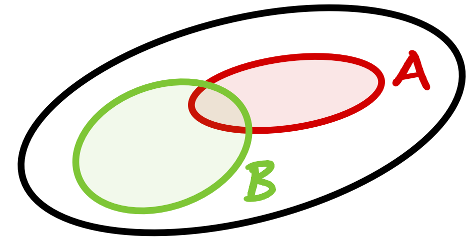
CS70 – Spring 2024

Lecture 17 – March 14

Review of Previous Lecture

- **Conditional Probability**



$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

- **Correlation & Independence**

$\Pr[A|B] > \Pr[A] \implies A, B$ positively correlated

$\Pr[A|B] < \Pr[A] \implies A, B$ negatively correlated

$\Pr[A|B] = \Pr[A] \implies A, B$ independent

equivalently: $\Pr[A \cap B] = \Pr[A]\Pr[B]$

Review (cont.)

- **Intersections of Events : Product Rule**

$$\Pr[A \cap B] = \Pr[B]\Pr[A \mid B]$$

$\Pr[A \cap B] = \Pr[A]\Pr[B \mid A]$

$$\Pr\left[\bigcap_{i=1}^{n} A_i\right] = \Pr[A_1] \times \Pr[A_2 \mid A_1] \times \Pr[A_3 \mid A_1 \cap A_2] \times \cdots \times \Pr[A_n \mid A_1 \cap \cdots \cap A_{n-1}]$$

- **Unions of Events : Inclusion-Exclusion**

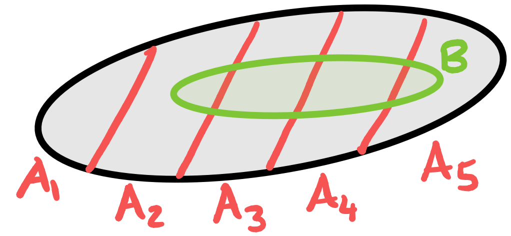
$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$$

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_i \Pr[A_i] - \sum_{i<j} \Pr[A_i \cap A_j] + \sum_{i<j<k} \Pr[A_i \cap A_j \cap A_k] - \cdots$$

- **Union Bound**: $\Pr\left[\bigcup_{i=1}^{n} A_i\right] \le \sum_{i=1}^{n} \Pr[A_i]$

Review (cont.)

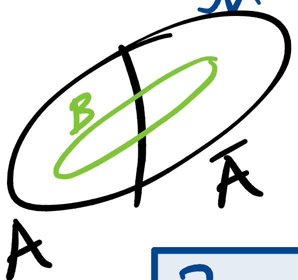
- **Law of Total Probability**



If $A_1 \ldots A_n$ partition Ω then

$$\Pr[B] = \sum_i \Pr[B \cap A_i] = \sum_i \Pr[B|A_i]\Pr[A_i]$$

In particular:



$$\Pr(B) = \Pr[B|A]\Pr[A] + \Pr(B|\bar{A}]\Pr[\bar{A}]$$

- **Bayes Rule**

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\underbrace{\Pr[B|A]\Pr[A] + \Pr[B|\bar{A}]\Pr[\bar{A}]}_{\text{can be computed if we know } \Pr[B|A],\ \Pr[B|\bar{A}],\ \Pr[A]}}$$

Today

Some applications of basic probability:

- Hashing (& Birthday "Paradox")
- Coupon Collecting
- Load Balancing

We will use:

- Concepts from last lecture (Union Bound, Product Rule, ...)
- Asymptotics (large-n approximations)

<u>Balls & Bins Model</u>

Throw m balls uniformly at random & independently into n bins

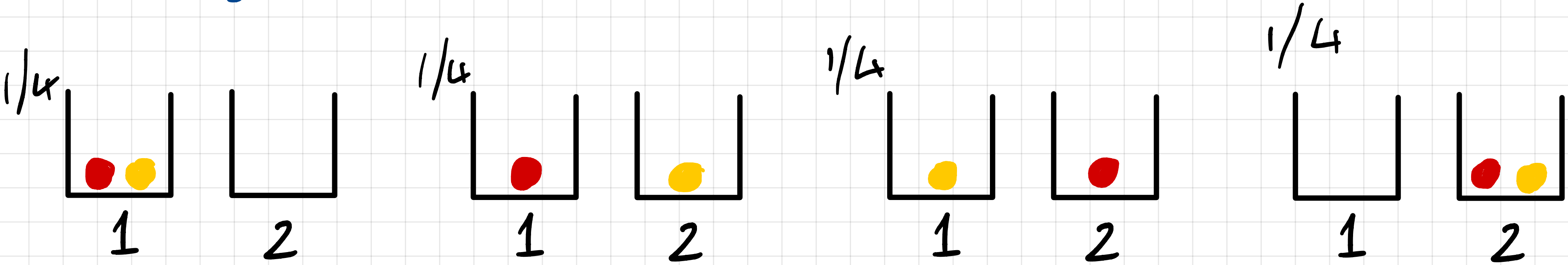
$$\Omega = \underbrace{\{1,\dots,n\} \times \{1,\dots,n\} \times \dots \times \{1,\dots,n\}}_{m \text{ times}}$$

[Each ball has choice of n bins]

$$|\Omega| = n^m$$

Probability space is <u>uniform</u>: for every $\omega = (b_1, \dots, b_m)$, $\Pr[\omega] = \frac{1}{|\Omega|} = \frac{1}{n^m}$.

E.g. $n = m = 2$ $\qquad |\Omega| = 2^2 = 4$



Events in Balls & Bins

E.g. E = "bin 1 is empty"

(i) Calulating $\Pr[E]$ using counting

Since prob. space is uniform, we have

$$\Pr[E] = \frac{|E|}{|\Omega|} = \frac{|E|}{n^m}$$

$|E|$ = # of ways of arranging balls s.t. Bin 1 is empty

$= (n-1)^m$

each ball now has only $n-1$ choices

So $\Pr[E] = \frac{(n-1)^m}{n^m} = \boxed{\left(1-\frac{1}{n}\right)^m}$

Example: If $m=n$ then $\Pr[E] = \left(1-\frac{1}{n}\right)^n \sim \boxed{\frac{1}{e} \approx 0.37}$

Events in Balls & Bins

E.g. E = "bin 1 is empty"

(ii) Calculating $\Pr[E]$ using Product Rule

Define A_i = "ith ball doesn't go to bin 1"

$\Pr[A_i] = 1 - \frac{1}{n}$ for all i

$$E = \bigcap_{i=1}^{m} A_i$$

$$\Pr[E] = \Pr[A_1] \times \Pr[A_2 | A_1] \times \Pr[A_3 | A_1 \cap A_2] \times \dots \times \Pr[A_m | A_1 \cap \dots A_{m-1}]$$

$$= \Pr[A_1] \times \Pr[A_2] \times \dots \times \Pr[A_m]$$

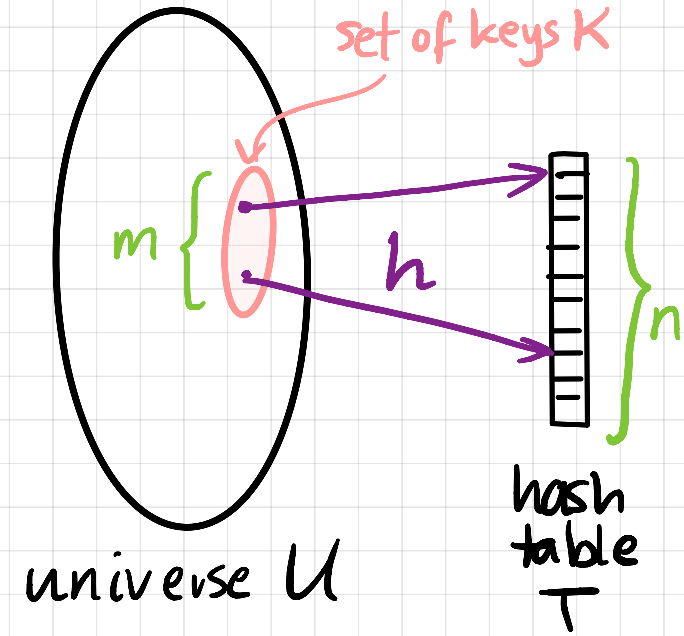
because the A_i are mutually independent!

$$= \boxed{\left(1 - \frac{1}{n}\right)^m}$$

same as before!

Application 1 : Hashing

Suppose we want to hash m keys into a hash table of size n



Use a <u>random</u> hash function h that sends keys independently & u.a.r. to table locations

$h : U \to T$

To ADD a key $x \in U$: store x at location $h(x)$ (using linked list if necessary)

To DELETE a key $x \in U$: remove x from location $h(x)$

To perform a MEMBER query for $x \in U$: check if x is stored at location $h(x)$

Goal: Avoid <u>collisions</u> ($\to$ linked lists)

$\boxed{Q}$: How large can m be (as a function of n) so that the probability of collisions is small?

<u>Analysis</u>: Balls & bins!

Keys = balls, Table locations = bins
m n

$\boxed{Q}$: In balls & bins with m balls, n bins, how large can m be so that (with good probability) no two balls land in same bin?

For now, "with good probability" = "with prob. $\geq 1/2$"

Rough calculation : Union Bound

For each (unordered) pair of balls $\{i,j\}$ with $i \neq j$, let $C_{\{i,j\}}$ denote the event that i,j land in same bin

Then $\Pr[C_{\{i,j\}}] = \frac{1}{n}$

[imagine i chooses bin first
$\Pr[j \text{ chooses same bin}] = \frac{1}{n}$]

$$\Pr[i,j \text{ go same place}] = \sum_k \Pr[i \to k] \underbrace{\Pr[j \to k \mid i \to k]}_{1/n} = \frac{1}{n}$$

Number of pairs $\{i,j\} = \binom{m}{2}$

Note that $\Pr[\text{some collision occurs}] = \Pr\left[\bigcup_{\{i,j\}} C_{\{i,j\}}\right]$

Union bound:

$$\Pr\left[\bigcup_{\{i,j\}} C_{\{i,j\}}\right] \leq \sum_{\{i,j\}} \Pr[C_{\{i,j\}}] = \binom{m}{2} \times \frac{1}{n} \leq \boxed{\frac{m^2}{2n}}$$

Union bound:

$$\Pr\Big[\bigcup_{\{i,j\}} C_{\{i,j\}}\Big] \le \sum_{\{i,j\}} \Pr[C_{\{i,j\}}] = \binom{m}{2} \times \frac{1}{n} \le \boxed{\frac{m^2}{2n}}$$

We want this prob. to be small (say, $\le 1/2$)

So we want $$\frac{m^2}{2n} \le \frac{1}{2}$$

i.e., $\boxed{m \le \sqrt{n}}$ (or $n \ge m^2$)

To get smaller collision prob. ε, just take $\boxed{m \le \sqrt{2\varepsilon n}}$

Bottom line: If the size of our hash table is roughly the square of the number of keys to be stored, then we're likely to have no collisions

More accurate calculation

Let A be the event "no collision occurs"

Then we can calculate $Pr[A]$ <u>exactly</u> as:

$$Pr[A] = \frac{|A|}{|\Omega|} = \frac{|A|}{n^m}$$

Q: What is $|A|$?

A: Number of ways of arranging the m balls in <u>different</u> bins

$= \#$ ways of choosing m items out of n <u>without</u> replacement

$= n \times (n-1) \times (n-2) \times \cdots \times (n-m+1)$

So

$$Pr[A] = \frac{n(n-1)(n-2)\cdots(n-m+1)}{n^m} = 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)$$

Alternatively, using Product Rule :

Let A_i = "ball i chooses different bin from balls $1, \ldots, i-1$"

Then $A = A_1 \cap A_2 \cap \cdots \cap A_m$

And $\Pr[A] = \Pr\left[\bigcap_{i=1}^{m} A_i\right]$

$$= \Pr[A_1] \times \Pr[A_2 \mid A_1] \times \Pr[A_3 \mid A_1 \cap A_2] \times \cdots \times \Pr[A_m \mid A_1 \cap \cdots \cap A_{m-1}]$$

$$= 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

same as above (phew!)

Since this is an exact formula for $\Pr[A]$, we can just fix any n and compute it for larger & larger values of m until $\Pr[A]$ drops to $1/2$ (or $1-\varepsilon$)

n	10	20	50	100	200	365	500	1000	10^4	10^5	10^6
m_0	4	5	8	12	16	22	26	37	118	372	1177

m_0 = largest m for which collision prob. remains below $1/2$

Can we get a <u>formula</u> for m_0 ?

$$\Pr[A] = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \cdots \left(1-\frac{m-1}{n}\right)$$

$$\ln \Pr[A] = \ln\left(1-\frac{1}{n}\right) + \ln\left(1-\frac{2}{n}\right) + \cdots + \ln\left(1-\frac{m-1}{n}\right)$$

$\ln(1-x) \approx -x$ for x small

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

$$= -\frac{1}{n} \sum_{i=1}^{m-1} i$$

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

$$\approx -\frac{m^2}{2n}$$

Hence $\boxed{\Pr[A] \approx e^{-m^2/2n}}$

$\Pr[A] \approx e^{-m^2/2n}$

Want $\Pr[A] = 1/2$ (or $\Pr[A] = 1-\varepsilon$)

This means

$$e^{-m^2/2n} = \frac{1}{2}$$

$$m^2 = (2\ln 2)n$$

So a more accurate bound is $m \leq \sqrt{(2\ln 2)n}$

$$\approx \boxed{1.177\sqrt{n}}$$

More generally (for collision prob. ε) $m \leq \sqrt{2\ln\left(\frac{1}{1-\varepsilon}\right)} \cdot \sqrt{n}$

n	10	20	50	100	200	365	500	1000	10^4	10^5	10^6
m_0	4	5	8	12	16	22	26	37	118	372	1177
$1.177\sqrt{n}$	3.7	5.3	8.3	11.8	16.6	22.5	26.3	37.3	118	372	1177

$m_0 =$ (exact) largest m for which collision prob. remains below $1/2$

$1.177\sqrt{n}$ = our approximation of m_0

Q: Why is 365 in the table?

Birthday "Paradox" / Birthday Problem

Q: In a room with m people, how large does m have to be so that $\Pr[2 \text{ people share a birthday}] \geq \frac{1}{2}$?

A: 10

20

50

100

300

Birthday "Paradox" / Birthday Problem

Q: In a room with m people, how large does m have to be so that $\Pr[\text{2 people share a birthday}] \geq \frac{1}{2}$?

A: This is exactly the collision problem for balls & bins!

\# bins $n = 365$

\# balls $m = $ \# people

(assumes all birthdays equally likely; ignores leap years)

From table, answer is $\boxed{m = 23}$

With $m = 60$, $\Pr[\text{2 people share a birthday}] > 99\%$

Application 2 : Coupon Collecting

There are n different baseball cards

Each box of cereal contains a uniformly random card

$\boxed{Q}$: How many boxes do we need to buy so that, with good probability, we have collected at least one copy of every card.

$\boxed{A}$: Balls & bins again!

Here we want to know how many balls we need to throw so that every bin contains at least 1 ball

Let A = "some bin is empty"

A_i = "bin i is empty"

Then $A = \bigcup_{i=1}^{n} A_i$

And $\Pr[A_i] = \left(1-\frac{1}{n}\right)^m$

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

(from earlier)

(using $\left(1-\frac{1}{n}\right)^n \xrightarrow[n\to\infty]{} e^{-1}$)

$\left(1-\frac{1}{n}\right)^n \sim e^{-1}$

$\left(1-\frac{1}{n}\right)^m \sim e^{-m/n}$

<u>Union Bound</u>:

$\Pr[A] \leq \sum_{i=1}^{n} \Pr[A_i] \approx ne^{-m/n}$

$ne^{-\left(\frac{n\ln n + n}{n}\right)}$

$= ne^{-(\ln n + 1)}$

$= \cancel{n} \times \cancel{\frac{1}{n}} \times e^{-1}$

So if we set $\boxed{m = n\ln n + n}$ we get

$\Pr[A] \leq e^{-1} < 1/2$

<u>Bottom line</u>: Need to buy about $\boxed{n\ln n}$ boxes !

E.g. for $n=100$, need to buy ~ 460 boxes

Application 3: Load Balancing

We have m jobs & n processors

We assign jobs independently and u.a.r. to processors

$\boxed{Q}$: What is the likely maximum load on a processor?

Obviously the max is at least $\lceil \frac{m}{n} \rceil$

But how much worse is it likely to be?

Focus on the case $\boxed{m = n}$ (#jobs = # processors)

Note: There will definitely be collisions since now $m \gg \sqrt{n}$

Strategy:

- Define A_k = "some processor has load $\geq k$"

  $\boxed{\text{Goal}}$: find smallest k s.t. $\Pr[A_k] \leq 1/2$ ← or ε

- Define $A_k(i)$ = "bin #i has load $\geq k$"

  $\boxed{\text{New goal}}$: find smallest k s.t. $\Pr[A_k(i)] \leq \frac{1}{2n}$

- Use Union Bound:

$$\Pr[A_k] = \Pr\left[\bigcup_{i=1}^{n} A_k(i)\right] \leq n \times \frac{1}{2n} = \frac{1}{2}$$

$\boxed{\text{New goal}}$: find smallest k s.t. $\Pr[A_k(i)] \le \frac{1}{2n}$

Focus on bin #i

For any subset $S \subseteq \{1, \dots, n\}$ of k balls, define

B_S = "all balls in S land in bin #i"

$\boxed{\text{Claim}}$: $A_k(i) = \bigcup_S B_S$

Union Bound (again!)

$$\Pr[A_k(i)] \le \sum_S \Pr[B_S]$$

And $\Pr[B_S] = \frac{1}{n^k}$; # of $S = \binom{n}{k}$

$\boxed{\text{So}}$: $\Pr[A_k(i)] \le \frac{1}{n^k}\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!\, n^k} \le \frac{1}{k!}$

$\boxed{\text{New goal}}$: find smallest k s.t. $\Pr[A_k(i)] \leq \frac{1}{2n}$

$$\Pr[A_k(i)] \leq \frac{1}{n^k}\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{k!\,n^k} \leq \frac{1}{k!}$$

Finally: We want

$$\frac{1}{k!} \leq \frac{1}{2n}$$

Taking logs: $\ln(k!) \geq \ln(2n)$

Standard approximation (Stirling): $\ln(k!) \simeq k\ln k - k$
(for large k)

So we want:

$$k\ln k - k \geq \ln(2n)$$

Solution: $\boxed{k \simeq \frac{\ln n}{\ln\ln n}}$ (for large n)

Bottom line: With prob. $\geq 1/2$, max. load is $\lesssim \frac{\ln n}{\ln\ln n}$

$\boxed{\text{Bottom line}}$: With prob. $\geq 1/2$, max. load is $\lesssim \dfrac{\ln n}{\ln\ln n}$

This bound is valid for very large values of n

For realistic values of n, we need to increase it a bit to allow for lower-order terms in our approximations — a more careful analysis leads to

$$\boxed{k \geq \frac{2\ln n}{\ln\ln n}}$$

n	10	20	50	100	500	1000	10^4	10^5	10^6	10^7	10^8	10^{15}
$\frac{2\ln n}{\ln\ln n}$	5.5	5.5	5.7	6.0	6.8	7.2	8.2	9.4	10.6	11.6	12.6	20

E.g.: Send 350 pieces of mail randomly to US population
Unlikely any one person gets more than ~13 pieces!

Next lecture

- Random variables [= functions on prob. spaces]
- Expectation [= mean/average]