] < \frac{\mathrm{Var}[X]}{c^2}.$$

$$\Pr\left[\left|\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i\right| > \varepsilon\|\boldsymbol{a}\|_2\right]$$

$$< \frac{\mathrm{Var}[\hat{\boldsymbol{a}}_i]}{\left(\varepsilon\|\boldsymbol{a}\|_2\right)^2}$$

$$\leq \frac{\|\boldsymbol{a}\|_2^2}{w} \cdot \frac{1}{\left(\varepsilon\|\boldsymbol{a}\|_2\right)^2}$$

$$\boxed{\mathrm{Var}[\hat{\boldsymbol{a}}_i] \leq \frac{\|\boldsymbol{a}\|_2^2}{w}}$$

$$
\begin{aligned}
& \Pr\left[\left|\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i\right| > \varepsilon \|\boldsymbol{a}\|_2\right] \\
< \quad & \frac{\mathrm{Var}[\hat{\boldsymbol{a}}_i]}{\left(\varepsilon \|\boldsymbol{a}\|_2\right)^2} \\
\leq \quad & \frac{\|\boldsymbol{a}\|_2^2}{w} \cdot \frac{1}{\left(\varepsilon \|\boldsymbol{a}\|_2\right)^2} \\
= \quad & \frac{1}{w\varepsilon^2}
\end{aligned}
$$

Goal: Make an estimator $\hat{\boldsymbol{a}}$ for some quantity $\boldsymbol{a}$ where

With probability at least $1 - \delta$, **Probably**

$|\hat{\boldsymbol{a}} - \boldsymbol{a}| \leq \varepsilon \cdot \mathit{size}(\mathit{input})$ **Approximately Correct**

for some measure of input size.

$$\Pr[|\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i| > \varepsilon \|\boldsymbol{a}\|_2] \leq \frac{1}{w\varepsilon^2}$$

Pick $w = e \cdot \varepsilon^{-2}$. Then

$$\Pr[|\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i| > \varepsilon \|\boldsymbol{a}\|_2] \leq e^{-1}.$$

We now have a single estimator with a not-so-great chance of giving a good estimate.

How do we fix this?

Running in Parallel

- Let's suppose that we run ***d*** independent copies of this data structure. Each has its own independently randomly chosen hash function.
- To ***increment***(x) in the overall structure, we call ***increment***(x) on each of the underlying data structures.
- The probability that at least one of them provides a good estimate is quite high.
- ***Question:*** How do you know which one?

Estimator 1:	*Estimator 2:*	*Estimator 3:*	*Estimator 4:*	*Estimator 5:*
137	271	166	103	261

Working with the Median

- ***Claim:*** If we output the median estimate given by the data structures, we have high probability of giving the right answer.
- ***Intuition:*** The only way we report an answer more than $\varepsilon\|\boldsymbol{a}\|_2$ is if at least half of the data structures output an answer that is more than $\varepsilon\|\boldsymbol{a}\|_2$ from the true answer.
- Each individual data structure is wrong with probability at most e^{-1}, so this is highly unlikely.

The Setup

- Let D denote a random variable equal to the number of data structures that produce an answer *not* within $\varepsilon\|\boldsymbol{a}\|_2$ of the true answer.
- Since each independent data structure has failure probability at most e^{-1}, we can upper-bound D with a Binom(d, e^{-1}) variable.
- We want to know $\Pr[D > d / 2]$.
- How can we determine this?

Chernoff Bounds

- The ***Chernoff bound*** says that if $X \sim \text{Binom}(n, p)$ and $p < 1/2$, then

$$\Pr[X > n/2] < e^{\frac{-n(1/2-p)^2}{2p}}$$

- In our case, $D \sim \text{Binom}(d, 1/e)$, so we know that

$$\Pr\left[D > \frac{d}{2}\right] \leq e^{\frac{-d(1/2-1/e)^2}{2(1/e)}}$$

$$= e^{-k \cdot d} \quad \textit{(for some constant k)}$$

- Therefore, choosing $d = k^{-1} \cdot \log \delta^{-1}$ ensures that $\Pr[D > d / 2] \leq \delta$.
- Therefore, the success probability is at least $1 - \delta$.

Chernoff Bounds

- The ***Chernoff bound*** says that if $X \sim \text{Binom}(n, p)$ and $p < 1/2$, then

$$\Pr[X > n/2] < e^{\frac{-n(1/2-p)^2}{2p}}$$

- In our case, $D \sim \text{Binom}(d, 1/e)$, so we know that

$$e^{\frac{-d(1/2-1/e)^2}{2(1/e)}}$$

$$e^{-k \cdot d} \qquad \textit{(for some constant k)}$$

The specific constant factor here matters, since it's an exponent! To implement this data structure, you'll need to work out the exact value.

- Therefore, choosing $d = k^{-1} \cdot \log \delta^{-1}$ ensures that $\Pr[D > d / 2] \leq \delta$.
- Therefore, the success probability is at least $1 - \delta$.

The Overall Construction

- The ***count sketch*** is the data structure given as follows.
- Given ε and δ, choose

$$w = \lceil e / \varepsilon^2 \rceil \qquad d = \Theta(\log \delta^{-1})$$

- Create an array `count` of $w \times d$ counters.
- Choose hash functions h_i and s_i for each of the d rows.
- To ***increment***(x), add $s_i(x)$ to `count`$[i][h_i(x)]$ for each row i.
- To ***estimate***(x), return the median of $s_i(x) \cdot$ `count`$[i][h_i(x)]$ for each row i.

The Final Analysis

- With probability at least $1 - \delta$, all estimates are accurate to within a factor of $\varepsilon \| \boldsymbol{a} \|_2$.
- Space usage is $\Theta(w \cdot d)$, which we've seen to be $\Theta(\varepsilon^{-2} \cdot \log \delta^{-1})$.
- Updates and queries run in time $\Theta(\delta^{-1})$.
- Compared to the Count-Min Sketch:
  - Accuracy guarantees are relative to $\| \boldsymbol{a} \|_2$ versus $\| \boldsymbol{a} \|_1$.
  - Uses a factor of ε^{-1} additional space.
- ***Question to ponder:*** Which would you prefer if your elements are more uniform? Which would you prefer if a few elements are extremely common?

Time-Out for Announcements!

Problem Set Four

- Problem Set Four is due one week from today.
- As usual, get in touch with us if you have any questions!
  - Ask on Piazza!
  - Stop by office hours!

Project Checkpoints

- Project checkpoints are due this Thursday at 2:30PM.
  - ***No late periods may be used***; we're hoping to get feedback released ASAP.
- As a reminder, you should
  - summarize your progress so far in understanding your topic,
  - address the questions we sent over email, and
  - describe your proposal for your "interesting" component.

Back to CS166!

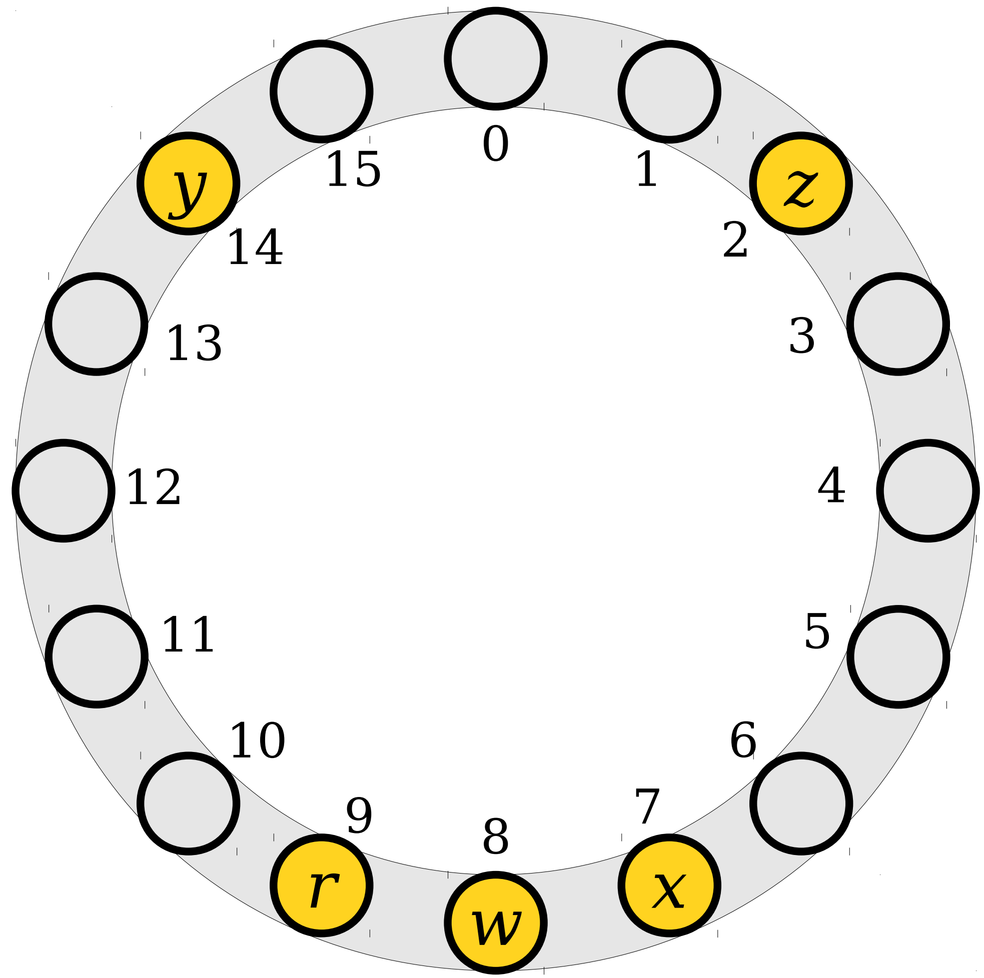
Hash Tables

Hashing Strategies

- All hash table implementations need to address what happens when collisions occur.
- Common strategies:
  - ***Closed addressing:*** Store all elements with hash collisions in a secondary data structure (linked list, BST, etc.)
  - ***Perfect hashing:*** Choose hash functions to ensure that collisions don't happen, and rehash or move elements when they do.
  - ***Open addressing:*** Allow elements to "leak out" from their preferred position and spill over into other positions.
- Linear probing is an example of open addressing.
- We'll see a type of perfect hashing (***cuckoo hashing***) on Thursday.

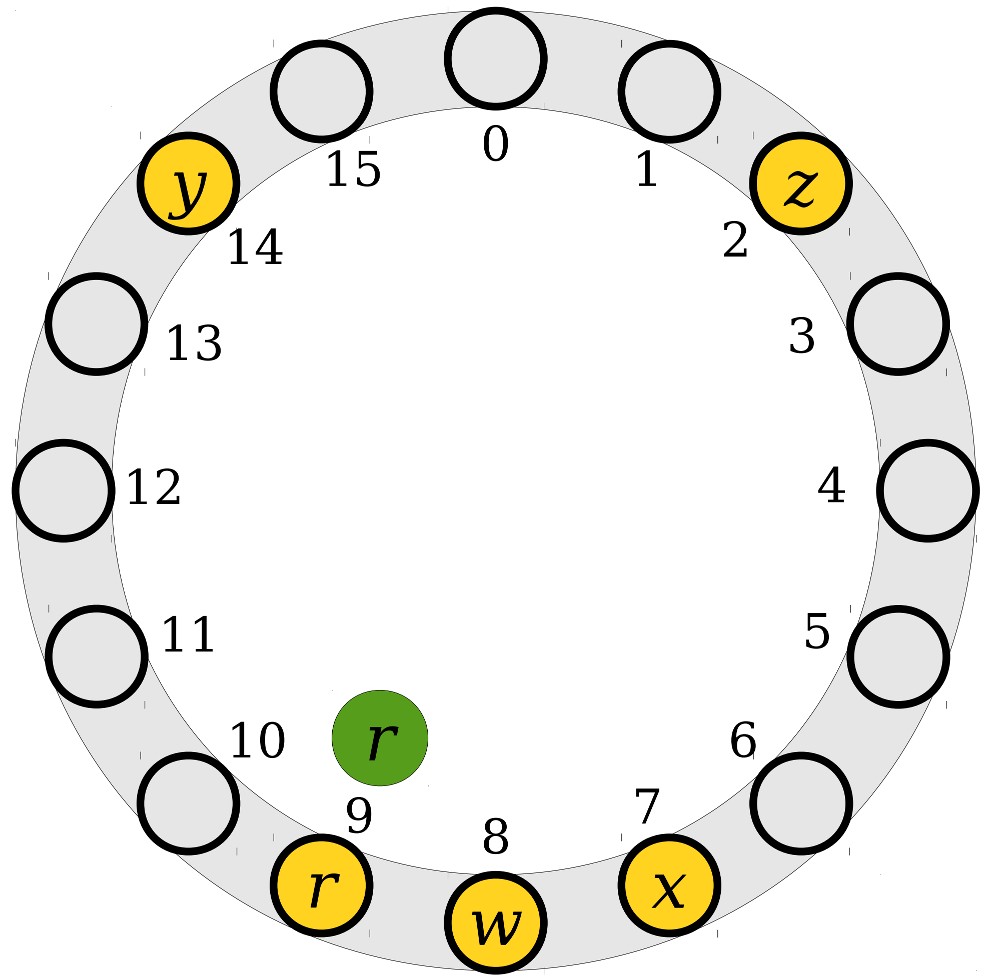
Linear Probing

- ***Linear probing*** is a simple open-addressing hashing strategy.
- To insert an element x, compute $h(x)$ and try to place x there.
- If that spot is occupied, keep moving through the array, wrapping around at the end, until a free spot is found.



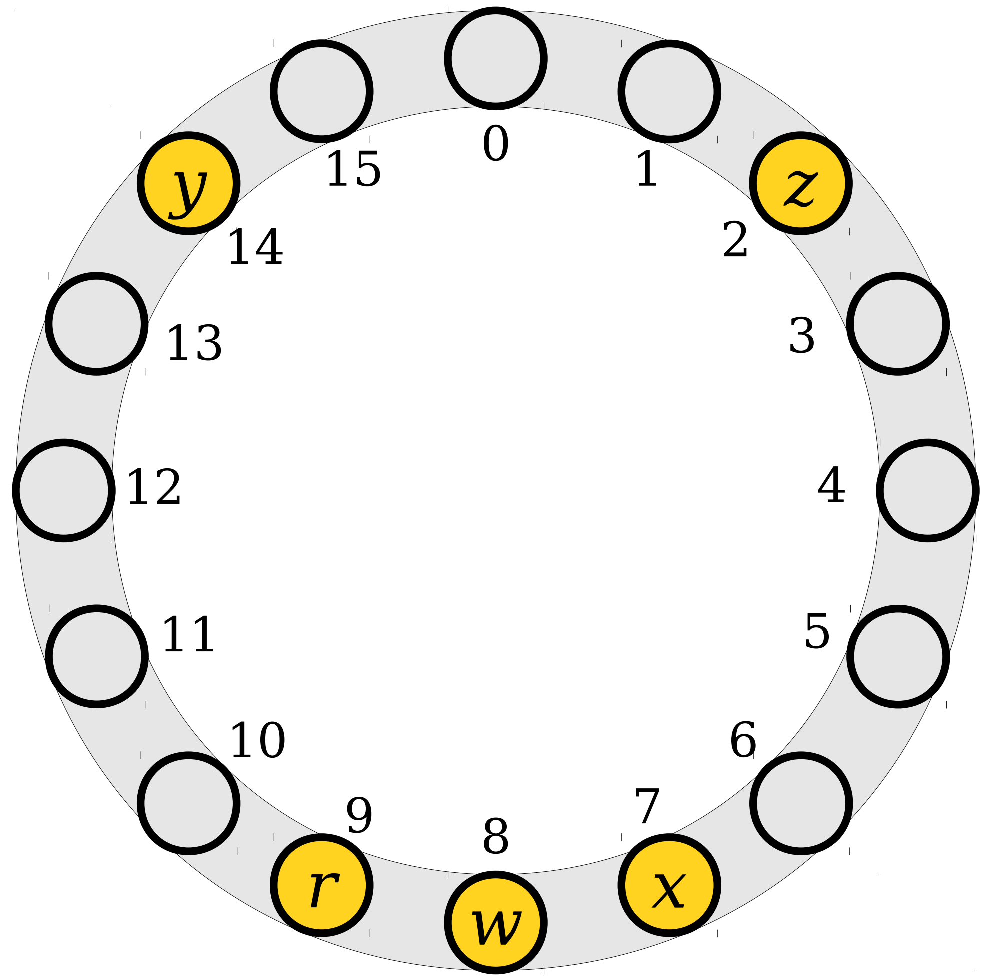
Linear Probing

- To look up an element x, compute $h(x)$ and start looking there.
- Move around the ring until either the element is found or a blank spot is detected.
- (We'll assume the load factor prohibits us from inserting so many elements that there are no free spaces.)



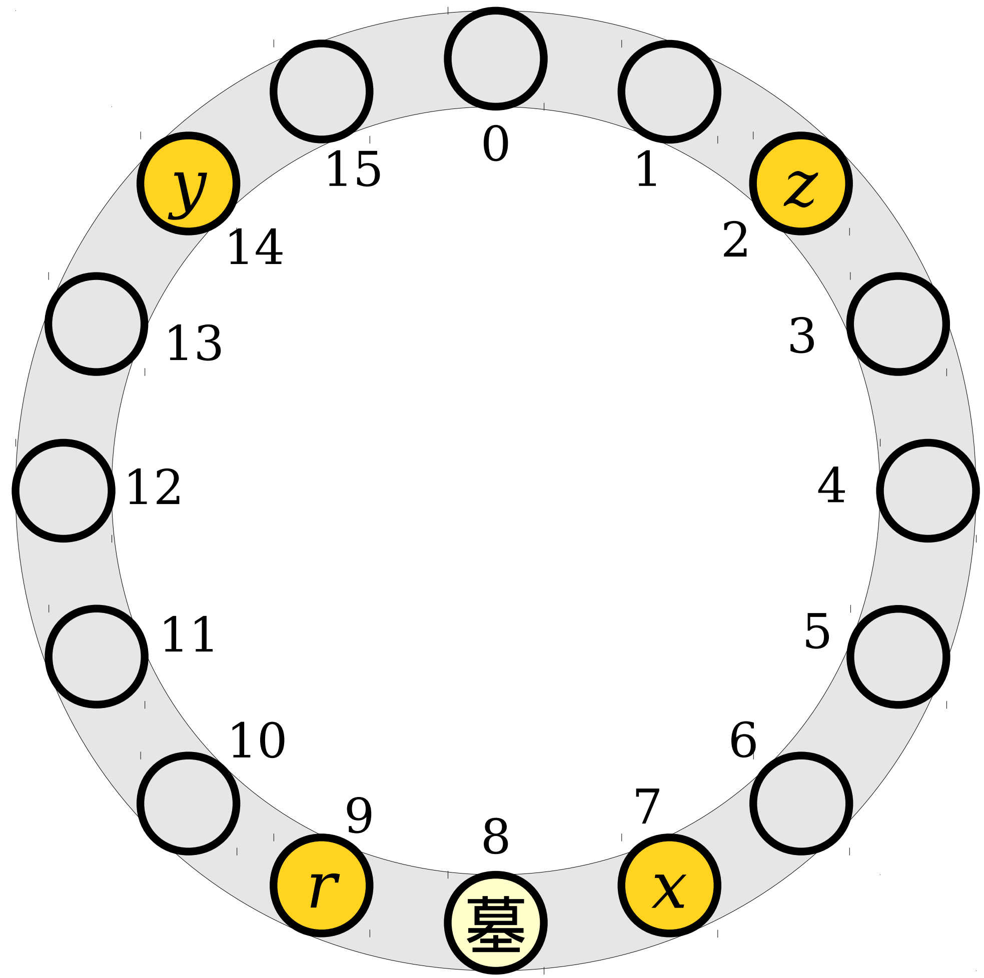
Linear Probing

- Deletions are a bit trickier than in chained hashing.
- We cannot just do a search and remove the element where we find it.
- Why?



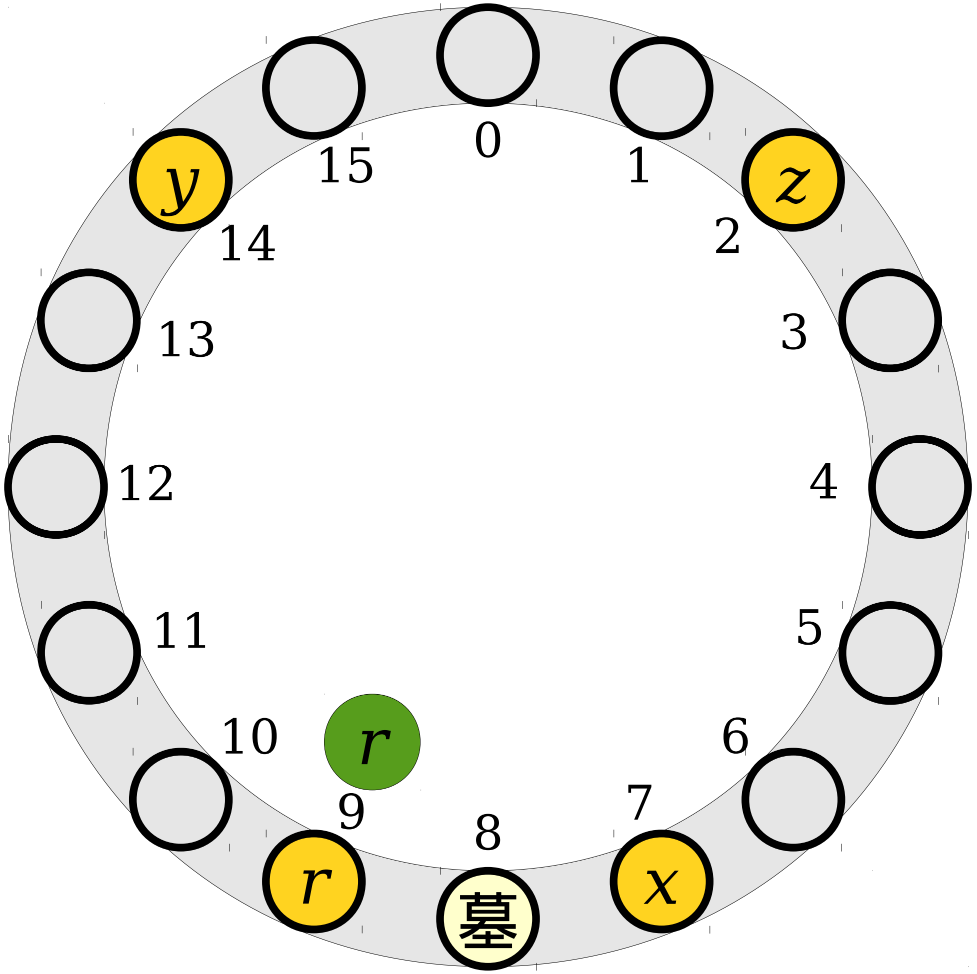
Linear Probing

- Deletions are often implemented using ***tombstones***.
- When removing an element, mark that the cell is empty and was previously occupied.
- When doing a lookup, don't stop at a tombstone. Instead, keep the search going.
- If there are “too many” tombstones, rebuild the table from scratch.



Linear Probing

- Deletions are often implemented using ***tombstones***.
- When removing an element, mark that the cell is empty and was previously occupied.
- When doing a lookup, don't stop at a tombstone. Instead, keep the search going.
- If there are “too many” tombstones, rebuild the table from scratch.

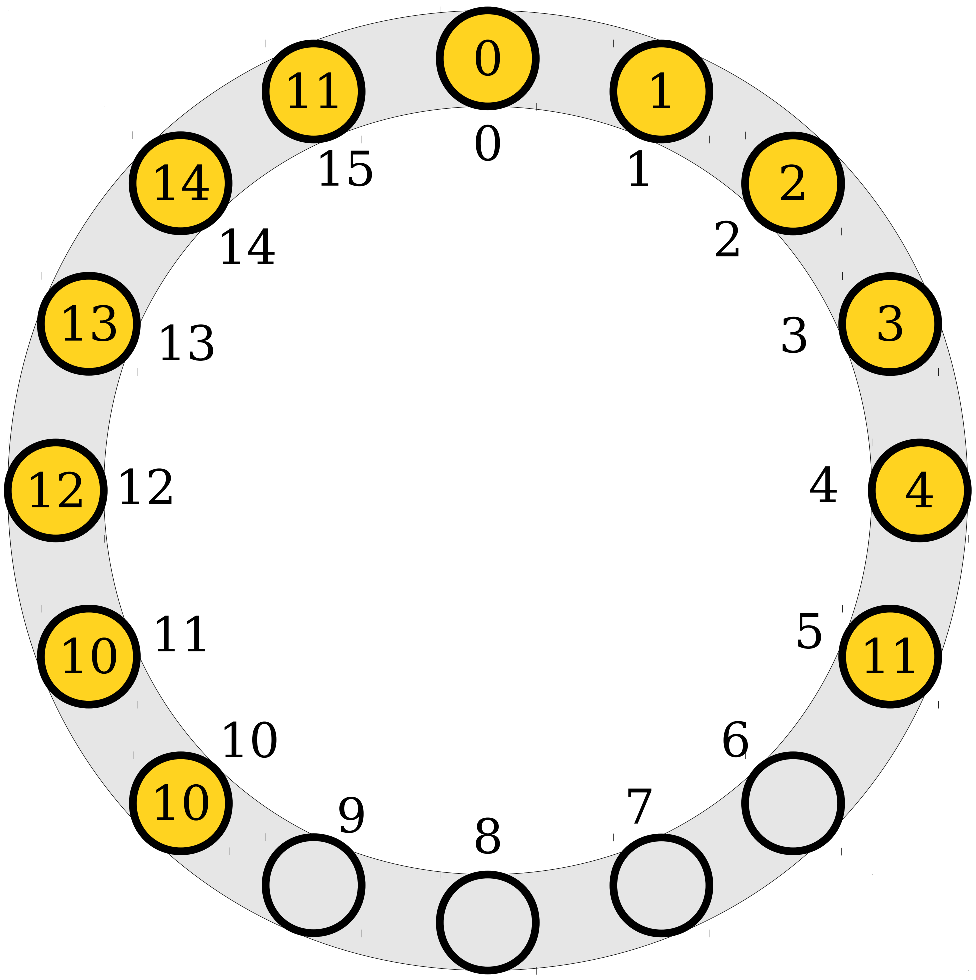


Linear Probing in Practice

- In practice, linear probing is one of the fastest general-purpose hashing strategies available.
- This is surprising – it was originally invented in 1954! It's amazing that it still holds up so well.
- Why is this?
  - ***Low memory overhead:*** just need an array and a hash function.
  - ***Excellent locality:*** when collisions occur, we only search in adjacent locations in the array.
  - ***Great cache performance:*** a combination of the above two factors.

The Weakness

- Linear probing exhibits severe performance degradations when the load factor gets high.
- The number of collisions tends to grow as a function of the number of existing collisions.
- This is called ***primary clustering***.



So… how fast is linear probing?

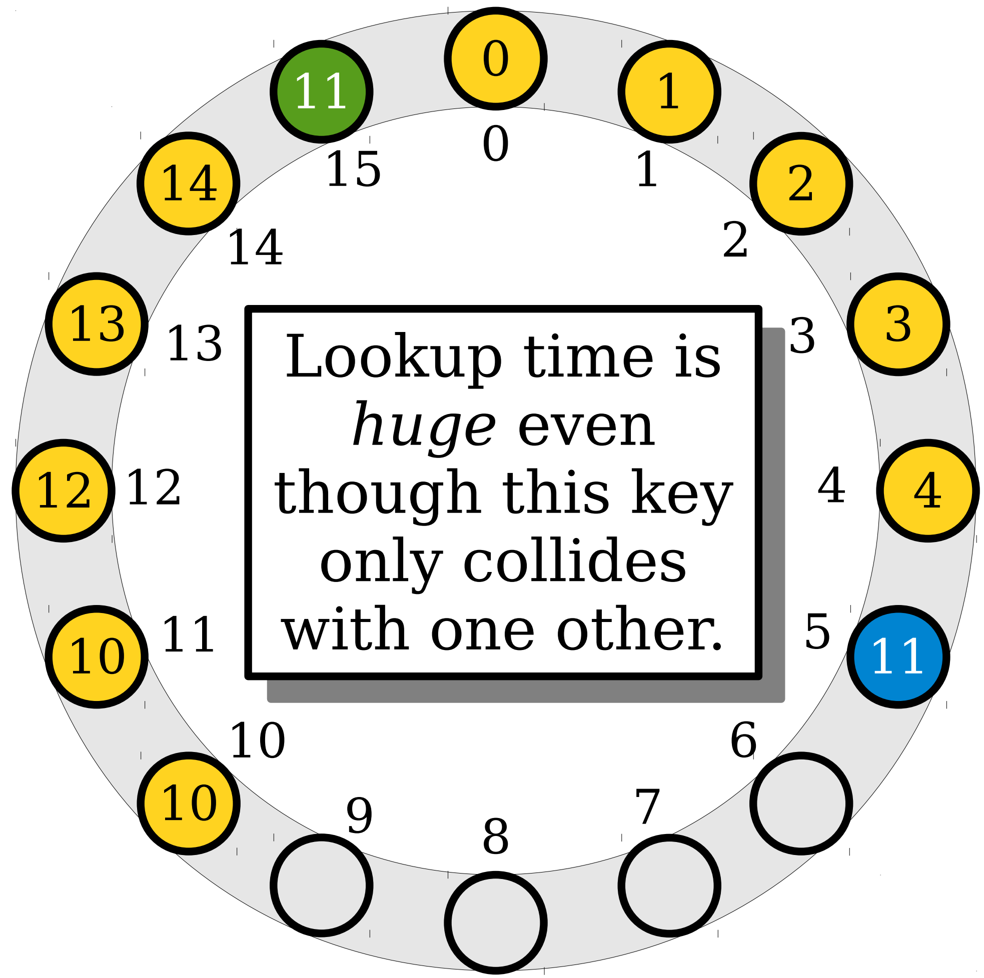
Analyzing Linear Probing

You probably saw an analysis of chained hash tables in CS161.

What makes linear probing different, interesting, or noteworthy?

Why Linear Probing is Different

- In chained hashing, collisions only occur when two values have exactly the same hash code.
- In linear probing, collisions can occur between elements with entirely different hash codes.
- To analyze linear probing, we need to know more than just how many elements collide with us.



Where We're Going

- The key question we need to answer for linear probing is the following:

How likely is it that a consecutive span of slots in a linear probing table has "too many things" hashing to it?

- We're going to investigate this in the abstract, but these answers directly translate into runtime bounds for a linear probing table.
- Check Thorup's lecture notes for details.

Where We're Going

- 2-independent hashing is useful because it leads to a small number of direct collisions.
- We need more than this: specifically, we want to avoid having elements "bunch up" in certain areas.
- ***Key idea:*** Slight increases to the strengths of our hash functions can lead to marked improvements in our guarantees.

k-Independent Hashing

Distribution Property:
Each element should have an equal probability of being placed in each slot.

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

Independence Property:
Where one element is placed shouldn't impact where a second goes.

For any distinct $x, y \in U$ and random $h \in \mathcal{H}$, the values $h(x)$ and $h(y)$ are independent random variables.

A family of hash functions $\mathscr{H}$ is called ***2-independent*** if it satisfies the distribution and independence properties.

Distribution Property:
Each element should have an equal probability of being placed in each slot.

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

Independence Property:
Where k-1 elements are placed shouldn't impact where a kth goes.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

A family of hash functions $\mathscr{H}$ is called ***k-independent*** if it satisfies the distribution and independence properties.

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

Suppose we hash n items with a k-independent hash function. On expectation, how many will be in the first b slots of the table?

Let X_i indicate whether $0 \le h(x_i) < b$.

Let $Y = \sum_{i=1}^{n} X_i$.

$$
\begin{aligned}
E[Y] &= E\left[\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} E[X_i] \\
&= \sum_{i=1}^{n} \frac{b}{m} \\
&= b \cdot \frac{n}{m} \\
&= \alpha \cdot b
\end{aligned}
$$

$\alpha = n/m$ is the ***load factor*** of the table.

0	1	2	3	…	b-1	b	b+1	...	m-1

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

What's the probability at least $2\alpha b$ elements end are the first b slots with a hash family that's ***1-independent***?

Let X_i indicate whether $0 \le h(x_i) < b$.

Let $Y = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} & \Pr[Y \geq 2\alpha \cdot b] \\ = \; & \Pr[Y \geq 2\mathrm{E}[Y]] \\ \leq \; & \frac{\mathrm{E}[Y]}{2\mathrm{E}[Y]} \\ = \; & \frac{1}{2} \end{aligned}$$

The best tool we can use is Markov's inequality, since with 1-independence we can control **E[*Y*]** but not **Var[*Y*]**.

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

What's the probability at least $2\alpha b$ elements end are the first b slots with a hash family that's ***2-independent***?

Let X_i indicate whether $0 \le h(x_i) < b$.

Let $Y = \sum_{i=1}^{n} X_i$.

Intuition: 2-indep. lets us control Var[Y].

$$
\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} \mathrm{Var}[X_i] \\
&\le \sum_{i=1}^{n} \mathrm{E}[X_i^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[X_i] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \mathrm{E}[Y]
\end{aligned}
$$

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

What's the probability at least $2\alpha b$ elements end are the first b slots with a hash family that's ***2-independent***?

Let X_i indicate whether $0 \le h(x_i) < b$.

Let $Y = \sum_{i=1}^{n} X_i$.

$$
\begin{aligned}
& \Pr[Y \ge 2\alpha \cdot b] \\
= \; & \Pr[Y \ge 2\,\mathrm{E}[Y]] \\
= \; & \Pr[Y - \mathrm{E}[Y] \ge \mathrm{E}[Y]] \\
\le \; & \Pr[|Y - \mathrm{E}[Y]| \ge \mathrm{E}[Y]] \\
\le \; & \frac{\mathrm{Var}[Y]}{\mathrm{E}[Y]^2} \\
\le \; & \frac{\mathrm{E}[Y]}{\mathrm{E}[Y]^2} \\
= \; & \frac{1}{\mathrm{E}[Y]}
\end{aligned}
$$

Intuiting Independence

	… lets us control …	… so we use …	… to bound $\Pr[Y \geq 2E[Y]]$ at
1-indep.	$E[Y]$	Markov	$O(1) \cdot E[Y]^0$
2-indep.	$\mathrm{Var}[Y]$	Chebyshev	$O(1) \cdot E[Y]^{-1}$
3-indep.	??	??	??
4-indep.	??	??	??

Question: Can we generalize the idea of variance?

Central Moments

- The ***kth central moment*** of a random variable X is given by

$$\mathbf{E[(}\boldsymbol{X}\mathbf{ - E[}\boldsymbol{X}\mathbf{])}^{\boldsymbol{k}}\mathbf{]}$$

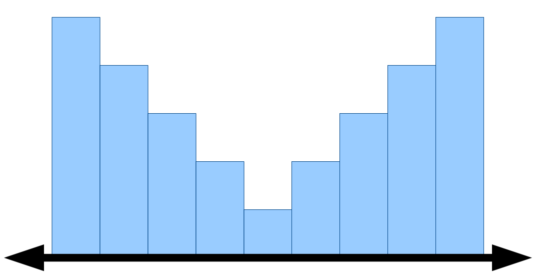
- What is the zeroth central moment of X?
  - ***Answer:*** 1.
- What is the first central moment of X?
  - ***Answer:*** 0.
- What is the second central moment of X?
  - ***Answer:*** Var[X].

Central Moments

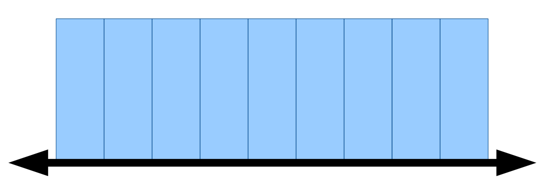
- The ***fourth central moment*** of a random variable X, denoted **$\mathbf{4th}[X]$**, is defined as

$$4\text{th}[X] = \text{E}[(X - \text{E}[X])^4].$$

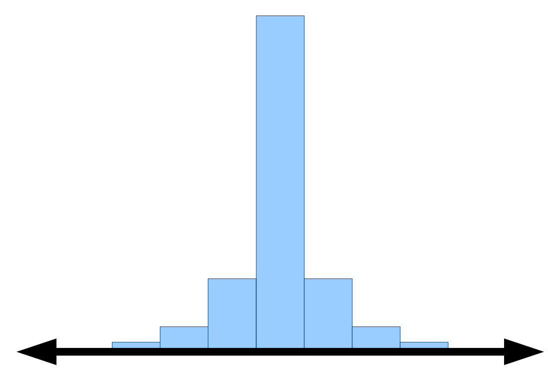
- ***Intuition:*** $4\text{th}[X]$ is similar to variance, but is significantly more sensitive to outliers.
- The actual values of $4\text{th}[X]$ and $\text{Var}[X]$ for the distributions here depend on scale. Assuming each bar has width one:



$4\text{th}[X] \gg \text{Var}[X]$



$4\text{th}[X] \gg \text{Var}[X]$



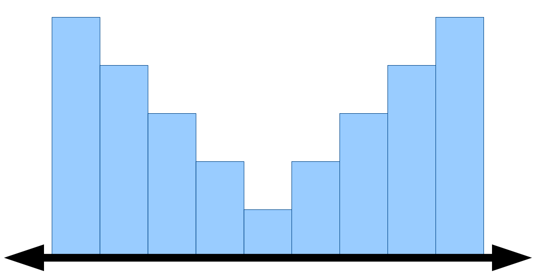
$4\text{th}[X] \ll \text{Var}[X]$

Central Moments

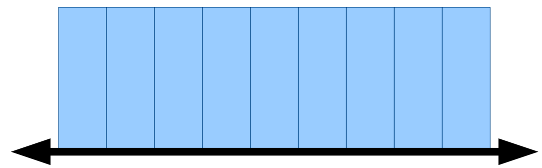
- The ***fourth central moment*** of a random variable X, denoted **4th[X]**, is defined as

$$\mathbf{4th}[X] = \mathbf{E}[(X - \mathbf{E}[X])^4].$$

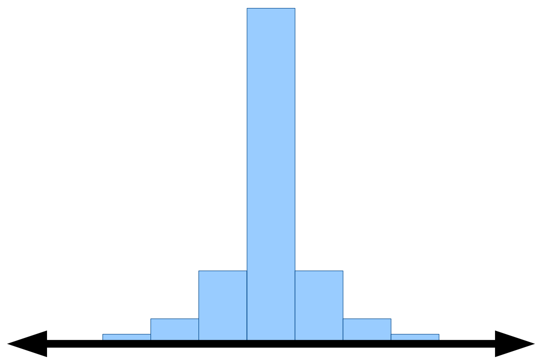
- ***Intuition:*** 4th[X] is similar to variance, but is significantly more sensitive to outliers.
- The actual values of 4th[X] and Var[X] for the distributions here depend on scale. Assuming each bar has width two:



$\text{4th}[X] \gg \text{Var}[X]$



$\text{4th}[X] \gg \text{Var}[X]$



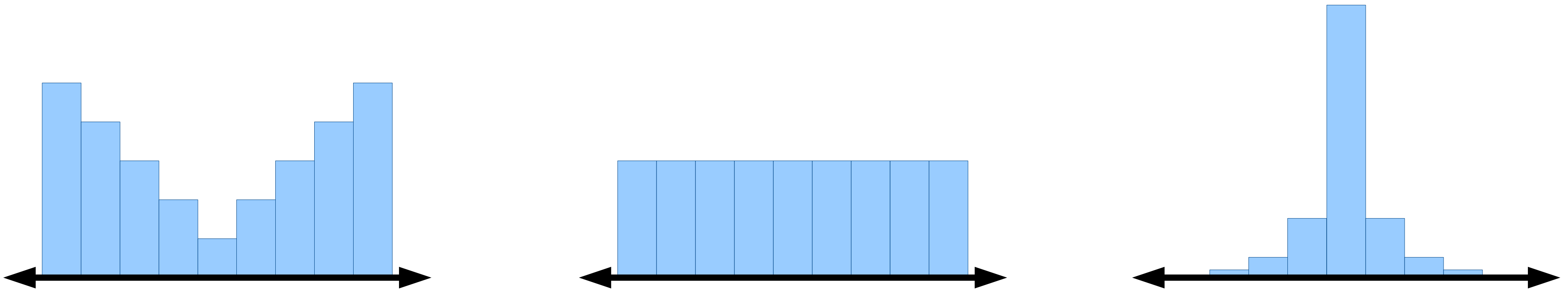
$\text{4th}[X] \gg \text{Var}[X]$

Central Moments

- Chebyshev's inequality says that

$$\mathbf{Pr[\ |X - E[X]| > c\ ]\ <\ Var[X] / c^2}.$$

- ***Intuition:*** The lower the variance of *X*, the less probability mass there is as you move away from the expected value.
- ***Question:*** Is there an analogy of Chebyshev's inequality for fourth moments?

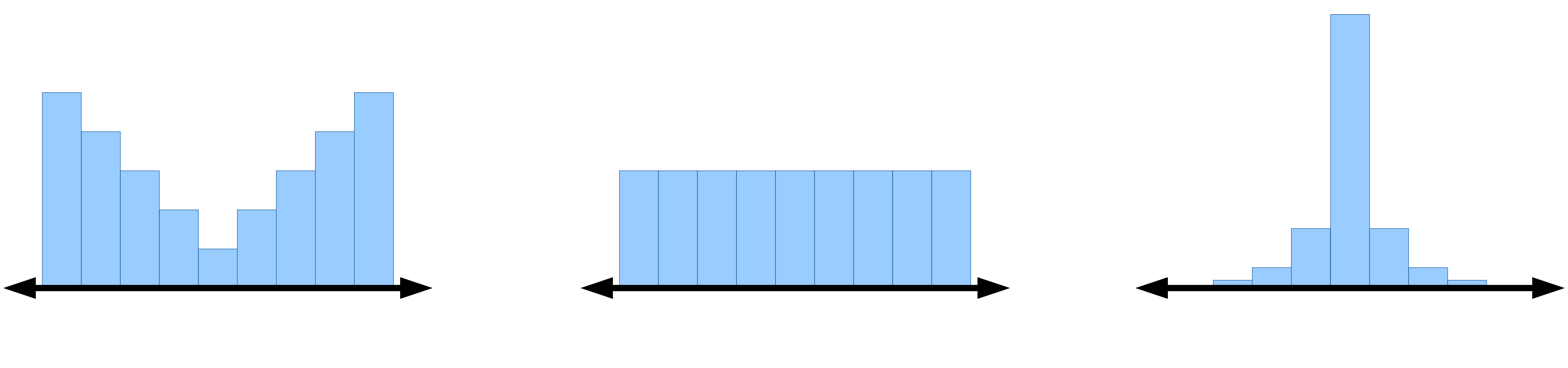


Central Moments

- The ***fourth moment inequality*** says that

$$\mathbf{Pr[\ |}\boldsymbol{X}\mathbf{\ -\ E[}\boldsymbol{X}\mathbf{]|\ >\ }\boldsymbol{c}\mathbf{\ ]\ \ <\ \ 4th[}\boldsymbol{X}\mathbf{]\ /\ }\boldsymbol{c}^{\mathbf{4}}.$$

- ***Intuition:*** The fourth moment of *X* is so sensitive to outliers that, if 4th[*X*] is low, it's extremely hard to move far from E[*X*].



Central Moments

- The ***fourth moment inequality*** says that

$$\mathbf{Pr[\ |X - E[X]| > c\ ]\ <\ 4th[X] / c^4}.$$

- ***Proof:*** Let X be a random variable. Then

$$\Pr[\ |X - E[X]| > c\ ] \ = \ \Pr[\ (X - E[X])^4 > c^4\ ].$$

Let $Y = (X - E[X])^4$. Notice that

$$E[Y] = E[(X - E[X])^4] = 4th[X],$$

so via Markov's inequality, we have

$$\begin{aligned} \Pr[\ |X - E[X]| > c\ ] \ &= \ \Pr[\ Y > c^4\ ] \\ &< \ E[Y] / c^4 \\ &= \ \mathbf{4th[X] / c^4}. \ \blacksquare \end{aligned}$$

Good question to ponder: why doesn't this work for the third central moment, where $3rd[X] = (X - E[X])^3$?

Intuiting Independence

	… lets us control	… so we can use
1-independence	$\mathrm{E}[Y]$	Markov's inequality
2-independence	$\mathrm{Var}[Y]$	Chebyshev's inequality
3-independence	??	??
4-independence	$\mathrm{4th}[Y]$	4^{th} moment inequality

For any $x \in U$ and random $h \in \mathcal{H}$, the value of $h(x)$ is uniform over $[m]$.

For any distinct $x_1, \ldots, x_k \in U$ and random $h \in \mathcal{H}$, the values $h(x_1), \ldots,$ and $h(x_k)$ are independent random variables.

What's the probability at least $2\alpha b$ elements end are the first b slots with a hash family that's ***4-independent***?

Let X_i indicate whether $0 \le h(x_i) < b$.

Let $Y = \sum_{i=1}^{n} X_i$.

Theorem:
$4\text{th}[Y] \le 4\mathrm{E}[Y]^2$.

$$
\begin{aligned}
& \Pr[Y \ge 2\alpha \cdot b] \\
= \; & \Pr[Y \ge 2\,\mathrm{E}[Y]] \\
\le \; & \Pr[|Y - \mathrm{E}[Y]| \ge \mathrm{E}[Y]] \\
\le \; & \frac{4\text{th}[Y]}{\mathrm{E}[Y]^4} \\
\le \; & \frac{4\,\mathrm{E}[Y]^2}{\mathrm{E}[Y]^4} \\
\le \; & \frac{4}{\mathrm{E}[Y]^2}
\end{aligned}
$$

To Summarize

	... lets us control	... so we use ...	... to bound $\Pr[Y \geq 2\mathrm{E}[Y]]$ at
1-indep.	$\mathrm{E}[Y]$	Markov	$O(1) \cdot \mathrm{E}[Y]^{0}$
2-indep.	$\mathrm{Var}[Y]$	Chebyshev	$O(1) \cdot \mathrm{E}[Y]^{-1}$
4-indep.	$4\mathrm{th}[Y]$	4^{th} Moment	$O(1) \cdot \mathrm{E}[Y]^{-2}$

Key intuition: Modest increases to the degree of independence lead to strong increases in the bounds of our error probabilities.

Theorem: The expected cost of a lookup in linear probing is

… **O(*n*)** if we use 2-independent hashing,

… **O(log *n*)** if we use 3-independent hashing, and

… **O(1)** if we use 5-independent hashing,

and these bounds can be made tight.

Proof idea: Imagine you know where some query hashes to. This uses up one degree of independence of the hash function. Define some regions near where the query lands. Then,

… if your hash function is 2-independent, you only have one degree of independence left. That gives weak (Markov) bounds on the odds that those regions have too many elements.

… if your hash function is 3-independent, you have two degrees of independence left. That gives modest (Chebyshev) bounds on the odds that those regions have too many elements.

… if your hash function is 5-independent, you have four degrees of independence left. That gives strong (fourth moment) bounds on the odds that those regions have too many elements.

The lower bounds come from technical adversarial arguments that are way above our pay grade. 😃 ■

Next Time

- ***Cuckoo Hashing***
  - Brood parasitism meets hashing.
- ***The Cuckoo Graph***
  - Random graphs for fun and profit.
- ***Subcritical Galton-Watson Processes***
  - Noble names and fast hashing.

Appendix: Bounding fourth moments of sums of indicator variables.

Or: How I learned to quit worrying and love the math.

Earlier, we claimed that

$$4\text{th}[Y] \leq 4\text{E}[Y].$$

Where does that result come from?

Step 1: Determine 4th[X_i], where X_i is one of our indicators from earlier.

Step 2: Determine 4th[Y], knowing that Y is the sum of all the X_i's.

Proceed slowly here.

The math involved isn't too tricky, but it does require some attention to detail.

Generalizing Indicator Variance

- ***Theorem:*** If X is an indicator variable for the event $\mathcal{E}$, then $4\text{th}[X] \leq \mathrm{E}[X]$.
- ***Proof:*** X takes on value 1 with probability $\Pr[\mathcal{E}]$ and 0 with probability $1 - \Pr[\mathcal{E}]$. Therefore, we have

$$\begin{aligned}
4\text{th}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^4] \\
&= (1 - \Pr[\mathcal{E}])^4 \cdot \Pr[\mathcal{E}] + \Pr[\mathcal{E}]^4(1 - \Pr[\mathcal{E}]) \\
&\leq (1 - \Pr[\mathcal{E}])^3 \cdot \Pr[\mathcal{E}] + \Pr[\mathcal{E}]^4 \\
&= \Pr[\mathcal{E}] - \Pr[\mathcal{E}]^4 + \Pr[\mathcal{E}]^4 \\
&= \Pr[\mathcal{E}] \\
&= \mathrm{E}[X]. \blacksquare
\end{aligned}$$

The Limits of Our Generalization

- There's a lovely little expression for $\mathrm{Var}[X]$:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

- That's because

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2 - 2X \cdot \mathrm{E}[X] + \mathrm{E}[X]^2] \\
&= \mathrm{E}[X^2] - 2\mathrm{E}[X] \cdot \mathrm{E}[X] + \mathrm{E}[X]^2 \\
&= \mathrm{E}[X^2] - 2\mathrm{E}[X]^2 + \mathrm{E}[X]^2 \\
&= \mathbf{E}[X^2] - \mathbf{E}[X]^2.
\end{aligned}$$

- We can try this for fourth moments, but, well, um…

$$\begin{aligned}
\mathrm{4th}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^4] \\
&= \mathrm{E}[X^4 - 4X^3 \cdot \mathrm{E}[X] + 6X^2 \cdot \mathrm{E}[X]^2 - 4X \cdot \mathrm{E}[X]^3 + \mathrm{E}[X]^4] \\
&= \mathrm{E}[X^4] - 4\mathrm{E}[X] \cdot \mathrm{E}[X^3] + 6\mathrm{E}[X]^2\mathrm{E}[X^2] - 4\mathrm{E}[X] \cdot \mathrm{E}[X]^3 + \mathrm{E}[X^4] \\
&= \mathrm{E}[X^4] - 4\mathrm{E}[X] \cdot \mathrm{E}[X^3] + 6\mathrm{E}[X]^2\mathrm{E}[X^2] - 3\mathrm{E}[X]^4
\end{aligned}$$

= **¯_(ツ)_/¯**

Looks like we'll need to compute 4th[Y] directly from the definition.

$$
\begin{aligned}
& \text{4th}[Y] \\
= \;& \mathrm{E}[(Y - \mathrm{E}[Y])^4] \\
= \;& \mathrm{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}[X_i]\right)^4\right] \\
= \;& \mathrm{E}\left[\left(\sum_{i=1}^{n} (X_i - \mathrm{E}[X_i])\right)^4\right] \\
= \;& \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} (X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])\right] \\
= \;& \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_{ls} - \mathrm{E}[X_l])]
\end{aligned}
$$

We "just" need to simplify this last expression.

Exploring this Summation

- The terms of this summation might sometimes range over the same variables at the same time:

$$4\text{th}[Y] = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])]$$

- ***Claim:*** Any term in the above summation where X_i is a different random variable than X_j, X_k, and X_l is zero.
- ***Proof:*** Suppose that X_i is a different random variable from the others. Then since X_i, X_j, X_k, and X_l are independent, we have

$$\begin{aligned} & \mathrm{E}[\ (X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])\ ] \\ = & \ \mathrm{E}[X_i - \mathrm{E}[X_i]] \cdot \mathrm{E}[(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])\ ] \\ = & \ 0 \cdot \mathrm{E}[(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])\ ] \\ = & \ \mathbf{0} \end{aligned}$$

Exploring this Summation

- The terms of this summation might sometimes range over the same variables at the same time:

$$4\text{th}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])]$$

- ***Claim:*** Every term in this sum is zero *except* for the following:
    - Terms where $i = j = k = l$.
    - Terms where two of i, j, k, and l refer to one value and the other two of i, j, k, and l refer to another.
- ***Proof:*** If a variable appears exactly one time, then by our previous logic the term evaluates to zero. If a variable appears exactly three times, then the other variable appears exactly once and the term evaluates to zero. That leaves behind the two remaining cases here.

Exploring this Summation

- The terms of this summation might sometimes range over the same variables at the same time:

$$4\text{th}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])]$$

- ***Claim:*** Every term in this sum is zero *except* for the following:
  - Terms where $i = j = k = l$.
  - Terms where two of i, j, k, and l refer to one value and the other two of i, j, k, and l refer to another.

$$\sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4]$$

Exploring this Summation

- The terms of this summation might sometimes range over the same variables at the same time:

$$4\text{th}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])]$$

- ***Claim:*** Every term in this sum is zero *except* for the following:
  - Terms where $i = j = k = l$.
  - Terms where two of $i, j, k,$ and l refer to one value and the other two of $i, j, k,$ and l refer to another.

$$\sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{p=1}^{n} \sum_{q=p+1}^{n} \mathrm{E}[(X_p - \mathrm{E}[X_p])^2 (X_q - \mathrm{E}[X_q])^2]$$

Which of $i, j, k,$ and l refer to the first value?

What's the first value?

What's the second? (It must be different than the first!)

Exploring this Summation

- The terms of this summation might sometimes range over the same variables at the same time:

$$4\text{th}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])(X_k - \mathrm{E}[X_k])(X_l - \mathrm{E}[X_l])]$$

- ***Claim:*** Every term in this sum is zero *except* for the following:
  - Terms where $i = j = k = l$.
  - Terms where two of i, j, k, and l refer to one value and the other two of i, j, k, and l refer to another.

$$\sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2]$$

We'll use *i* and *j* as our summation variables, since that's easier to read.

$$
\begin{aligned}
4\mathrm{th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2]\,\mathrm{E}[(X_j - \mathrm{E}[X_j])^2]
\end{aligned}
$$

Since h is 4-independent, these are independent random variables.

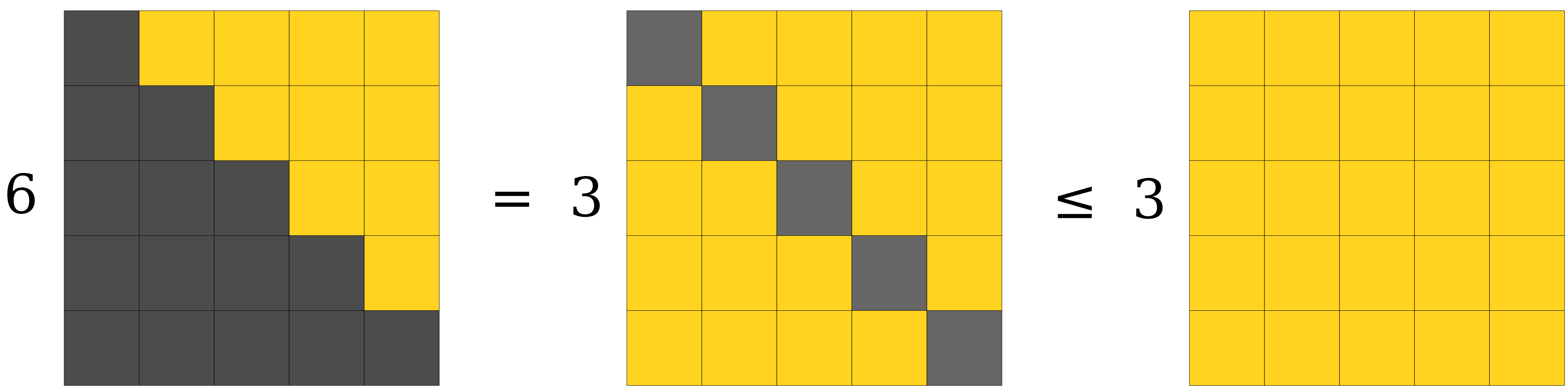
$$
\begin{aligned}
\mathrm{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2] \mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Var}[X_i] \mathrm{Var}[X_j]
\end{aligned}
$$

This is the definition of the fourth central moment.

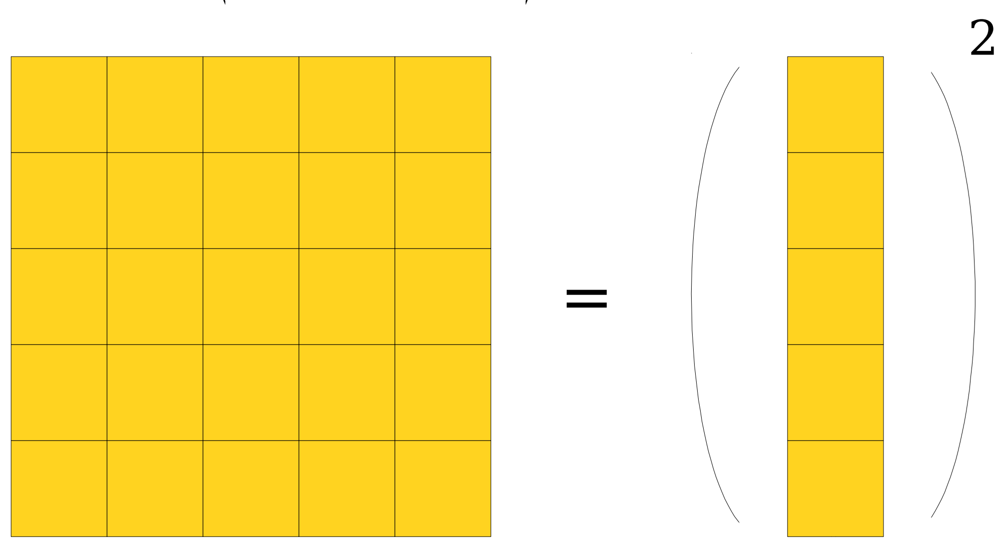
This is the definition of variance.

So is this.

$$
\begin{aligned}
\mathrm{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2]\mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&\leq \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j]
\end{aligned}
$$



$$
\begin{aligned}
\text{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2]\mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \text{4th}[X_i] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&\leq \sum_{i=1}^{n} \text{4th}[X_i] + 3 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&= \sum_{i=1}^{n} \text{4th}[X_i] + 3 \left(\sum_{i=1}^{n} \mathrm{Var}[X_i] \right)^2
\end{aligned}
$$



$$
\begin{aligned}
\mathrm{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2]\mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&\leq \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3 \left(\sum_{i=1}^{n} \mathrm{Var}[X_i] \right)^2 \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3\mathrm{Var}[Y]^2
\end{aligned}
$$

$$
\boxed{\sum_{i=1}^{n} \mathrm{Var}[X_i] = \mathrm{Var}[\sum_{i=1}^{n} X_i] = \mathrm{Var}[Y]}
$$

$$
\begin{aligned}
\mathrm{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2] \mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&\leq \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3 \left(\sum_{i=1}^{n} \mathrm{Var}[X_i] \right)^2 \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3\mathrm{Var}[Y]^2 \\
&\leq \sum_{i=1}^{n} \mathrm{E}[X_i] + 3\mathrm{E}[Y]^2
\end{aligned}
$$

If X is an indicator, then $\mathrm{4th}[X] \leq \mathrm{E}[X]$.

We know from our 2-independence analysis that $\mathrm{Var}[Y] \leq \mathrm{E}[Y]$

$$
\begin{aligned}
\mathrm{4th}[Y] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_{js} - \mathrm{E}[X_j])(X_k - E[X_k])(X_l - E[X_l])] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_{is} - \mathrm{E}[X_i])^4] + \binom{4}{2} \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2 (X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^4] + 6 \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{E}[(X_i - \mathrm{E}[X_i])^2]\mathrm{E}[(X_j - \mathrm{E}[X_j])^2] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 6\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&\le \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Var}[X_i]\mathrm{Var}[X_j] \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3\left(\sum_{i=1}^{n} \mathrm{Var}[X_i]\right)^2 \\
&= \sum_{i=1}^{n} \mathrm{4th}[X_i] + 3\mathrm{Var}[Y]^2 \\
&\le \sum_{i=1}^{n} \mathrm{E}[X_i] + 3\mathrm{E}[Y]^2 \\
&= \mathrm{E}[Y] + 3\mathrm{E}[Y]^2 \\
&\le \mathbf{4E[Y]^2}
\end{aligned}
$$

(As long as E[Y] ≥ 1, which we can assume if we're talking about sufficiently large sums.)