Advanced Algorithms

Lecture 16: Balls and bins

Announcements

- **HW 4 out — due next Friday** (Graphs & shortest paths).
- Mid-term poll: discussion
- Office hrs today — 1 PM → 2 PM.

Randomized algorithms: analysis

- In some cases: trade-off between running time & probability of correctness (e.g., identity testing, …)

- Las Vegas algorithms: always correct, but running time can sometimes be large (e.g., quick sort)

(running times are random variables).

Expected running time

- Run time is a random variable — e.g., quick sort
  - Choose random number of the array as the “pivot”, divide array into two parts, sort recursively
    - As long as split is “roughly balanced”, problem size reduces significantly
- Can write a **recurrence** for the expected running time

$f(n)$: expected runtime on array of size n.

- Law of “conditional expectation”

$$\mathbb{E}[X] = p(F) \cdot \mathbb{E}[X|F] + (1 - p(F)) \cdot \mathbb{E}[X|\overline{F}]$$

$\hookrightarrow \int x \Pr[X = x] \, dx$

Expectation good enough?

- Suppose that the expected running time is O(n log n)
- Can we upper bound probability that it is (say) n^2 ?
- What about $n^{1.5}$?

$$Pr\{\text{runtime} > n^{1.5}\} < \frac{1}{100} \; ?$$

General result

Markov's inequality: let X be a non-negative random variable with expectation C. Then prob[X > tC] <= 1/t. $\forall\, t \geq 1$

$\Pr[X > 10 \cdot C] < \frac{1}{10}$ ..

- Implication for quick sort?

- "Boosting probability"

what if we want success prob. of 0.9999

$\rightsquigarrow$ $C = 4n\log n$

X: running time of quicksort on instance of size n.

$$\Pr\left[X > 10 \cdot 4n\log n\right] \leq \frac{1}{10}.$$

[w.p. ≥ 0.5, running time $\leq 2 \cdot C$.

What is a bound on $\text{Prob}\left[X \geq \frac{n^2}{2}\right]$.

$$\mathbb{E}[X] := C = n \log n$$

$$t = \frac{n}{2 \log n}$$

$$\Pr\left[X \geq t \cdot C\right] \leq \frac{1}{t} \implies \Pr\left[X \geq \frac{n^2}{2}\right] \leq \frac{2 \log n}{n}$$

What is a bound on $\Pr\left[X > 2n\right]$? → no meaningful bound.

Goal: get much higher success prob.

Amplification by repetition

"Basic" Markov: $\text{Prob}[X > 2 \cdot n \log n] \leq \frac{1}{2}$.

→ Impatient quicksort:

- try 100 times:
  - run quicksort for $2n \log n$ steps
  - if it doesn't complete, terminate
  - if it completes, break;

prob. $\leq \frac{1}{2}$

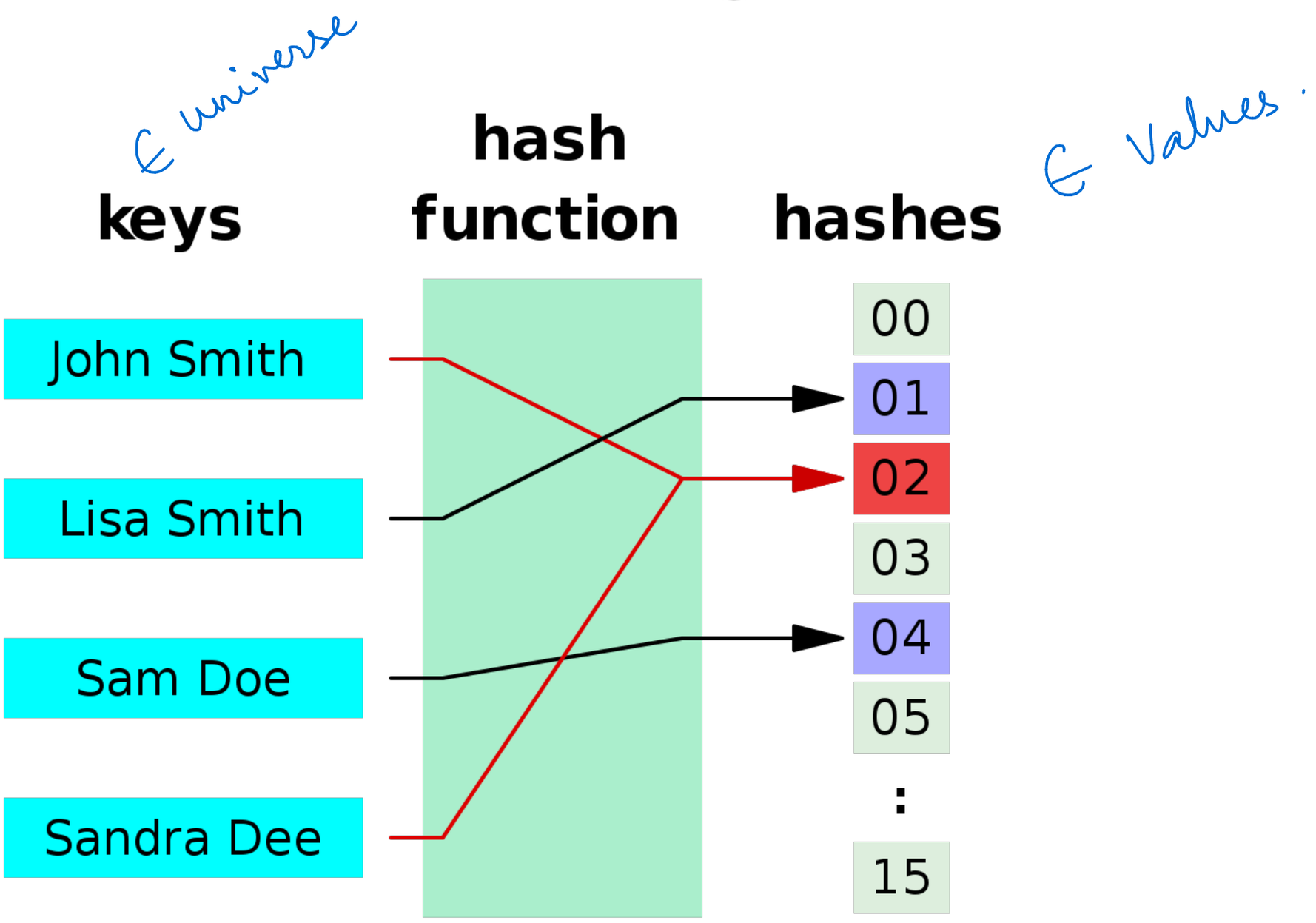
total running time = $200 \cdot n \log n$

Prob that we fail all the time $\leq \frac{1}{2^{100}}$

Today's plan

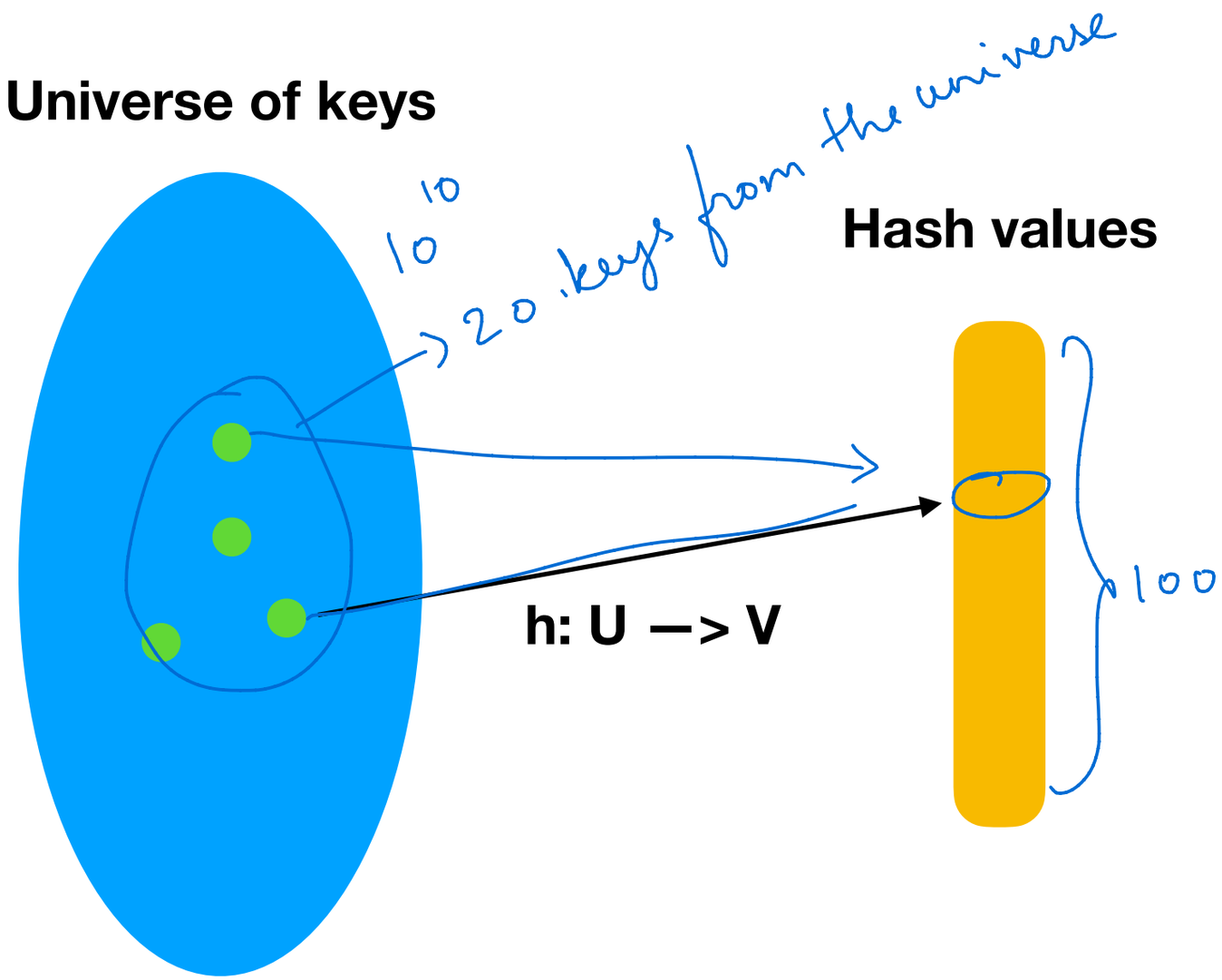
- Understanding hashing via "balls and bins"
- Linearity of expectation

Hashing



(Src: wikipedia)

Hashing



What one hopes:

- Don't know keys beforehand
- Small set of hash values (~ # of keys)
- Few "collisions"

Designing hash functions can be tricky…

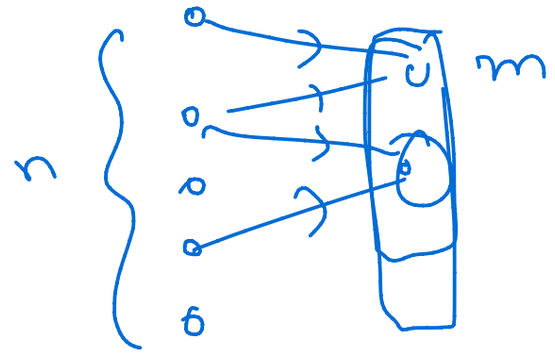
Balls and bins

Some questions

$B_1, B_2, \ldots, B_m$

Problem: suppose we have *n* balls and *m* bins. Imagine throwing the balls into bins, independently and uniformly at random.

n/m

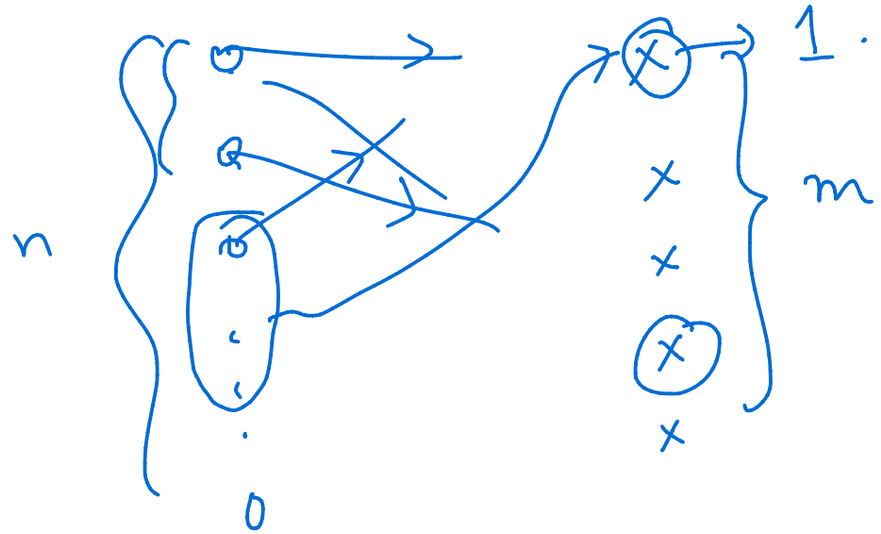


- What is the expected size of each bin?
- Suppose $n = m$; What is the expected number of bins with exactly 4 balls?

$c \cdot n$
$\log n$
$\sqrt{n}$
1

- Suppose $n = m$; What is the probability that there exists a bin with (log n) balls?

Expected size of ~~a~~ bin 1.



Definition of expectation:

B: random variable which is the # balls landing in bin 1.

$$\mathbb{E}[B] = \sum_{k=1}^{n} k \cdot \Pr[B = k].$$

Expected size of a bin

Definition of expectation:

Moral from last week: never compute expectations using the definition!

Linearity of expectation

→ Let X, Y be any two random variables (✱ they could be dependent). Then

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].$$

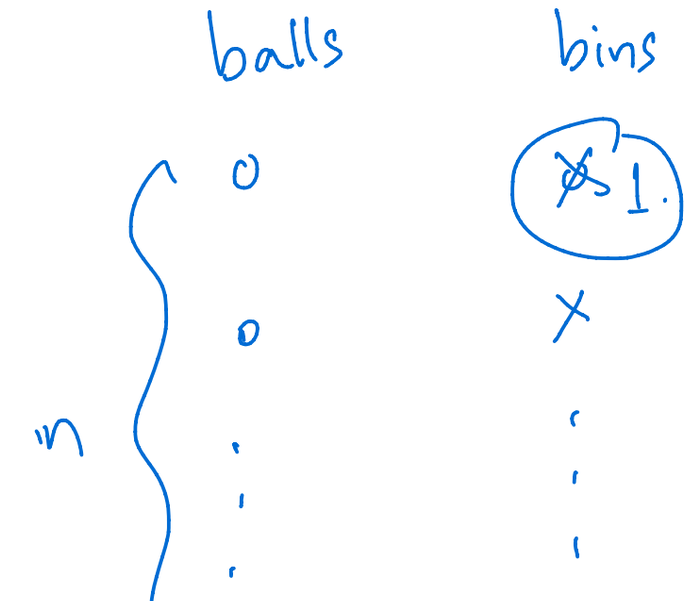
$$\mathbb{E}[XY] \stackrel{?}{=} \mathbb{E}[X] \cdot \mathbb{E}[Y] \; ?$$

Main "use": We can often write random variables as sum of "simpler" r.v.s.

Expected size of a bin

"Decomposing" into a sum of random variables...

$B_i \equiv$ random variable that is the # of balls in bin i



$$B_i = n - \sum_{j \neq i} B_j$$

$$\mathbb{E}[B_i] = n - \sum_{j \neq i} \mathbb{E}[B_j]$$

$\sum_j \mathbb{E}[B_j] = n$ $\rightsquigarrow$ using symmetry, all $\mathbb{E}[B_j]$ vals are equal; $\Rightarrow \mathbb{E}[B_j] = \frac{n}{m}$.

- Take bin 1 ; let B be the r.v. denoting # of balls in it.

- define $X_1 = \begin{cases} 1 & \text{if ball 1 went to bin 1} \\ 0 & \text{o/wise.} \end{cases}$

$X_2 = \begin{cases} 1 & \text{if ball 2 } \dots \text{ 1} \\ 0 & \text{o/wise.} \end{cases}$

by defn, $B = X_1 + X_2 + \dots + X_n \longrightarrow \star$.

What is $\mathbb{E}[X_i]$? $= 0 \cdot \Pr[X_i = 0] + 1 \cdot \Pr[X_i = 1] = \frac{1}{m}$.

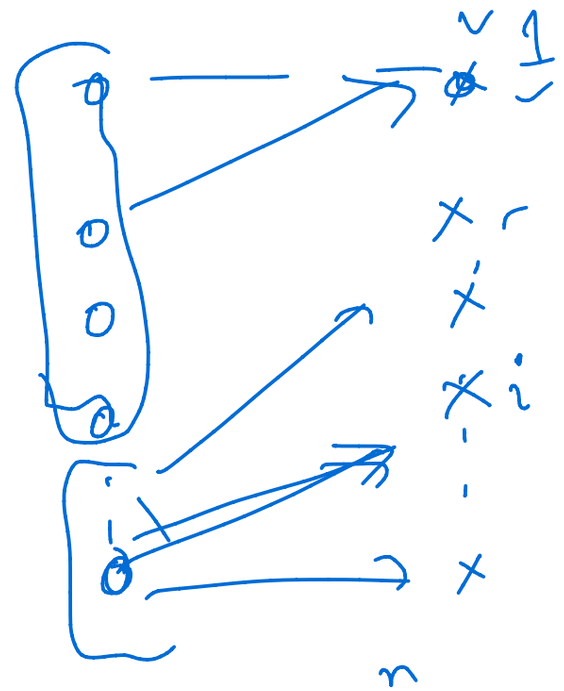
Using linearity of exp. at $\star$, we get $\mathbb{E}[B] = \frac{n}{m}$.

$(n = m)$.

Expected # of bins with 4 balls

Y: # of bins with precisely 4 balls

$$X_i = \begin{cases} 1 & \text{if bin } i \text{ receives precisely 4 balls} \\ 0 & \text{o/wise.} \end{cases}$$



by definition, $Y = X_1 + X_2 + \cdots + X_n \Rightarrow \mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i]$.

what is $\mathbb{E}[X_i]$? $= \Pr[X_i = 1] = \binom{n}{4} \cdot \frac{1}{n^4} \cdot \left(\frac{n-1}{n}\right)^{n-4}$

$$\mathbb{E}[Y] \approx \frac{n}{80} = \frac{n(n-1)(n-2)(n-3)}{4!} \cdot \frac{(n-1)^{n-4}}{n^{n-4} \cdot n^4} \approx \frac{1}{e} \cdot \frac{1}{24} \approx \frac{1}{80}.$$

Expected # of bins with 4 balls

Some questions

Problem: suppose we have *n* balls and *m* bins. Imagine throwing the balls into bins, independently and uniformly at random.

- What is the expected size of each bin?
- Suppose $n = m$; What is the expected number of bins with exactly 4 balls?
- Suppose $n = m$; What is the probability that there exists a bin with (log n) balls?

The union bound

The union bound