Lecture 7: k-wise independence

Top 4 probability facts:

- "Union Bound": $\Pr[A \text{ OR } B] \le \Pr[A] + \Pr[B]$
- "Linearity of Expectation": average of $X+Y$ = avg. of X + avg of Y
  $(E[X+Y])$ $(= E[X] + E[Y])$

no matter if A,B ~~[illegible]~~ X,Y independent, dependent, whatever

- Markov's Inequality: If avg. of some positive #'s is 100, at most $\frac{1}{3}$ of them are ≥ 300.

  If $X \geq 0$ and $E[X] = \mu$, $\Pr[X \geq c\mu] \le \frac{1}{c}$ $(c \geq 1)$

- $1 + x \approx e^x$ if x tiny [not about probability :)]

〜

[Why do we care about probability so much in an algs. class?
Life/alg. philosophy: if you don't know what choice to make, make a random one. At least this way it's hard for a devious adversary/ial input to force you into bad outcomes. Like when playing R.P.S.: if you play randomly, you know that no element of psychology can make your winning chances < 1/3...].

〜

[Let's talk about one of the most famous ~~[illegible]~~ probability chestnuts of all time:] "Birthday Paradox".

[actually play it!]

Balls & bins ver.: m balls (people)
n bins (birthdays). "$m \ll n$" [unlike $n=n$ cases studied previously]

(2)

distinct birthdays $\Longleftrightarrow$ max load 1.

Prob. is .... $\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)$

[Pr[2nd ball doesn't "collide" & 3rd ball doesn't collide & 4th ball...]

Rem: Using that all m ball locations are independent ($\Pr[A \& B \& C \& \ldots]$
$= \Pr[A]\cdot\Pr[B]\cdot\Pr[C]\cdots$ if all indep.)

$m \ll n \Rightarrow -\frac{1}{n}, \ldots, -\frac{m-1}{n}$ "tiny" [I'm being heuristic/ nonrigorous here. I can explain how to be more precise upon request!]

$\therefore \Pr[\text{max load } 1] \approx e^{-\frac{1}{n}} e^{-\frac{2}{n}} \cdots e^{-\frac{m-1}{n}}$

$$= e^{-\frac{1}{n}\left(1+2+\cdots+(m-1)\right)} \qquad \frac{m(m-1)}{2} \approx \frac{m^2}{2}$$

$$\approx e^{-m^2/2n}$$

Takeaway: • $m^2 \ll n \Rightarrow \frac{m^2}{2n}$ "tiny" $\Rightarrow \Pr[\text{all distinct}] \approx 1-\frac{m^2}{2n}$.

• $m^2 \gg n \Rightarrow \frac{m^2}{2n}$ "huge" $\Rightarrow \Pr[\text{all distinct}]$ expon. small

• $m = c\cdot\sqrt{n} \Rightarrow \frac{m^2}{2n} = \frac{c^2}{2} \Rightarrow \Pr[\text{all dist.}] \approx e^{-c^2/2}$
$= \frac{1}{2}$ if $c = \sqrt{2\ln 2}$
≈ 1.18

$(m = 1.18\sqrt{365} \approx 22.5)$.

$\therefore$ to ensure no collisions "with high probability", ~~need~~ $n \gg m^2$. is necessary & sufficient.

[Great, so if you really want no hash collisions, you're in good shape if you take $n \gg m^2$...
under SUHA.]

Beyond SUHA

[Now I'd like to "stop cheating" :) . We'll now try to ask...]

How to choose hash funcs. that provably achieve properties SUHA achieves?

[e.g. no collisions if $n >> m^2$]

[Let's change notation a little...]

hash function : {~~strings~~} $\longrightarrow \{0, 1, \ldots, n-1\}$

numbers [← not a big deal, every string can be naturally converted to a # (ascii, whatever) & vice versa]

$\{0, 1, 2, \ldots, U-1\}$ [← just putting an upper bound on the size of the #'s].

"Universe" size: think of as huge, like $U = 2^u$, $u = 1000s$, e.g.

all possible hash fcns:

	0	1	2	...		U-1
h_0	0	0	0	...		0
h_1	0	0	0	...		1
⋮						
h_{n-1}	0	0	0	...		n-1
h_n	0	0	0	...	1	0
⋮						
h_{n^U-1}	n-1	n-1	n-1	...	n-1	n-1

SUHA ≡ Choose h at random from all possible $h_0, \ldots, h_{n^U-1}$

Can't literally do! Storing h takes $U \cdot \lg n$ bits [maybe 2^{1000}]

Idea: Find a small set $\mathcal{H}$ of hash funcs (rows), all "easy to compute", only choose h randomly from $\mathcal{H}$.

Hope h is "random enough".

One notion of "random enough":

$\mathcal{H}$ is "universal" [bad term, but anyway] if...

- If you first fix any two elements $x, y < U$, ... $x \neq y$
- Then you choose $h \in \mathcal{H}$ at random ..
- Then $\Pr[h(x)=h(y)] \leq \frac{1}{n}$.

"collision"

Rem: SUHA $\equiv$ $\mathcal{H}$ is all possible is universal: $\Pr_h[h(x)=h(y)] = \frac{1}{n}$.

e.g. $U=3$, $n=2$. ← objs to hash / balls

0, 1 are "bins"

	0	1	2
h_0	0	0	0
h_1	0	1	1
h_2	1	0	1
h_3	1	1	0

weird hash function, hashes everything to 0!

$\mathcal{H} = \{h_0, h_1, h_2, h_3\}$ is universal. [And only uses 4 out of 8 poss. hash fcns.]

↳ for any 2 distinct columns, pick a random row, chance of same "bin" is $= \frac{1}{2} \leq \frac{1}{2}$ ✓.

Fact [we'll prove later]: For any $U \leq 2^u$, there is a simple universal family $\mathcal{H}$ of size 2^{2u}, hash functions "named" $h_{a,b}$ for all possible u-bit strings a, b; $h_{a,b}(x)$ = simple function of a, b, x.

[To implement your hash table, initially pick a, b at random: just $2u$ (a few thousand?) random bits. Then hash with $h_{a,b}$.]

[Let's assume for now...]

Q: Is universality "random enough"?

A: For some properties achievable with SUHA, yes.

[Let's go back to birthday paradox / balls & bins...] ⑤

Say m balls, n bins, $n \gg m^2$. SUHA $\Rightarrow$ max load 1 with high prob.

Proof breaks down without SUHA: $(1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{m-1}{n})$

assumed independence of $h(x_1), h(x_2), \ldots, h(x_m)$ [don't have any more]

Different proof [more or less, up to constant factor] works!

Say you are hashing $x_1, \ldots, x_m$ using h drawn randomly from universal $\mathcal{H}$.

$h(x_1), \ldots, h(x_m)$ are #'s ("bins") in $0 \ldots n-1$.

Define a ~~boolean~~ integer (random variable) $C_{1,2} = \begin{cases} 1 & \text{if } h(x_1) = h(x_2) \\ 0 & \text{if } h(x_1) \neq h(x_2) \end{cases}$ "balls 1 & 2 collide"

$$E[C_{1,2}] = \Pr[h(x_1) = h(x_2)] \cdot 1 + \Pr[h(x_1) \neq h(x_2)] \cdot 0$$
$$= \Pr_h[h(x_1) = h(x_2)]$$
$$\leq \frac{1}{n} \text{ by "universality".}$$

[expectation of "indicator" of event A = $\Pr[A]$]

Similarly define $C_{i,j}$ for $1 \leq i < j \leq m$, $\begin{cases} 1 & \text{if "ball } i,j \text{ collide"} \\ 0 & \text{if not.} \end{cases}$

Universality $\Rightarrow E[C_{i,j}] \leq \frac{1}{n}$.

Define $C = C_{1,2} + C_{1,3} + \cdots + C_{m-1,m} = \sum_{i<j} C_{i,j}$.

$= \underline{\text{\# of pairs of balls that "collide"}}$

Note: $C = 0 \iff$ no collisions / max load 1.

$C \geq 1 \iff$ at least one bin has load ≥ 2

③ ② ①

$C_{1,2} = 0$
$C_{1,3} = 1$
$C_{2,3} = 0$
$C = 1$.

$$E[C] = E[C_{1,2}] + E[C_{1,3}] + \cdots + E[C_{m-1,m}].$$

"linearity of expectation!"

$$\leq \underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{\binom{m}{2} \text{ times}}$$
$$= \frac{1}{n}\binom{m}{2} = \frac{1}{n}\frac{m(m-1)}{2} \leq \frac{m^2}{2n}$$

[remember that quantity?]

Say n chosen $\geq 10m^2$.

Then $\mathbb{E}[C] \leq \frac{m^2}{2 \cdot m} \leq \frac{m^2}{2 \cdot 10m^2} = \frac{1}{20} = .05$.

[Now think for a sec. C is either $0, 1, 2, 3, \ldots$
And supposedly "on avg." it's $\leq .05$. Then it's probably usually 0!]

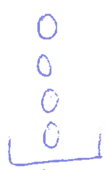
"Markov's inequality": $\underbrace{\Pr[C \geq 1]}_{\Pr[\text{collisions}]} \leq .05$ (otherwise, certainly $\mathbb{E}[C] > .05$)

$\therefore$ universality is enough to conclude: $\geq 95\%$ chance of max load 1, provided $n \geq 10m^2$.

[What about our "usual" setting, $m = n$?] $\xRightarrow{\text{Markov}}$ $\Pr[C \geq 10m] \leq \frac{1}{20} = .05$. ⊛.

If $n = m$, $\mathbb{E}[C] \leq \frac{m^2}{2m} \leq \frac{m}{2}$. [not amazing].

Q: If ith bin has load L_i, what will C be (at least)?



$L_i = 4 \rightsquigarrow$ contributes $\binom{4}{2}$ to C. $C = \sum_{i=0}^{n-1} \binom{L_i}{2}$.

i

$\binom{L_i}{2} \approx \frac{L_i^2}{2}$; $\geq \frac{L_i^2}{4}$ for $L_i \geq 2$. $\therefore C \geq \sum_{i=0}^{n-1} \frac{L_i^2}{4}$.

⊛ says $\geq 95\%$ of the time, $10m \geq C$
$\geq \sum_i \frac{L_i^2}{4} \Leftrightarrow \sum_i L_i^2 \leq 40m$.

[What does this imply about max load?] $\Rightarrow$ every L_i has $L_i^2 \leq 40m$
$\Rightarrow L_i \leq 6.33\sqrt{m}$.

$\geq 95\%$ chance of "max load $\leq O(\sqrt{m})$" is not so great:
SUHA would have $O\left(\frac{\log m}{\log\log m}\right)$.

[Sad thing is, this could happen.] [Good news $\rightarrow$ we'll see a cool trick next time, "double hashing", to work around it.]