Advanced Algorithms

Lecture 17: Balls and bins (contd.), sampling

Announcements

- **HW 4 due next Friday**

Last lecture

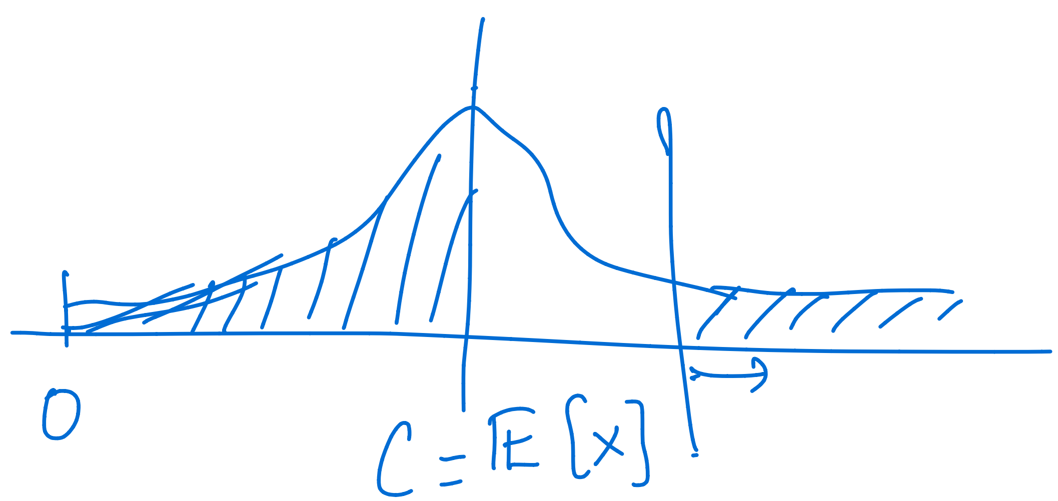
$\Pr\left[\begin{array}{c}\text{random} \\ \text{var is} > t \cdot \mathbb{E}\end{array}\right] \le \frac{1}{t}$

- Markov's inequality and why "expectation" analysis often suffices
- Hashing, throwing balls into bins
- Key analysis methods:
  - define appropriate random variables
  - linearity of expectation
  - expectations easy to compute if r.v.s are "binary"

Recap: Markov's inequality

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

- <u>Note:</u> Markov's inequality does not give any bounds on ~~whether~~ the prob. that r.v. ~~can be~~ is **much smaller** than expectation



Today's plan

- Answer final question about balls-and-bins
- Union bound
- More comments on hashing
- Sampling — estimation, variance

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? $(\overset{1}{\overset{\shortparallel}{m/n}})$ (when $m = n$).

- Suppose $n = m$; What is the expected number of bins with exactly 4 balls? [~ n/(24e) ~ n/65]

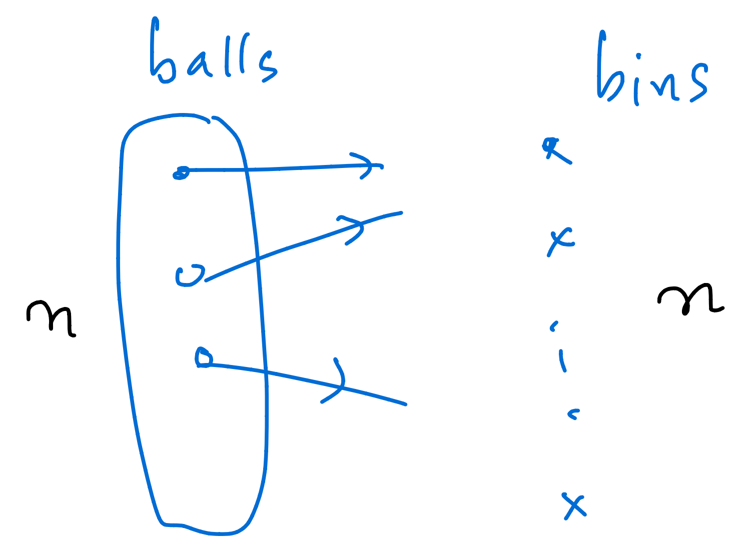
- Suppose $n = m$; What is the probability that there exists a bin with (log n) balls?

$$\leq \frac{1}{n}.$$

Number of bins with log n balls

$X \equiv$

$\Pr[X \geq 1]$? $= 1 - \Pr[X = 0]$



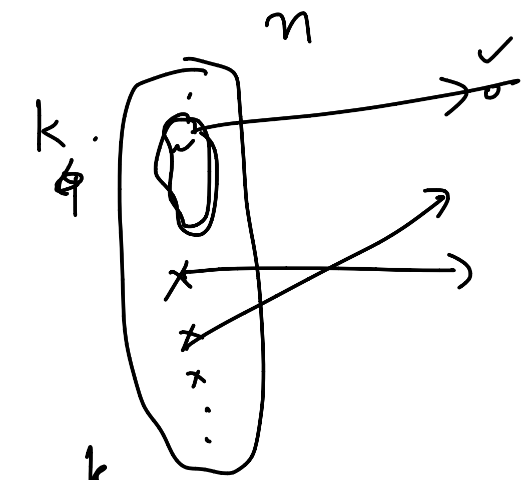
$Y_i :=$ r. v. that indicates if bin i receives $\log n$ balls.

By defn, $X = Y_1 + Y_2 + \cdots + Y_n$.

$$\Pr[X \geq 1] = \Pr\left[(Y_1 = 1) \underbrace{\vee}_{= \text{OR}} (Y_2 = 1) \vee \cdots \vee (Y_n = 1)\right].$$

$$\leq \Pr[(Y_1 = 1)] + \Pr[Y_2 = 1] + \cdots + \Pr[Y_n = 1]$$

P[one bin having *k* balls]



What is $\Pr[Y_1 = 1]$?

$\parallel$

$$\binom{n}{k}\cdot\left(\frac{1}{n}\right)^{k}\underline{\left(\frac{n-1}{n}\right)^{n-k}} \sim \le \left(\frac{e}{k}\right)^{k}$$

Stirling approximation: $\binom{n}{k} \le \left(\frac{n\cdot e}{k}\right)^{k}$

(Conseq. of)

$$\le \left(\frac{\cancel{n}\, e}{k}\right)^{k} \cdot \boxed{\frac{(n-1)^{n-k}}{n^{n-k}}} \rightsquigarrow \left(\frac{2.718}{k}\right)^{k} \rightsquigarrow \left(\frac{3}{100}\right)^{100} \ll 1$$

$\approx \frac{1}{e}$

$k=100$

What if $k = \log n$?

$\approx \frac{1}{n}$ when $k = \frac{\log n}{\log \log n}$.

$$\left(\frac{e}{k}\right)^k = \left(\frac{e}{\log n}\right)^{\log n}$$

For large enough n, $\frac{e}{\log n} < \frac{1}{10}$

$$\left(\frac{e}{\log n}\right)^{\log n} < \left(\frac{1}{10}\right)^{\log n} < \frac{1}{n^2}.$$

$$\Pr[Y_1 = 1] < \frac{1}{n^2} \Rightarrow \Pr[Y_1=1] + \Pr[Y_2=1] + \dots + \Pr[Y_n=1]$$

$$\le n \cdot \frac{1}{n^2} \le \frac{1}{n}.$$

$$X = Y_1 + Y_2 + \cdots + Y_n .$$

$$\Pr[Y_1 = 1] \leq \frac{1}{n^2} \Rightarrow \mathbb{E}[Y_1] \leq \frac{1}{n^2}$$

$$\Rightarrow \mathbb{E}[X] \leq \frac{1}{n} .$$

$$\Pr\left[X \geq n \cdot \mathbb{E}[X]\right] \leq \frac{1}{n} .$$

$$\Downarrow$$

$$\Pr[X \geq 1] \leq \frac{1}{n} .$$

The union bound

Suppose $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ are events in some prob. space. Then $\Pr[\underline{\underline{\mathcal{E}_1}} \vee \underline{\underline{\mathcal{E}_2}} \vee \ldots \vee \mathcal{E}_n] \leq \sum_{i=1}^{n} \Pr[\mathcal{E}_i]$.

→ Becomes equality iff $\mathcal{E}_i$ are all <u>disjoint</u>. (ie, $\Pr[\mathcal{E}_i \wedge \mathcal{E}_j] = 0$). for all i,j.

$n = 3$

→ <u>Inclusion/Exclusion formula</u>: (for all $\mathcal{E}_i$).

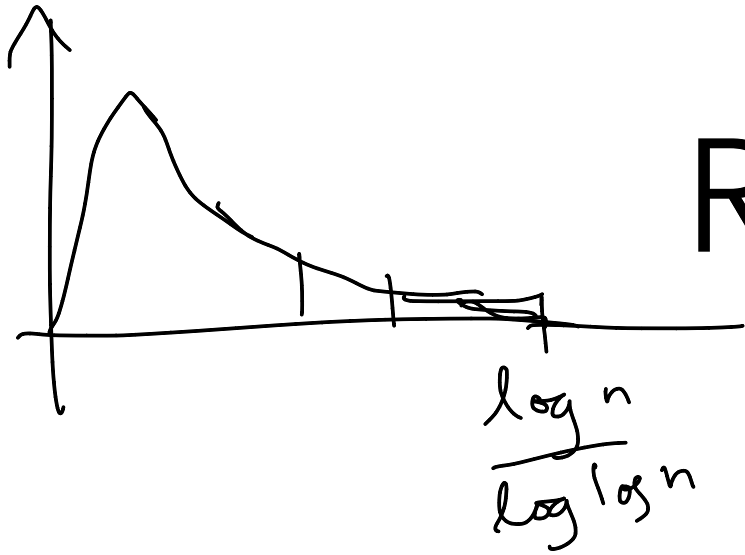
$$\Pr[\mathcal{E}_1 \vee \mathcal{E}_2 \vee \mathcal{E}_3] = \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] + \Pr[\mathcal{E}_3] - \Pr[\mathcal{E}_1 \wedge \mathcal{E}_2] - \Pr[\mathcal{E}_1 \wedge \mathcal{E}_3] - \Pr[\mathcal{E}_2 \wedge \mathcal{E}_3] + \Pr[\mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \mathcal{E}_3]$$

Conclusions

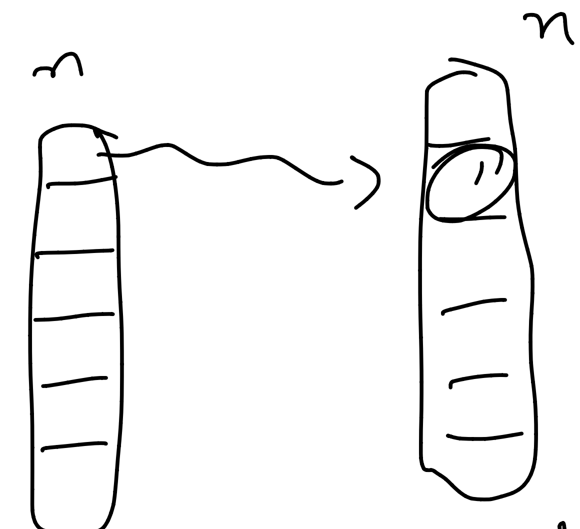
Suppose n = m:

- What is the expected size of each bin? (1) ✓
- What is the expected number of bins with exactly 4 balls? [~ n/(24e) ~ n/65] ✓
- What is the probability that there exists a bin with (log n) balls? $\leq \frac{1}{n}$.
- Maximum "load" = log n / (log log n)

[typically, there exists a bin with $\frac{\log n}{\log\log n}$ balls.]
w.h.p.



Random hashes



20 .

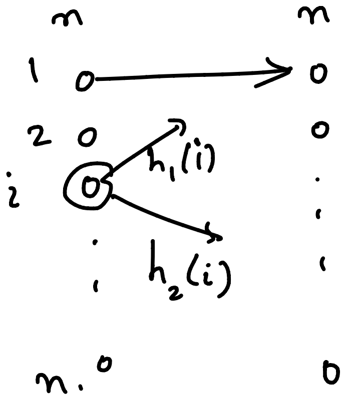
- Hash a set of n elements into memory of size n
- Size of max bin = log n / (log log n) ✱. (linear probing takes ~ log n time in the worst case.) .
- How large should memory be, so that max load is 1? 4?

. . $n^{5/4}$ . .

- Better than random assignment? (power of two choices)

[Azar, Broder, . . .] .

b diff. hash fns .



h_1, h_2 } random hash fns.

h': defined iteratively

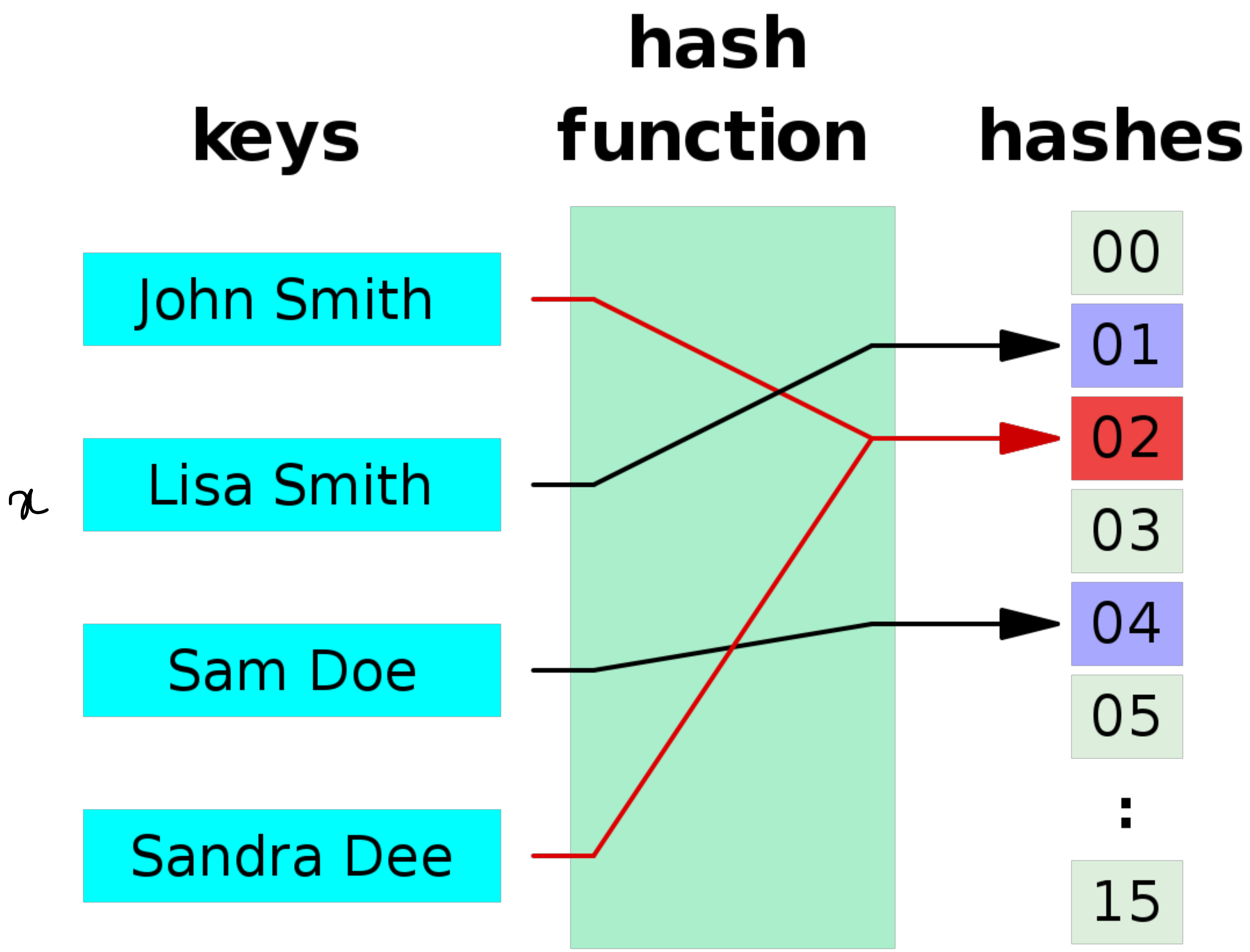
$h'(i) :=$ "less-crowded" of $h_1(i)$ & $h_2(i)$

max. bin size $= \log \log n$

$$\frac{\log\log n}{\log b}.$$

Balls and bins vs. hashing

$h(x) \rightarrow$ uniformly random bin.



- Function defined over the whole universe
- Function eval must be "cheap"
- Must do as well as random assignment

(Src: wikipedia)

Sampling / estimation

Sum of elements in array

Problem: let A be an array with n elements, each in interval [~~0~~ -1,1]. Find sum of all elements.

Sum $\in [-n, n]$.

- <u>Twist:</u> suppose we are OK with a little bit of error ($\sim$ 0.01 n)

Alg: Sample k elements w/replacement;

- compute sum of sample
- rescale by factor $\frac{n}{k}$.

Sum of elements in array

Problem: let A be an array with n elements, each in interval [0,1]. Find sum of all elements.

- Twist: suppose we are OK with a little bit of error ($\sim 0.01\, n$)
- Natural idea: sampling and re-scaling
- Questions: how bad can error be? With what probability? "confidence intervals"

Trade-offs

Key quantities:

- Number of samples (k)
- Error in result (more samples ⇒ smaller error).
- Confidence in result (" ⇒ more confidence.).

Formalizing sampling

Variance

Sample size and variance

Chebychev’s inequality