CSTO – Spring 2024
Lecture 17 – March 14
Review of Previous Lecture

- **Conditional Probability**
  \[ P(A|B) = \frac{P(A \cap B)}{P(B)} \]

- **Correlation & Independence**
  \[ P(A|B) > P(A) \quad \Rightarrow A, B \text{ positively correlated} \]
  \[ P(A|B) < P(A) \quad \Rightarrow A, B \text{ negatively correlated} \]
  \[ P(A|B) = P(A) \quad \Rightarrow A, B \text{ independent} \]
  \( \iff \) equivalently: \[ P(A \cap B) = P(A)P(B) \]
Review (cont.)

Intersections of Events: Product Rule

\[ P(A \cap B) = P(B) \cdot P(A|B) \]
\[ P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cap \ldots \cap A_{n-1}) \]

Unions of Events: Inclusion-Exclusion

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
\[ P(\bigcup_{i=1}^{n} A_i) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots \]

Union Bound: \[ P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \]
Review (cont.)

- **Law of Total Probability**

  If $A_1, \ldots, A_n$ partition $\mathcal{R}$ then
  \[
  \Pr[B] = \sum_i \Pr[B \cap A_i] = \sum_i \Pr[B | A_i] \Pr[A_i]
  \]

  In particular:
  \[
  \Pr[B] = \Pr[B | A] \Pr[A] + \Pr[B | A] \Pr[A]
  \]

- **Bayes Rule**

  \[
  \Pr[A | B] = \frac{\Pr[B | A] \Pr[A]}{\Pr[B]} = \frac{\Pr[B | A] \Pr[A]}{\Pr[B | A] \Pr[A] + \Pr[B | A] \Pr[A]}
  \]

  can be computed if we know $\Pr[B | A], \Pr[B | A], \Pr[A]$
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)
Balls & Bins Model

Throw \( m \) balls uniformly at random into \( n \) bins

\[ \Omega = \{1, \ldots, n\} \times \{1, \ldots, n\} \times \cdots \times \{1, \ldots, n\} \]

\( |\Omega| = n^m \)

[Each ball has choice of \( n \) bins]

Probability space is uniform:

for every \( \omega = (b_1, \ldots, b_m) \),

\[ \Pr[\omega] = \frac{1}{|\Omega|} = \frac{1}{n^m} . \]

E.g. \( n = m = 2 \)

\[ |\Omega| = 2^2 = 4 \]

\[ \begin{array}{cccc}
1 & 2 & 1 & 2 \\
\text{1} & \text{2} & \text{1} & \text{2} \\
\end{array} \]
Events in Balls & Bins

E.g. $E = \text{“bin 1 is empty”}$

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

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

E.g. \( E = \text{"bin 1 is empty"} \)

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

Define \( A_i = \text{"ith ball doesn't go to bin 1"} \)

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

\[ E = \bigcap_{i=1}^{\infty} A_i \]

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

\[ = \Pr[A_1] \times \Pr[A_2] \times \ldots \times \Pr[A_m] \]

because the \( A_i \) are mutually independent!

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

same as before!
Application 1 : Hashing

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

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

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

**Analysis:** Balls & bins!

- Keys = balls, Table locations = bins

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. \( \geq \frac{1}{2} \)”
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 \( \Pr [C_{i,j}] = \frac{1}{n} \).

\[ \Pr [\bigcup C_{i,j}] \leq \sum_{\{i,j\}} \Pr [C_{i,j}] = \left( \binom{m}{2} \right) \times \frac{1}{n} \leq \frac{m^2}{2n} \]
Union bound:
\[
\Pr \left[ \bigcup_{\{i,j\}} C_{\{i,j\}} \right] \leq \sum_{\{i,j\}} \Pr \left[ C_{\{i,j\}} \right] = \left( \frac{m}{2} \right) \times \frac{1}{n} \leq \frac{m^2}{2n}
\]

We want this prob. to be small (say, \( \leq \frac{1}{2} \))
So we want \( \frac{m^2}{2n} \leq \frac{1}{2} \)

i.e., \( m \leq \sqrt{n} \) (or \( n \geq m^2 \))

To get smaller collision prob. \( \varepsilon \), just take \( m \leq \sqrt{2\varepsilon n} \)

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 $\Pr[A]$ exactly as:

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

Q: What is $|A|$?

A: Number of ways of arranging the $m$ balls in different bins
   = \# ways of choosing $m$ items out of $n$ without replacement
   = $n \times (n-1) \times (n-2) \times \ldots \times (n-m+1)$

So

$$\Pr[A] = \frac{n(n-1)(n-2)\ldots(n-m+1)}{n^m} = 1 \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)$$
Alternatively, using Product Rule:
Let $A_i$ = "ball $i$ chooses different bin from balls $1, \ldots, i-1$"
Then $A = A_1 \cap A_2 \cap \ldots \cap A_m$
And $\Pr[A] = \Pr \left[ \bigcap_{i=1}^{m} A_i \right]$

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

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

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 $\frac{1}{2}$ (phew!)
<table>
<thead>
<tr>
<th>$N$</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>365</th>
<th>500</th>
<th>1000</th>
<th>$10^4$</th>
<th>$10^5$</th>
<th>$10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_0$</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>22</td>
<td>26</td>
<td>37</td>
<td>118</td>
<td>372</td>
<td>1177</td>
</tr>
</tbody>
</table>

$m_0 = \text{largest } m \text{ for which collision prob. remains below } \frac{1}{2}$
Can we get a formula for \( m_0 \)?

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

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

\[
\ln(1-x) \approx -x \quad \text{for } x \text{ small}
\]

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

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

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

\[
\approx -\frac{m^2}{2n}
\]

Hence

\[
Pr[A] \approx \mathcal{C}^{-\frac{m^2}{2n}}
\]
\[ \Pr[A] = e^{-\frac{m^2}{2n}} \]

Want \( \Pr[A] = \frac{1}{2} \) (or \( \Pr[A] = 1 - \varepsilon \))

This means

\[ e^{-\frac{m^2}{2n}} = \frac{1}{2} \]

\[ m^2 = (2\ln 2)n \]

So a more accurate bound is

\[ m \leq \sqrt{(2\ln 2)n} \]

\approx 1.177\sqrt{n}

More generally (for collision probability \( \varepsilon \))

\[ m \leq \sqrt{2\ln \left( \frac{1}{\varepsilon} \right)} \cdot \sqrt{n} \]
<table>
<thead>
<tr>
<th>$n$</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>365</th>
<th>500</th>
<th>1000</th>
<th>$10^4$</th>
<th>$10^5$</th>
<th>$10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>22</td>
<td>26</td>
<td>37</td>
<td>118</td>
<td>372</td>
<td>1177</td>
</tr>
<tr>
<td>$1.177\sqrt{n}$</td>
<td>3.7</td>
<td>5.3</td>
<td>8.3</td>
<td>11.8</td>
<td>16.6</td>
<td>22.5</td>
<td>26.3</td>
<td>37.3</td>
<td>118</td>
<td>372</td>
<td>1177</td>
</tr>
</tbody>
</table>

$M_0$ = the largest $m$ for which collision probability remains below $\frac{1}{2}$

$1.177\sqrt{n}$ = 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 \( \Pr \left[ 2 \text{ people share a birthday} \right] \geq \frac{1}{2} \)?

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 \( m = 23 \)

With \( m = 60 \), \( \Pr \left[ 2 \text{ people share a birthday} \right] > 99\% \)
Application 2: Coupon Collecting

There are \( n \) different baseball cards. Each box of cereal contains a uniformly random card.

Q: How many boxes do we need to buy so that, with good probability, we have collected at least one copy of every card?

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 $A = \text{"some bin is empty"}$
$A_i = \text{"bin } i \text{ is empty"}$

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

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

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

Union Bound:

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

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

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

Bottom line: Need to buy about $n \ln n$ boxes!
E.g. for $n = 100$, need to buy $\sim 460$ boxes
**Application 3: Load Balancing**

We have $m$ jobs & $n$ processors
We assign jobs independently and u.a.r. to processors

**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 $A_k = \text{"some processor has load } \geq k\text{"}$
  
  Goal: find smallest $k$ s.t. $Pr[A_k] \leq \frac{1}{2}$

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

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

- Use Union Bound:
  
  $$Pr[A_k] = Pr[\bigcup_{i=1}^{n} A_k(i)] \leq n \times \frac{1}{2^n} = \frac{1}{2}$$
**New goal**: find smallest $k$ s.t. $Pr[A_k(i)] \leq \frac{1}{2^n}$

Focus on bin #i
For any subset $S = \{1, \ldots, n\}$ of $k$ balls, define $B_S = \{\text{all balls in } S \text{ land in bin } #i\}$

**Claim**: $A_k(i) = \bigcup_S B_S$

Union Bound (again!)
$$Pr[A_k(i)] \leq \sum_S Pr[B_S]$$

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

**So**: $Pr[A_k(i)] \leq \frac{1}{n^k} \binom{n}{k} = \frac{n(n-1) \ldots (n-k+1)}{k! n^k} \leq \frac{1}{k!}$
New goal: find smallest $K$ s.t. $Pr[A_k(i)] \leq \frac{1}{2^n}$

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

Finally: We want

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

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

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

So we want:

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

Solution: $K \approx \frac{\ln n}{\ln \ln n}$ (for large $n$)

Bottom line: With prob. $\geq 1/2$, max. load is $\leq \frac{\ln n}{\ln \ln n}$
**Bottom line:** With $p_{50} > \frac{1}{2}$, max. load is $\leq \frac{\ln n}{\ln \ln n}$

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

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

<table>
<thead>
<tr>
<th>$N$</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>$10^4$</th>
<th>$10^5$</th>
<th>$10^6$</th>
<th>$10^7$</th>
<th>$10^8$</th>
<th>$10^{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{2 \ln n}{\ln \ln n}$</td>
<td>55</td>
<td>55</td>
<td>5.7</td>
<td>6.0</td>
<td>6.8</td>
<td>7.2</td>
<td>8.2</td>
<td>9.4</td>
<td>10.6</td>
<td>11.6</td>
<td>12.6</td>
<td>20</td>
</tr>
</tbody>
</table>

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

- Random variables \([= \text{functions on probability spaces}]\)
- Expectation \([= \text{mean/average}]\)