2/8/23 HS1090

classes: M, Tu, W, Th (opt-)

Eval

Quiz 1,2 - 15% × 2 = 30%

Endsem - 40%

Spoken assignment - 10% (task, record video) [4-5 mins.]

Group task - 20% (in English, groups of 3-4)
- some aspect of Germany, culture, lit., philosophy, history
- banned topics: Hitler, cars, beer, football
↳ presented to class after quiz 2

Kaiser - Czar - Caesar - King

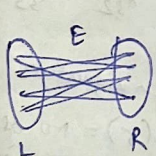
Liechtenstein
2 syllables

* Oder-Neisse line

2/8/23 CS6170

Perfect matching in bipartite graphs

$G(L, R, E)$



A matching $M \subseteq E$ s.t. $\forall u \in L \cup R$, $\exists$ at most one $e \in M$ that is incident on u

A perfect matching is a bijection ($|L| = |R|$)

Q: Given $G(L, R, E)$, check if G has a perfect matching.

There exist polynomial-time deterministic algos., but we're interested in looking at a rand. one

Q': Given $G(L, R, E)$, compute a p.m. if one exists.

We can consider some edge & check if it is present in any p.m. by deleting it & recursing on the smaller graph ⟹ property called self-reducibility

We form a soln to this by solving a diff prob.

Polynomial Identity Testing (PIT)

I/P: Program C that computes an n-variate degree d poly.

$$p(x_1, \ldots, x_n)$$

$$C(a_1, \ldots, a_n) = p(a_1, \ldots, a_n)$$

Q: check if $p \equiv 0$.

i.e., $\forall\, \alpha_1, \ldots, \alpha_n$, $p(\alpha_1, \ldots, \alpha_n) = 0$

No. of coefficients $= \binom{n+d-1}{n} \approx n^d$

How are these probs. related?

For $G(L, R, E)$, consider the bipartite adj. mat.

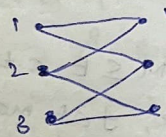
$$A \in \{0,1\}^{n \times n}, \quad n = |L| = |R|$$

$$A(i,j) = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{o.w.} \end{cases} \qquad [i \in L, j \in R]$$

$$P(i,j) = x_{ij}\, A(i,j)$$

(variable)

For



$$P = \begin{bmatrix} x_{11} & x_{12} & 0 \\ x_{21} & 0 & x_{23} \\ 0 & x_{32} & x_{33} \end{bmatrix}$$

$$\det(P) = \sum_{\sigma \in S_n} (-1)^{\operatorname{sgn}(\sigma)} \prod_{i=1}^{n} P(i, \sigma(i))$$

S_n: set of bijections from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$

$\operatorname{sgn}(\sigma)$: signature

$\operatorname{sgn}(\sigma)$ = No. of inversions in σ

$= |\{(i,j) \mid i<j,\ \sigma(i) > \sigma(j)\}|$

$\prod_{i=1}^{n} P(i, \sigma(i))$: $|E|$-variate degree n poly w/ at most $n!$ terms

Each surviving terms corresponds to a valid perfect matching

Obs: G has a p.m. iff $\det(P) \neq 0$

$\Rightarrow$ solving PIT solves P.M.! (w/ a n. simple algo)

Solving PIT

DeMillo-Lipton-Schwartz-Zippel lemma:

Let p be a non-zero n-variate deg-d poly. over $\mathbb{R}$.

Let $S \subseteq \mathbb{R}$. Then

$$\Pr_{a_1, a_2, \ldots, a_n \in S}\left[p(a_1, a_2, \ldots, a_n) = 0\right] \leq \frac{d}{|S|} \quad \text{(indep. of } n!)$$

($a_1, a_2, \ldots, a_n$: chosen indep. & uniformly at random)

3/8/23

Here, sample space $\Omega = \underbrace{S \times S \times \cdots \times S}_{n \text{ times}} = S^n$

The event of interest is the set of zeros,

i.e., $Z \subseteq S^n = \{(a_1, a_2, \ldots, a_n) \in S^n \mid p(a_1, \ldots, a_n) = 0\}$

$\Pr(Z) = ?$

The proba. space is $(\Omega, \mathcal{F}, \Pr)$

$\mathcal{F} = \mathcal{P}(\Omega) = 2^{\Omega}$ (set of all events)

$\Pr : \Omega \to [0,1]$ (proba. of choosing each tuple)

$\Pr(a_1, \ldots, a_n) = \frac{1}{|S|^n}$ (indep. & u.a.r)

It satisfies:

1. $\forall E \in \mathcal{F} \quad 0 \leq \Pr(E) \leq 1$
2. $\Pr(\Omega) = 1$
3. If $E_1, E_2, \ldots, E_n$ are pairwise disjoint,
$$\Pr(\cup E_i) = \sum \Pr(E_i)$$

Some facts:

* Let $E_1, E_2, \ldots, E_n$ be any n events

$$\Pr(\cup E_i) \leq \sum_{i=1}^{n} \Pr(E_i) \quad \text{[union bound]}$$

* **Conditional proba:**

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

* E & F are indep. if $\Pr(E|F) = \Pr(E)$

$\Rightarrow \Pr(E \cap F) = \Pr(E)\Pr(F)$

* **Law of total proba:**

$E_1, \ldots, E_n$ form a partition of Ω for disjoint events $E_1, E_2, \ldots, E_n$ s.t. $\cup E_i = \Omega$,

$$\Pr(F) = \sum_{i=1}^{n} \Pr(F \cap E_i)$$

Proof of DLSZ-lemma

[By induction on n]

Base case: $n = 1$

$\Pr_{a \in S}[p(a) = 0] \leq \frac{d}{|S|}$

$Z = \{a \in S \mid p(a) = 0\}$

$|Z| \leq d$

$\Rightarrow \Pr(Z) \leq \frac{d}{|S|}$

Algo: (A)

Sample $a_1, \ldots, a_n \in_r S$ (indep. & u.a.r.)

If $p(a_1, \ldots, a_n) = 0$, say $p \equiv 0$

else, say $p \not\equiv 0$

Note:

① If $p \equiv 0$, then $\Pr(A \text{ says } p \equiv 0) = 1$

② If $p \not\equiv 0$, then

$\Pr(A \text{ says } p \equiv 0) \leq \frac{d}{|S|}$ (error proba.)

$\Rightarrow$ one-sided error algo.

Induction step

$p(x_1, \ldots, x_n) = \sum_{i=0}^{d} x_1^i \, p_i(x_2, \ldots, x_n)$

Find largest i s.t. $p_i \not\equiv 0$

Z_i = set of tuples on which p_i evals. to 0

deg. $p_i \leq d - i$

$\Rightarrow \Pr(Z_i) \leq \frac{d-i}{|S|}$ (induction hypothesis)

$\Pr(Z) = \Pr(Z \cap Z_i) + \Pr(Z \cap \overline{Z_i})$

$= \Pr(Z_i) \cdot \Pr(Z \mid Z_i) + \Pr(\overline{Z_i}) \cdot \Pr(Z \mid \overline{Z_i})$

$\Pr(Z_i) \leq 1$; $\Pr(Z \mid \overline{Z_i}) \leq \frac{i}{|S|}$

we have a lower bound, but we want an upper bound, we can just choose 1

[p_i's are evaluated, now we have a poly. on x_i, w/ deg. i]

$\leq \frac{d-i}{|S|} \cdot 1 + 1 \cdot \frac{i}{|S|}$

$\leq \frac{d}{|S|}$

We can boost the success proba. by repeating the sampling.

i.e., A'

Repeat k times

- Sample $a_1, \ldots, a_n \in_r S$
- If $p(a_1, \ldots, a_n) \neq 0$, say $p \not\equiv 0$

say $p \equiv 0$

Error proba. now is $\left(\frac{d}{|S|}\right)^k$

Running time inc. by factor of k

If $p \equiv 0$, $\Pr(A' \text{ says } p = 0) = 1$

If $p \not\equiv 0$, $\Pr(A' \text{ says } p = 0)$

$= \Pr(p = 0 \text{ in every iteration})$

$= \Pr(Z_1 \cap Z_2 \cap \ldots \cap Z_k) = \Pr(Z_1) \cdot \Pr(Z_2) \cdots \Pr(Z_k)$

$\leq (d/|S|)^k$

4/8/23

Verifying matrix multiplication (Frievald's algorithm)

Given $n \times n$ matrices A, B, C,

Check if $A \cdot B = C$

- can you do this asymptotically faster than matrix-multiplication?

$O(n^3)$ - naive

$O(n^{2.37})$ - complex

conjectured to be $O(n^{2+\epsilon})$ for v. small ϵ

Frievald's algo. → randomized $O(n^2)$

Algo

- choose $\bar{r} \in \{0,1\}^n$
- check if $\underbrace{ABr}_{\underbrace{A(B\bar{r})}_{O(n^2)}} = C\bar{r}$

$\Omega = \{0,1\}^n$

$Pr(\bar{a}) = 1/2^n \quad \forall\, \bar{a} \in \Omega$

Theorem

If $A \cdot B \neq C$, → non-zero matrix $D = AB - C$

$$Pr_{\bar{r} \in \{0,1\}^n}[D\bar{r} = 0]$$

$$Pr_{\bar{r} \in \{0,1\}^n}[AB\bar{r} = C\bar{r}] \leq 1/2$$

Consider that the first row of D is non-zero (WLOG) & $D_{1,1} \neq 0$

$$(D_{1,1}\ D_{1,2}\ \cdots\ D_{1,n}) \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

$$Pr_{r \in \{0,1\}^n}\left(\sum_{i=1}^{n} D_{1,i} r_i = 0\right)$$

$\leq 1/2$ by DLSZ-lemma, where variables are r_i, $i = 1, \ldots, n$ & identically zero implies $AB_{1,:} = C_{1,:}$

$$= Pr_{r \in \{0,1\}^n}\left(r_1 = \frac{-\sum_{i=2}^{n} D_{1,i} r_i}{D_{1,1}}\right)$$

or

$r_1, r_2, \ldots, r_n \in \{0,1\}$ → we choose $r_2, \ldots, r_n$ & finally choose r_1

[Principle of deferred decisions]

$\leq 1/2$

$$= \sum_{b_2, \ldots, b_n \in \{0,1\}} Pr\left(r_1 = \frac{-\sum_{i=2}^{n} D_{1,i} \cdot b_i}{D_{1,1}} \,\middle|\, r_2, \ldots, r_n = b_2, \ldots, b_n\right) \cdot Pr(r_2 = b_2, \ldots, r_n = b_n)$$

$\leq 1/2$

If we want to repeat it until error proba. is ϵ,

we want $(1/2)^n = \epsilon$

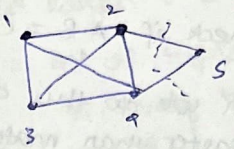
$$n = -\log_2 \epsilon = \log_2 1/\epsilon$$

Min-cut in graphs

$G(V, E)$

A cut set is a set of edges whose removal disconnects G

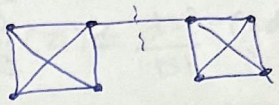
Min-cut is a cut set of smallest size/cardinality



min. cut size 2

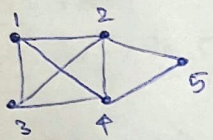
- A rand. algo. (Karger '93)

repeat until 2 vertices remain:
- choose $e \in E$ u.a.r.
- contract e
  $G \leftarrow G \setminus \{e\}$
  & keep parallel edges

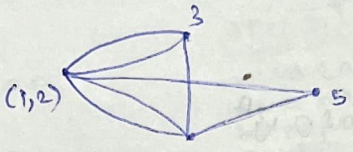


min cut size 1

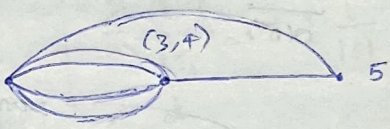
- remaining edges belong to min cut



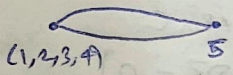
$(1,2)$



$(3,4)$



$(1,2),(3,4)$



min-cut size = 2

If k is size of min. cut,

$|E| \geq \frac{nk}{2}$ (degree of each vertex $\geq k$)

Obs: If G has min cut size k, then G has at least $\frac{kn}{2}$ edges

[If min-cut size $\geq k$, then
$\forall u \in V$, $\deg(u) \geq k$
no. of edges $= \frac{1}{2} \sum_{u \in V} \deg(u) \geq \frac{nk}{2}$]

Fix a min-cut C of size k

$\Pr(\text{no edge from } C \text{ is sampled}) \geq 1 - \frac{2}{n}$

7/8/23

E_i : Event that an edge from C was not sampled in the i^{th} step

$\Pr(E_1) \geq 1 - \frac{2}{n}$

Assume that E_1 occurred

$G' = G \setminus \{e\}$

obs: Every cut in G' is a cut in G

$\Rightarrow$ The min-cut size in G' is also k

$F_i = \bigcap_{j=1}^{i} E_j$ = the event that no edge from C was sampled in the first i iterations

We are interested in $\Pr(F_{n-2})$

$\Pr(E_2 | F_1) \geq 1 - \frac{2}{n-1}$ [no of edges in $G \setminus \{e\} \geq \frac{(n-1)k}{2}$]

$\Pr(E_i | F_{i-1}) \geq 1 - \frac{2}{n-i+1}$

$\Pr(F_{n-2}) = \Pr(F_{n-3}) \cdot \Pr(E_{n-2} | F_{n-3})$

$= \Pr(E_{n-2} | F_{n-3}) \cdot \Pr(E_{n-3} | F_{n-4}) \cdot \dots \cdot \Pr(E_1)$

$\geq \prod_{i=1}^{n-2} \frac{n-i-1}{n-i+1} = \frac{1 \times 2}{n(n-1)}$

$\geq \frac{2}{n(n-1)}$ → v. low success proba.

for some fixed min cut ⇒ $O(n^2)$ mincuts possible, more not possible

$\Pr(C \text{ was not outputted}) \leq 1 - \frac{2}{n(n-1)}$

Repeat k times

$\Pr(C \text{ was not output in any iteration}) \leq \left[1 - \frac{2}{n(n-1)}\right]^k$ ($\star$ $1-x \leq e^{-x}$)

$\leq e^{\frac{-2k}{n(n-1)}}$ $\quad k = \log 1/\epsilon \cdot \frac{n(n-1)}{2}$

⇒ success proba $1-\epsilon$

Max-cut

$G(V, E)$

Partition $V = V_1 \cup V_2$ s.t.

$|E(V_1, V_2)|$ is maximized

no. of edges across the partition/cut

we analyse a $\frac{1}{\alpha}$-approx

- $O(2^n)$ brute-force algo.
- NP-hard
- α-approximation
  - Algo. that outputs a cut of size $\geq \alpha$ (max.), $\alpha < 1$
  - Best known α for poly. time: $\alpha = 0.8$
  - $\alpha \geq 0.94$ not possible unless P = NP

Algo. A

$\forall v \in V$,

put v in V_i w.p. $1/2$ ind. of other vertices

$A(x, r)$ → rand. var.

$A : \{0,1\}^n \to \mathbb{R}$

(n-bit str. denoting a partition → cut size)

Theorem: $E(x) \geq \frac{m}{2}$ (size of the rand. cut; m = total no. of edges)

9/8/23

Expectation (AKA first moment)

- Most basic parameter for a r.v.

$$E(X) = \sum_{\omega \in \Omega} Pr(\omega) \cdot X(\omega)$$

Note: $Pr(X=a) = \sum_{\omega \in \Omega,\ X(\omega)=a} Pr(\omega)$

- Linearity of expectations:

If $X_1, X_2, \ldots, X_n$ are r.v.s, $X = \sum_{i=1}^{n} X_i$, then

$$E(X) = \sum_{i=1}^{n} E(X_i)$$

- Independence:

Criterion: $Pr(X=a \cap Y=b) = Pr(X=a) \cdot Pr(Y=b) \quad \forall a, b$

Analysis of A

$\forall e \in E$, define X_e = indicator r.v.s → $E(X) = Pr(X=1)$ for those

$$X_e = \begin{cases} 1, & \text{if } e \in \text{cut} \\ 0, & \text{o.w.} \end{cases}$$

indep.

$$\text{Cut-size} = X = \sum_{e \in E} X_e$$

$$= \sum_{\substack{e \in E \\ e=(u,v)}} Pr(u \in V_1 \ \& \ v \in V_2) + Pr(u \in V_2 \ \& \ v \in V_1)$$

($Pr(u \in V_1 \ \& \ v \in V_2) = 1/2$, $Pr(u \in V_2 \ \& \ v \in V_1) = 1/2$)

$$= m \times \frac{1}{2} = \frac{m}{2}$$

- doesn't necessarily hold for 3 or more r.v.s ⇒ pairwise indep. r.v.s

$\{X_{e_1}$ is indep of X_{e_2} $\forall e_1, e_2 \in E$

- if e_1 & e_2 are both incident on some $v \in V$, knowing X_{e_1} does not influence X_{e_2}
- if they don't share a vertex, they are indep.

Theorem: Every graph G w/ m edges contains a cut of size $\geq \frac{m}{2}$

Pf: $E(X) = \sum_{\text{cuts } C} Pr(C) \cdot \text{size}(C) = \frac{m}{2}$

If $\text{size}(C) < \frac{m}{2} \ \forall C$

$\Rightarrow E(X) < \frac{m}{2}$

$\Rightarrow \exists C$ w/ size $\geq m/2$ [Probabilistic method]

Derandomizing the algo.

→ naive: Brute force over the 2^n possible random choices

Obs: We have assumed indep. of choosing $V_i \ \forall v \in V$, however, we only reqd. pairwise indep.

Thm: We can construct $n = 2^b - 1$ pairwise-indep. rand. bits from b completely rand. bits

Example:

$b = 3, n = 7$

rand. bits: Y_1, Y_2, Y_3

001	$X_1 = Y_1$
010	$X_2 = Y_2$
011	$X_3 = Y_1 \oplus Y_2$
100	$X_4 = Y_3$
101	$X_5 = Y_3 \oplus Y_1$
110	$X_6 = Y_3 \oplus Y_2$
111	$X_7 = Y_3 \oplus Y_1 \oplus Y_2$

(At any pt., use the newest var. & do XOR w/ a previous combo not containing it)

Modified algo.

- sample $\log n$ bits u.a.r.
- construct n pairwise indep rand-bits defining a cut
- output the size of the cut

No. of rand. choices $= 2^{\log n} = n$

$\Rightarrow$ Brute-forcing over all $\log n$ bit strings obtains a poly-time deterministic algo.

. same results as before

$\mathbb{E}(X) = m/2$ ↳ over the $\log n$ bits

10/8/23

Alternate procedure for derandomization

Method of conditional expectation

---

conditional expectation: Given r.v.s X, Y,

$$\mathbb{E}(X \mid Y = y) = \sum_x x \cdot \Pr(X = x \mid Y = y)$$

Fact: 1) $\mathbb{E}(X) = \sum_y \Pr(Y = y)\, \mathbb{E}(X \mid Y = y)$

2) $\mathbb{E}\left(\sum_{i=1}^{n} X_i \mid Y = y\right) = \sum_{i=1}^{n} \mathbb{E}(X_i \mid Y = y)$

$$\mathbb{E}(X) = \Pr(v_1 \in V_1) \cdot \mathbb{E}(X \mid v_1 \in V_1) + \Pr(v_1 \in V_2) \cdot \mathbb{E}(X \mid v_1 \in V_2)$$

$$= \frac{1}{2} \cdot \mathbb{E}(X \mid v_1 \in V_1) + \frac{1}{2} \cdot \mathbb{E}(X \mid v_1 \in V_2) = \frac{m}{2} \text{ (from analysis before)}$$

Put $v_1 \in V_1$ arbitrarily

$$\mathbb{E}(X \mid v_1 \in V_1) = \Pr(v_2 \in V_1) \cdot \mathbb{E}(X \mid v_1 \in V_1 \,\&\, v_2 \in V_1) + \Pr(v_2 \in V_2) \cdot \mathbb{E}(X \mid v_1 \in V_1 \,\&\, v_2 \in V_2)$$

($\Pr(v_2 \in V_1) = 1/2$, $\Pr(v_2 \in V_2) = 1/2$; $\mathbb{E}(X \mid v_1 \in V_1) = m/2$)

The larger of these must be $\geq m/2$ ↳ keep choosing the larger

$$\mathbb{E}(X \mid v_1, v_2, \ldots, v_{i-1}) = \Pr(v_i \in V_1) \cdot \mathbb{E}(X \mid v_1, v_2, \ldots, v_{i-1}, v_i \in V_i) + \Pr(v_i \in V_2) \cdot \mathbb{E}(X \mid v_1, v_2, \ldots, v_{i-1}, v_i \in V_2)$$

↳ $\geq \frac{m}{2}$ (from prev. steps)

↳ No. of cut edges b/w $v_1, \ldots, v_{i-1}$ + No. of cut edges incident on v_i + $\frac{1}{2} \times$ No. of other edges (almost one vertex incident on $v_1, \ldots, v_n$)

↳ recursing

sequential greedy algo:

- Fix some order of vertices $v_1, \ldots, v_n$
- For $i = 1, \ldots, n$:
  - Add v_i to V_1 if no. of cut edges from v_i to $v_1, \ldots, v_{i-1}$ is larger
  - Else, add v_i to V_2

11/9/23

Quicksort

- Randomly choose a pivot
- Partition array based on pivot $\longrightarrow O(n)$
- Recursively sort the two parts

Running time:

worst-case $O(n^2)$

Expected $\Theta(n \log n)$

Running time $\propto$ no. of comparisons

$\bar{a} = a_1 a_2 \ldots a_n \xrightarrow{\text{sorted}} b_1 b_2 \ldots b_n$

$$X_{ij} = \begin{cases} 1, & \text{if } b_i \text{ \& } b_j \text{ were compared at some pt. in Quicksort} \\ 0, & \text{o.w.} \end{cases}$$

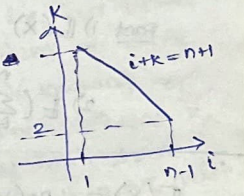
$\longrightarrow$ this occurs when the first pivot in $b_i, b_{i+1}, \ldots, b_{j-1}, b_j$ is b_i or b_j

$$X = \sum_{i<j} X_{ij}$$

$$E(X_{ij}) = Pr(X_{ij} = 1) = \frac{2}{j-i+1}$$

$$E(X) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1}$$

$$= \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} = \sum_{k=2}^{n} \sum_{i=1}^{n-k+1} \frac{2}{k} = \sum_{k=2}^{n} (n-k+1)\frac{2}{k}$$



$$= 2(n+1)\sum_{k=2}^{n} \frac{1}{k} - 2(n-1)$$

$\log n + \Theta(1)$ ★ Harmonic sum

$$= 2n\log n + \Theta(n)$$

$\Rightarrow$ Expected $\Theta(n \log n)$

We will show later that

$$Pr(X > cn\log n) \leq 1/n$$

A similar analysis holds if we randomly permute the array & then choose the first/last/some fixed index element as the pivot repeatedly

A few standard r.v.s

① Bernoulli r.v. $\equiv$ Indicator r.v.

$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1-p \end{cases} \qquad E(X) = Pr(X=1) = p$$

② Binomial r.v.

Let $X_1, X_2, \ldots, X_n$ be Bernoulli trials, i.e., rand. expts. w/ success proba. p

A binomial r.v. counts the no. of successes in n Bernoulli trials, i.e., $X = \sum_{i=1}^{n} X_i$

Range of $X = \{0, 1, \ldots, n\}$

$$Pr(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \longrightarrow \text{proba. mass fn.}$$

$$E(X) = np$$

(3) Geometric r.v.

* Counts the no. of Bernoulli trials before the first success
* Range of the r.v. = $\mathbb{N}$
* $Pr(X = i) = (1-p)^{i-1} p$
* Memorylessness

$$Pr(X = n+k \mid \underbrace{X > k}_{\text{no success in first } k \text{ trials}}) = Pr(X = n)$$

* $E(X) = Pr(Y=1) \cdot \underbrace{E(X|Y=1)}_{1} + Pr(Y=0) \cdot E(X|Y=0)$ — Y: first trial
  - $Pr(Y=1) = p$
  - $Pr(Y=0) = 1-p$
  - $E(X|Y=0)$: $X = 1 + Z$, Z being an identical geometric r.v. [memorylessness]

$$= p + (1-p)(E(1+Z))$$
$$= 1 + (1-p)E(Z) = 1 + (1-p)E(X)$$

$$\Rightarrow pE(X) = 1 \Rightarrow E(X) = \frac{1}{p}$$

Coupon collector problem

n coupons

A box of chocolate has one of the coupons u.a.r.

Q. How many boxes must be collected to obtain one copy of every coupon?

X_i = No. of boxes that are bought after $i-1$ coupons are obtained before obtaining the i^{th} coupon

X_i is a geometric r.v. with success proba. $\frac{n-i+1}{n}$

No. of boxes, $X = \sum_{i=1}^{n} X_i$

$$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{j=1}^{n} \frac{1}{j}$$

$$= n \log n + \Theta(n)$$

1/15/23

Markov's inequality

If X is a non-neg. r.v. & $a > 0$, then

$$Pr(X \geq a) \leq \frac{E(X)}{a}$$

i.e., $Pr(X \geq k E(X)) \leq \frac{1}{k}$

Pf: Consider the indicator r.v. $I = \begin{cases} 1, & X \geq a \\ 0, & \text{o.w.} \end{cases}$

$Pr(X \geq a) = Pr(I = 1) = E(I)$

Note that $I \leq \frac{X}{a}$

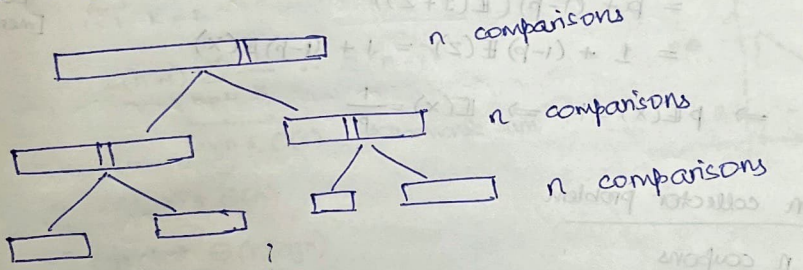
$\Rightarrow E(I) \leq E\left(\frac{X}{a}\right) = \frac{E(X)}{a}$

$\therefore Pr(X \geq a) \leq \frac{E(X)}{a}$

A bound for Quicksort

X = No. of comparisons, $E(X) = 2n\log n + \Theta(n)$

Thm: $Pr(X > k n \log n) \leq \frac{1}{n}$, k is a constant indep. of n



Each level performs n comparisons

$\Rightarrow$ Bounding no. of levels bounds no. of comparisons

Fix an element a of the array

X_i = size of the array containing a in the i^{th} level of recursion

$X_0 = n$

$$E(X_i) = \sum_{t > 0} Pr(X_{i-1} = t) \cdot E(X_i \mid X_{i-1} = t)$$

We want this to be $\alpha \cdot X_{i-1}$ to bound levels as $\Theta(\log n)$

Choosing an element in 25^{th} - 75^{th} percentile bounds the size of the array with a to be at most $\frac{3}{4} X_{i-1}$

$$\mathbb{E}(x_i \mid x_{i-1} = t) \leq \frac{1}{2} \cdot \frac{3}{4} x_i + \frac{1}{2} \cdot x_i = \frac{7}{8} t$$

$$\Rightarrow \mathbb{E}(x_i) \leq \sum_{t>0} \Pr(x_{i-1} = t) \cdot \frac{7}{8} t = \frac{7}{8} \sum_{t>0} \Pr(x_{i-1} = t) \cdot t$$

$$\mathbb{E}(x_i) \leq \frac{7}{8} \mathbb{E}(x_{i-1})$$

$$\Rightarrow \mathbb{E}(x_i) \leq \left(\frac{7}{8}\right)^i n$$

Choose $i = 3\log_{8/7} n$

$$\Rightarrow \mathbb{E}(x_i) \leq n \left(\frac{7}{8}\right)^{3\log_{8/7} n} = n\left(\frac{1}{n^3}\right)$$

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

We have only considered the bound for a single element a.

By Markov's inequality, for $a = 1$,

$$\Pr(x_i > 1) \leq \frac{1}{n^2} \quad \text{for } i = 3\log_{8/7} n$$

Let E_a be the event that a exists in a subarray of size > 1 after $i = 3\log_{8/7} n$ recursive steps.

$$\Pr(E_a) \leq \frac{1}{n^2}$$

$$\Pr(\exists a \text{ s.t. } E_a) \leq \sum \Pr(E_a) = n\left(\frac{1}{n^2}\right)$$

$$\leq 1/n$$

Variance of a r.v.

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2), \quad \mathbb{E}(X) = \mu$$

$$= \mathbb{E}(X^2 - 2\mu X + \mu^2)$$

$$= \mathbb{E}(X^2) - \mu^2$$

Properties

① $\text{var}(kX) = k^2 \text{var}(X)$

② $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$,

$\text{cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

If X & Y are independent,

$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

* $X \sim \text{Bin}(n, p)$

$X = X_1 + X_2 + \cdots + X_n$, $X_i \sim$ Bernoulli(p)

$\text{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = p - p^2 = p(1-p)$

$\text{Var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p)$

* $X \sim Geom(p)$

$$E(X^2) = p \cdot 1 + (1-p) \cdot E((X+1)^2) \quad \text{(memorylessness)}$$
$$= p + (1-p)\left[E(X^2) + 2\underbrace{E(X)}_{1/p} + 1\right]$$
$$= 1 + 2\left(\frac{1-p}{p}\right) + (1-p)E(X^2)$$

$$p\,E(X^2) = \frac{2-p}{p} \Rightarrow E(X^2) = \frac{2-p}{p^2}$$

$$Var(X) = E(X^2) - E(X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

Chebyshev's inequality

Let X be any r.v. & $a > 0$

$$Pr(|X-\mu| \geq a) \leq \frac{Var(X)}{a^2}$$

i.e., $Pr(\mu - a \leq X \leq \mu + a) \geq 1 - \frac{Var(X)}{a^2}$

Pf: Consider $Y = (X-\mu)^2$

Y is a non-neg. r.v.

$\Rightarrow$ By Markov's inequality, $Pr(Y \geq a^2) \leq \frac{E(Y)}{a^2}$

$$Pr(|X-\mu| \leq a) \leq \frac{E((X-\mu)^2)}{a^2} = \frac{Var(X)}{a^2}$$

Randomness-efficient probability amplification

Consider Freivald's algo.

$$\underbrace{AB \overset{?}{=} C}_{n \times n}$$

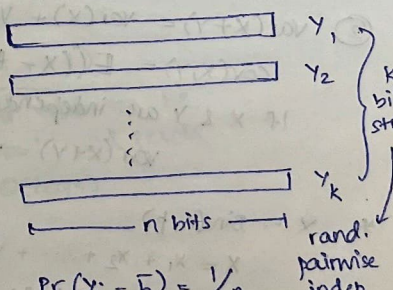
* We use n rand. bits
* $AB \neq C \Rightarrow Pr(\text{Error}) \leq 1/2$
* $AB = C \Rightarrow Pr(\text{Error}) = 0$
* Repeating the event k times, $Pr(\text{Error}) \leq 1/2^k$
  w/ no. of rand. bits used $= nk$

Alternative:

Repeating k times,

$Pr(\text{Error}) \leq 1/k$, $k \leq n$

No. of rand. bits $= 2n$



$Pr(Y_i = \bar{b}) = 1/2^n$

* The construction of k rand. pairwise indep. bit strings of len. n can be done w/ $n \log k$ bits
  - For each bit pos., generate $\log k$ purely rand. bits
  - Using these bits, gen. k pairwise indep. rand. bits
  - ~~Do the co~~ $\Rightarrow \log k \times n$

* Here, we can actually do it w/ ~~$n \log k$~~ $2n$ bits

Let $Y_1, Y_2, \dots, Y_k$ be pairwise indep. n-bit strings

$$X_i = \begin{cases} 1, & \text{if } ABY_i \neq CY_i \\ 0, & \text{o.w.} \end{cases}$$

Given $AB \neq C$, $E(X_i) \geq 1/2$

Consider $X = \sum_{i=1}^{k} X_i$

$E(X) \geq k/2$

Error proba. $= Pr(X = 0)$

$$Pr(X=0) \leq Pr\left(|X - E(X)| \geq k/2\right) \leq \frac{var(X)}{(k/2)^2}$$

$$Pr(X=0) \leq \frac{4}{k^2} var(X)$$

$$var(X) = \sum_{i=1}^{k} var(X_i) + 2 \sum_{i<j} Cov(X_i, X_j) \quad (Y_i, Y_j \text{ are pairwise indep.} \Rightarrow Cov(X_i, X_j) \to 0)$$

$$= \sum_{i=1}^{k} var(X_i)$$

$$X_i = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1-p \end{cases}$$

$$var(X_i) = p(1-p) \leq 1/4$$

$$\Rightarrow Pr(X=0) \leq \frac{4}{k^2} \cdot (k \times 1/4) = 1/k \quad \Rightarrow \text{Error bound proven}$$

* Consider $\mathbb{Z}_p = \{0, 1, \dots, p-1\}, +, \cdot$ ~~a finite field~~ modulo p

If p is prime, the following properties hold.

1. associativity
2. commutativity
3. additive identity
4. multiplicative identity
5. Unique additive inverse $(\forall z \in \mathbb{Z}_p)$
6. Unique multiplicative inverse $(\forall z \in \mathbb{Z}_p \setminus \{0\})$ → p's primality is reqd here

A set w/ these operations satisfying the above properties is a field.

$\mathbb{Z}_p$ is a finite field for prime p. There also exist finite fields $\mathbb{Z}_p$ for non-prime p

17/8/23

Theorem: Let p be a prime no. (for our proof, we want $p > 2$)

choose $a, b \in_r \mathbb{Z}_p = \{0, 1, \ldots, p-1\}$

$$x_i = \underbrace{a \cdot i + b}_{\text{operations modulo } p} \quad , \; i \le p$$

$\longrightarrow$ forms a finite field

① x_i is uniformly distributed over $\mathbb{Z}_p$

② $\{x_i\}$ are pairwise ind.

① i.e., $\forall c \in \mathbb{Z}_p$

$$\Pr_{a,b \in_r \mathbb{Z}_p}[ai + b = c] = 1/p$$

Fix $a \in \mathbb{Z}_p$

$\Rightarrow b = \underbrace{c - ai}_{\text{additive inverse}}$

$\Rightarrow p$ of p^2 combos yield c

② i.e., $\forall c_1, c_2 \in \mathbb{Z}_p$,

$$\Pr_{a,b \in_r \mathbb{Z}_p}[a \cdot i + b = c_1 \wedge a \cdot j + b = c_2]$$

$\longrightarrow$ yields unique a, b

$$a = \underbrace{\frac{c_1 - c_2}{i - j}}_{\text{multiplicative inverse}}$$

$$b = \frac{c_1 j - c_2 i}{i - j}$$

$$= \Pr_{a,b \in_r \mathbb{Z}_p}[a \cdot i + b = c_1] \cdot \Pr_{a,b \in_r \mathbb{Z}_p}[a \cdot j + b = c_2]$$

$$= 1/p \cdot 1/p$$

$$= 1/p^2$$

$\Rightarrow 1/p^2$

Theorem: For every prime p & integer k, $\exists$ a unique finite field of cardinality p^k, up to isomorphism

[Galois field, $GF(p^k)$]

Construction of $GF(4)$ [smallest non-prime finite field] $(p = k = 2)$

Consider polynomials of deg. ≤ 2 over $GF(2)$,

i.e., deg. ≤ 2 & coeffs. from $GF(2)$, i.e., $\{0, 1\}$

We have 8 polynomials:

0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1
000	001	010	011	100	101	110	111

Irreducible polynomials of deg. 2 over $GF(2)$

↳ cannot be resolved to a product of lesser polynomials

~~$x^2 + x$~~ $= x(x+1)$ not irreducible

$x^2 + 1$ has a root of 1, not irreducible

In fact, $(x+1)^2 = x^2 + 1 + \underbrace{(x + x)}_{0} = x^2 + 1$

x^2+x+1 is irreducible

$GF(4) = \{0, 1, x, x+1\}$ $+, \cdot$ performed modulo $1+x+x^2$ the irreducible poly. (of deg. 2)

$x \cdot x \equiv (1+x) \bmod (1+x+x^2)$

$x \cdot (1+x) \equiv 1 \bmod (1+x+x^2)$

$(1+x) \cdot (1+x) \equiv x \bmod (1+x+x^2)$

↳ this always exists, apparently

Fact: $\forall l \geq 0$, $x^{2 \cdot 3^l} + x^{3^l} + 1$ is irreducible over $GF(2)$

Chernoff - Hoeffding bounds

$X_1, X_2, \ldots, X_n$, $X_i = \begin{cases} 1, & \text{w.p. } p_i \\ 0, & \text{w.p. } 1-p_i \end{cases}$

$X = \sum_{i=1}^{n} X_i$, X_i's are indep., $\mu = E[X]$

① $\Pr(|X-\mu| \geq t) \leq e^{-2t^2/n}$, $t > 0$

② $\Pr(X \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/3}$

③ $\Pr(X \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/2}$

} $0 \leq \delta \leq 1$

Hoeffding's extension:

$E(X_i) = \mu$

X_i's are indep. & take values in $[a,b]$,

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

18/8/23

Consider a 2-sided error algo.,

w/ $\Pr(\text{error}) \leq \frac{1}{2} - \epsilon$

To boost the success proba., we

- repeat the algo. k times
- take the majority as the ans.

$X_1, X_2, \ldots, X_k$, $X_i = \begin{cases} 1, & \text{if } i^{th} \text{ iteration answers correctly} \\ 0, & \text{o.w.} \end{cases}$

$E(X_i) \geq \frac{1}{2} + \epsilon$

$X = \sum_{i=1}^{k} X_i$ = No. of successful trials

$E(X) \geq \frac{k}{2} + k\epsilon$

Proba. of error = $\Pr(X \leq k/2)$

$\Pr(E(X) - X > k\epsilon) \leq e^{-2k^2\epsilon^2/k}$ [Chernoff bound ①]

$\leq e^{-2k\epsilon^2}$

When $\epsilon = 1/n^c$, $k = \frac{1}{\epsilon^2}\log n$ is poly(n)

choosing $k = \frac{1}{\epsilon^2}\log n$, $\leq 1/n^2$

Even $k = \frac{n}{\epsilon^2}$ is poly(n), w/ error $\leq e^{-n}$!

Moment generating fn.

$$M_X(t) = E(e^{tX}) = E\left(\sum_{i\geq 0} \frac{t^i X^i}{i!}\right)$$

$$= \sum_{i\geq 0} \frac{t^i}{i!} \underbrace{E(X^i)}_{i^{th} \text{ moment of } X}$$

Pf of Chernoff bounds:

$$\Pr(X \geq m) = \Pr(\underbrace{e^{tX}}_{\text{non-neg. r.v.}} \geq e^{tm}), \quad t > 0$$

$$\leq \frac{E(e^{tX})}{e^{tm}} \quad \text{(Markov's ineq.)}$$

$$E(e^{tX}) = E\left(e^{t\sum_{i=1}^{n} X_i}\right) = E\left(\prod_{i=1}^{n} e^{tX_i}\right)$$

$$= \prod_{i=1}^{n} E(e^{tX_i}) \quad [\text{indep r.v.s}]$$

$$E(e^{tX_i}) = p \cdot e^{t\cdot 1} + \underbrace{(1-p)}_{q} \cdot 1 = pe^t + q = p(e^t - 1) + 1$$

$$E(e^{tX}) = (pe^t + q)^n$$

$$\Pr(X \geq m) \leq \frac{(pe^t + q)^n}{e^{tm}}$$

Let $m = (p+r)n$, $\mu = np$

$$\Pr(X \geq (p+r)n) \leq \frac{(pe^t + q)^n}{e^{t(p+r)n}}$$

calculus ↓

$$\leq \left(\left(\frac{p}{p+r}\right)^{p+r} \left(\frac{q}{q-r}\right)^{q-r}\right)^n$$

strongest version of the bound but difficult to apply

Weakening & simplifying w/ approximations:

$$E(e^{tX_i}) = p(e^t - 1) + 1 \leq e^{p(e^t-1)}$$

$$E(e^{tX}) = \prod_{i=1}^{n} E(e^{tX_i}) \leq e^{np(e^t-1)} = e^{\mu(e^t-1)}$$

$$\Rightarrow \Pr(X \geq (1+\delta)\mu) \leq \frac{e^{\mu(e^t-1)}}{e^{t\mu(1+\delta)}} \rightarrow \text{Minimized for } t = \ln(1+\delta)$$

$$\leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

calculus/analysis ↓

$$\leq e^{-\delta^2\mu/3}, \quad \delta < 1$$

* K servers, n jobs

Allocate jobs to servers in a decentralized way

simple way: Allocate each job to a server to a server uniformly at random

Expected load = n/k ⟶ "balls-in-bins"

21/8/23

Balls & bins

m balls, n bins

Throw the balls into the bins u.a.r.

Coupon collector prob. - what is the val. of m s.t. every bin contains at least one ball?

Hashing - How large should n be compared to m s.t. no bin has a lot of balls?

Birthday problem

what should m be so that the proba. that ∃ a bin w/ more than 1 ball is at least ½?

B_i - the event that the i^{th} ball lands in a bin of its own

we want $\Pr\left(\bigcap_{i=1}^{m} B_i\right)$.

$\Pr\left(\bigcap_{i=1}^{m} B_i\right)$ ⟶ we want $\Pr(B_i \mid B_{i-1}, \dots, B_1) = \frac{n-i+1}{n}$

$$= \prod_{i=1}^{m} \Pr(B_i \mid B_{i-1}, \dots, B_1) = \prod_{i=1}^{m} \frac{n-i+1}{n} = \prod_{i=1}^{m} \left(1 - \frac{i-1}{n}\right)$$

$$\leq \prod_{i=1}^{m} e^{-(i-1)/n} = e^{-\sum_{i=1}^{m} \frac{i-1}{n}} = e^{-m(m-1)/2n} < \frac{1}{2}$$

$$\Rightarrow m = \Theta(\sqrt{n})$$

For hashing, this tells that a hash table of size ~~smaller than~~ $o(n^2)$ is likely to find collisions (w.p. > ½)

E_i = event that the i^{th} ball doesn't land in ~~one of t.~~ a separate bin given that the first $i-1$ balls landed in separate bins

$\Pr(E_i) = \frac{i-1}{n}$ $\quad \Pr\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} \frac{i-1}{n} = \frac{m(m-1)}{2n}$

If $m < \sqrt{n}$, $\Pr\left(\exists \text{ a ball that doesn't land in a separate bin}\right) < ½$

If n balls are thrown into n bins, what is the max. load? ⊖ $\frac{3\log n}{\log\log n}$ w.p. $1 - 1/n$

max. no. of balls in some bin

Thm: If n balls are thrown into n bins u.a.r., then w.p. $\to 1 - 1/n$, the max. load is at most $\frac{3\log n}{\log\log n}$

Pf: Fix a bin i.

$$\Pr(\text{bin } i \text{ has } k \text{ balls}) \le \binom{n}{k}\left(\frac{1}{n}\right)^k$$

★ $\binom{n}{k} \le \left(\frac{ne}{k}\right)^k$

$$\le \left(\frac{ne}{k}\cdot\frac{1}{n}\right)^k = \left(\frac{e}{k}\right)^k$$

$$\Pr(\exists \text{ bin } i \text{ s.t. bin } i \text{ has } k \text{ balls}) \le n\left(\frac{e}{k}\right)^k \quad \text{(Union bound)}$$

$$\le 1/n \quad \text{(reqd.)}$$

what val. of k do we choose?

we can substitute $\frac{3\log n}{\log\log n}$ to verify it works.

$X_1, X_2, \ldots, X_n$ – r.v.s corresponding to no. of balls in individual bins

dependent, their sum is fixed → something indep would be nicer to work with

$$X_i \sim \text{Bin}(m, 1/n)$$

$$\Pr(X_i = k) = \binom{m}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{m-k}$$

$$= \frac{m!}{k!\,(m-k)!}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{m-k}$$

$$= \frac{(m-k+1)\cdots(m)}{k!\;n^k}\left(1-\frac{1}{n}\right)^{m-k}$$

$$\approx \left(\frac{m}{n}\right)^k\left(\frac{1}{k!}\right)e^{-m/n} \quad \text{(assuming } m, n \gg k\text{)}$$

$\lambda = m/n$

Poisson dbn.

$X \sim \text{Poi}(\lambda)$

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad E(X) = \lambda$$

Thm: If $X \sim \text{Bin}(n, p)$ & $\lim_{n\to\infty} np = \lambda$ (indep. of n), for every fixed k, $\lim_{n\to\infty} \Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$

23/6/23

$X_1 \sim \text{Poi}(\lambda_1)$, $X_2 \sim \text{Poi}(\lambda_2)$ [indep.]

then $X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$

Pf: $\Pr(X_1 + X_2 = j) = \sum_{k=0}^{j} \Pr(X_1 = k) \cdot \Pr(X_2 = j-k)$

$$= \sum_{k=0}^{j} \frac{e^{-\lambda_1} \cdot \lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2} \lambda_2^{j-k}}{(j-k)!}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{j!} \sum_{k=0}^{j} \lambda_1^k \lambda_2^{j-k} \frac{j!}{k!(j-k)!}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)} (\lambda_1+\lambda_2)^j}{j!}$$

Chernoff bounds for Poisson r.v.s

$X \sim \text{Poi}(\lambda)$

$$E(e^{tX}) = \sum_{k=0}^{\infty} \Pr(X=k) \cdot e^{tk} = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^k}{k!} \cdot e^{tk}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{e^{tk} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t \lambda)^k}{k!} \cdot e^{-\lambda e^t} \cdot e^{\lambda e^t}$$

$$= e^{\lambda(e^t-1)} \sum_{k=0}^{\infty} \frac{e^{-\lambda e^t} (e^t \lambda)^k}{k!}$$

$$= e^{\lambda(e^t-1)}$$

Let $x \geq \lambda$.

$$\Pr(X \geq x) = \Pr(e^{tX} \geq e^{tx})$$

$$\leq \frac{e^{\lambda(e^t-1)}}{e^{tx}}$$ Find t that minimizes this. $t = \log x/\lambda$

$$\leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}$$

Let $x \leq \lambda$

$$\Pr(X \leq x) = \Pr(e^{tX} \geq e^{tx}), \quad t < 0$$

$$\leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}$$

Thm: Let $\{X_i^m\}_{1\le i\le n}$ be the balls-in-bins dbn. w/ m balls & n bins

Let $\{Y_i^m\}_{1\le i\le n}$ be indep. ~~Poisson r.v.s w/~~ $\sim Poi(m/n)$

The dbn. $\{X_i^m\}_{1\le i\le n}$ conditioned on $\sum_{i=1}^n Y_i^m = k$

is identical to $\{X_i^k\}_{1\le i\le n}$

Pf: $\Pr(\bigcap_{i=1}^n (X_i = k_i))$ [multinomial dbn.] $(\sum_{i=1}^n k_i = k)$

$$= \binom{k}{k_1}\cdot\binom{k-k_1}{k_2}\cdot\cdots\cdot\binom{k-k_1-k_2-\cdots-k_{n-1}}{k_n}\cdot\left(\frac{1}{n}\right)^{k_1}\cdot\left(\frac{1}{n}\right)^{k_2}\cdot\cdots\cdot\left(\frac{1}{n}\right)^{k_n}$$

$$= \frac{k!}{k_1!\,k_2!\cdots k_n!}\left(\frac{1}{n}\right)^k$$

$$\Pr\left(\bigcap_{i=1}^n (Y_i^m = k_i) \,\middle|\, \sum_{i=1}^n Y_i^m = k\right) = \frac{\Pr\left[\bigcap_{i=1}^n (Y_i^m = k_i)\right]\cdot \Pr\left[\sum_{i=1}^n Y_i^m = k \,\middle|\, \bigcap_{i=1}^n (Y_i^m = k_i)\right]}{\Pr\left(\sum_{i=1}^n Y_i^m = k\right)}$$

indep. $\Rightarrow$ prod. ; $\sum Y_i^m = k$ is determined by $\bigcap (Y_i^m = k_i)$, so that factor $= 1$

↳ sum of Poisson $\Rightarrow Poi(m)$

$$= \frac{\prod_{i=1}^n \Pr(Y_i^m = k_i)}{\frac{e^{-m}m^k}{k!}} = \left(\prod_{i=1}^n \frac{e^{-m/n}\cdot\left(\frac{m}{n}\right)^{k_i}}{k_i!}\right)\cdot\frac{e^m k!}{m^k}$$

$$= \frac{e^{-m}\left(\frac{m}{n}\right)^k}{k_1!\,k_2!\cdots k_n!}\cdot\frac{e^m k!}{m^k} = \frac{k!}{k_1!\,k_2!\cdots k_n!}\left(\frac{1}{n}\right)^k$$

Thm: $\{X_i^m\}_{1\le i\le n}$, balls-in-bins dbn. w/ m balls & n bins

$\{Y_i^m\}_{1\le i\le n}$, indep. $Poi(m/n)$

Let f be a non-negative fn.

$$\mathbb{E}(f(X_1^m, X_2^m, \ldots, X_n^m)) \le e\sqrt{m}\,\underbrace{\mathbb{E}(f(Y_1^m, Y_2^m, \ldots, Y_n^m))}_{\text{no conditioning!}}$$

How to apply this?

Suppose we want to answer w/ what prob. no. of bins with no balls is $\ge k$

Define $f(x_1,\ldots,x_n) = \begin{cases} 1, \text{ if } |\{i : x_i = 0\}| \ge k \\ 0, \text{ o.w.} \end{cases}$

$\mathbb{E}(f(X_1^m,\ldots,X_n^m)) = \Pr(\text{no. of bins w/ no balls is} \ge k)$

24/8/23

Pf: $\mathbb{E}(f(Y_1^m, Y_2^m, \ldots, Y_n^m)) = \sum_{k\ge 0} \mathbb{E}(f(Y_1^m,\ldots,Y_n^m) \mid \sum_{i=1}^n Y_i^m = k)\cdot \Pr(\sum_{i=1}^n Y_i^m = k)$

$$\ge \mathbb{E}(f(Y_1^m,\ldots,Y_n^m) \mid \sum_{i=1}^n Y_i^m = m)\cdot \Pr(\sum_{i=1}^n Y_i^m = m)$$

$\Pr(\sum_{i=1}^n Y_i^m = m)$: $\sum Y_i^m \sim Poi(m)$

$$= \mathbb{E}(f(Y_1^m,\ldots,Y_n^m))\cdot\frac{e^{-m}m^m}{m!}$$

$$\mathbb{E}(f(Y_1^m, \dots, Y_n^m)) \geq \mathbb{E}(f(X_1^m, \dots, X_n^m)) \cdot \frac{e^{-m} \cdot m^m}{\sqrt{2\pi m}} \cdot \left(\frac{e}{m}\right)^m$$

$$\mathbb{E}(f(X_1^m, \dots, X_n^m)) \leq e\sqrt{m}\, \mathbb{E}(f(Y_1^m, \dots, Y_n^m))$$

$[e \leq \sqrt{2\pi}]$

Stirling's approx.
$m! \leq \sqrt{2\pi m}\left(\frac{m}{e}\right)^m$

Thm: If n balls are thrown into n bins u.a.r., then $\exists$ a bin containing $\geq \Omega\left(\frac{\log n}{\log\log n}\right)$ balls w.p. $\geq 1 - \frac{1}{n}$

Pf: $f(x_1, \dots, x_n) = \begin{cases} 1, & \text{if } x_i \leq k \ \forall\, i \in \{1, \dots, n\} \equiv \max\{x_1, \dots, x_n\} \leq k \\ 0, & \text{o.w.} \end{cases}$

$$\mathbb{E}(f(X_1^m, \dots, X_n^m)) = \Pr(\text{Every bin has} \leq k \text{ balls})$$

$$\Pr\left(\bigcap_{i=1}^n (Y_i^m \leq k)\right) = \left(\Pr(Y_i^n \leq k)\right)^n = \left(1 - \Pr(Y_i^n \geq k)\right)^n$$

$$\Pr(Y_i^n \geq k) \geq \Pr(Y_i^n = k) = \frac{e^{-1} \cdot 1^k}{k!} = \frac{1}{ek!}$$

($Y_i^n \hookrightarrow \text{Poi}(1)$)

$$\Pr\left(\bigcap_{i=1}^n (Y_i^n < k)\right) = \left(1 - \frac{1}{ek!}\right)^n \leq e^{\frac{-n}{ek!}}$$

$$\Pr(\text{Every bin has} < k \text{ balls}) \leq e\sqrt{n} \cdot e^{-n/ek!}$$

$$< \frac{1}{n} \text{ (req'd.)}$$

Verify that $k = \Omega\left(\frac{\log n}{\log\log n}\right)$ works

Static dictionary

Universe $\mathcal{U}$, set $S \subseteq \mathcal{U}$

S is static, i.e., known at the start & fixed

$m = |S| \ll |\mathcal{U}| = M$ → $O(|S|)$

we wish to compactly represent S so that membership queries ($x \overset{?}{\in} S$) can be answered efficiently → $O(1)$ [expected or worst-case]

$\log(1 - e^{-y}) - y + \frac{y}{e^y - 1}$ $\log(e^y - 1) - \frac{ye^y}{e^y - 1}$ $\log e^y - \log(e^y - 1) = ye^y$ $e^y = 1/2$ $y = -\log 2$

Bloom filter

Bit-array $A[1,2,\ldots,n]$ (i.e., size is n)

Choose k fns. $h_1,\ldots,h_k$ u.a.r. [assumes perfectly rand. hash fns.]

$\forall x \in S$, set $A[h_i(x)] = 1 \ \forall \ i \in \{1,\ldots,k\}$

i.e., $h_j(x) = i$ w.p. n^{-1}

Membership query resolution: Given x, check $A[h_i(x)] \ \forall \ i \in \{1,\ldots,k\}$

If all 1's, say $x \in S$

$x \in S \Rightarrow$ correct result guaranteed

$x \notin S \Rightarrow$ false positives may occur

Given $x \notin S$,

$\Pr[\text{we answer } x \in S] \leq \delta$

$\Pr[A(j) = 0] = \underbrace{\left(1 - \frac{1}{n}\right)^{mk}}_{=p} \approx e^{-mk/n}$ ← overall h_i's, k & x, m

For given $x \in S$ & h_i, $\Pr(h_i(x) = j) = \frac{1}{n}$

$A[j]$ & $A[j']$ are dependent (balls-in-bins w/ km balls & n bins)

$\Pr(Y_i \sim \text{Poi}(mk/n) = 0)$

25/8/23

What should be n, k so that proba. for false positives is $\leq \delta$?

Expected no. of empty posns. in the Bloom filter $= np$

Given $x \notin S$, assuming p proba. of bin being empty (frac. of empty bins), ↳ assumes independence

$\Pr(\text{Answer Yes}) = (1-p)^k$

$Y_i \sim \text{Poi}(km/n)$, $\Pr(Y_i = 0) = e^{-km/n}$

$Z_i = \begin{cases} 1, & \text{if } Y_i = 0 \\ 0, & \text{o.w.} \end{cases}$ $\quad Z_i \sim \text{Ber}(e^{-km/n})$

$Z = \sum_{i=1}^{n} Z_i \approx$ No. of ~~empty~~ posns. in A that are 0

Just to say they are close i.e., a good bound

$\Pr[|Z - E(Z)| \geq \epsilon n] \leq 2e^{-n\epsilon^2 e^{mk/n}/3}$, $[\delta = \epsilon n/\mu]$

With $f(x)$ as $|x - E(z)| \geq \epsilon n \Rightarrow 1$; 0 o.w.,

$\Pr[|X - \underbrace{E(z)}_{ne^{-km/n}}| \geq \epsilon n] \leq 2e\sqrt{km}\, e^{-n\epsilon^2 e^{mk/n}/3}$

$\Pr(\text{false positives}) \approx \left(1 - e^{-km/n}\right)^k \xrightarrow{\text{min. for } k = (\ln 2)\frac{n}{m}} = \left(\frac{1}{2}\right)^k \leq \delta$

$k = \log(1/\delta) \Rightarrow n = m\log(1/\delta)/\ln 2$

$\frac{k\log(1 - e^{-km/n})}{\log(1 - e^{-km/n}) + k\left(e^{-km/n}\right)\cdot\frac{m}{n}}$

Universal hash families (Carter-Wegman)

k-universal hash family:

Family of fns $\mathcal{H} = \{h : U \longrightarrow [n]\}$ s.t.

$\forall\ x_1, x_2, \ldots, x_k$ (all different),

$$\Pr_{h \in_r \mathcal{H}}\left[h(x_1) = h(x_2) = \cdots = h(x_k)\right] \leq \frac{1}{n^{k-1}}$$

If $\mathcal{H}$ = set of all fns. from U to $[n]$, $(|\mathcal{H}| = n^{|U|})$

$$\Pr_{h \in_r \mathcal{H}}\left[h(x_1) = \cdots = h(x_k)\right] = \frac{1}{n^{k-1}} \quad \forall k$$

(Ideally, we want $|\mathcal{H}| = \text{poly}(|U|)$)

k-wise indep. hash family: [strongly k-universal]

$\mathcal{H} = \{h : U \longrightarrow [n]\}$ s.t.

$\forall\ x_1, x_2, \ldots, x_k$ (distinct), $\forall\ y_1, y_2, \ldots, y_k \in \{1, \ldots, n\}$

$$\Pr_{h \in_r \mathcal{H}}\left[\bigcap_{i=1}^{k} (h(x_i) = y_i)\right] = \frac{1}{n^k}$$

$\Downarrow$

$\forall\ x_1, x_2, \ldots, x_k$ (distinct),

$$\Pr_{h \in_r \mathcal{H}}\left[h(x_1) = h(x_2) = \cdots = h(x_k)\right] = \frac{1}{n^{k-1}}$$

strongly k-universal $\Rightarrow$ k-universal

28/8/23

strongly k-universal $\equiv$ uniformity, i.e.,

① $\Pr_{h \in_r \mathcal{H}}[h(x) = y] = 1/n$ &

② $\Pr_{h \in_r \mathcal{H}}\left[\bigcap_{i=1}^{k} h_{x_i}(x_i) = y_i\right] = \prod_{i=1}^{k} \Pr(h_{x_i}(x_i) = y_i)$

* m balls into n bins using a 2-universal hash family

$\mathcal{H} = \{h : [m] \longrightarrow [n]\}$

i.e, choose $h \in_r \mathcal{H}$ & assign balls to bins using h

For any 2 balls i & j, let

$$X_{ij} = \begin{cases} 1, & \text{if } i \text{ \& } j \text{ land in the same bin} \\ 0, & \text{o.w.} \end{cases}$$

$$\Rightarrow \Pr(X_{ij} = 1) = \Pr_{h \in_r \mathcal{H}}[h(i) = h(j)] \leq 1/n$$

Total no. of collisions, $X = \sum_{i<j} X_{ij}$

~~collision~~ collision ~~collission~~ colision

$$E(X) \leq \binom{m}{2} \cdot \frac{1}{n} \leq \frac{m^2}{2n}$$

Let Y be the max. load

$X \geq \binom{Y}{2}$ = no. of collisions in the bin w/ max. load

$\Pr\left(X \geq \frac{m^2}{n}\right) \leq 1/2$ (Markov's ineq.)

$\Pr\left(\binom{Y}{2} \geq \frac{m^2}{n}\right) \leq 1/2$ (since $\binom{Y}{2} \leq X$)

$\Pr\left(Y \geq 1 + m\sqrt{\frac{2}{n}}\right) \leq 1/2$ $\left[\binom{Y}{2} \geq \frac{(Y-1)^2}{2}\right]$

If $n = \Theta(m^2)$, then every bin has $\Theta(1)$ balls w.p. $\geq 1/2$

for a static dict. on $S \subseteq U$, $[|S| = m]$

we can obtain a good hash fn. of size $\Theta(m^2)$

- choose a hash fn from a 2-universal hash family
- if no. of collisions is too high, resample

[Expected number of times to sample is 2 from our bound]

* Dynamic dictionary over U

- additions
- deletions
- query

(using H, a 2-universal hash family)

[we want $n = O(m)$ ideally ⇒ FKS (Fredman-Komlós-Szemerédi) hashing]

Set of all additions, $S \subseteq U$

$x \in U$, $h \in H$

No. of collisions of x w/ elements in S using hash fn. h,

$$\text{coll}(x, S, h) = |\{y \in S \mid h(x) = h(y)\}|$$

$$\mathop{E}_{h \in H}[\text{coll}(x, S, h)] = \frac{1}{|H|} \sum_{h \in H} \text{coll}(x, S, h)$$

$$= \frac{1}{|H|} \sum_{h \in H} \sum_{y \in S} 1_{[h(x)=h(y)]}$$ ← indicator fn

$$= \sum_{y \in S} \underbrace{\frac{1}{|H|} \sum_{h \in H} 1_{[h(x)=h(y)]}}_{\Pr_{h \in H}[h(x)=h(y)]}$$

$$\leq \sum_{y \in S} \frac{1}{n} = \frac{|S|}{n}$$ [$n \to$ size of hash table]

[If $n = \Theta(|S|)$, then $O(1)$ running time for query, but hash table could be sparse at times]

Explicit 2-universal family

U, $|U| = m \longrightarrow n$ sized table

Prime $p > m$, $p = \Theta(m)$ [$\exists$ prime p b/w m & $2m$]

$$\mathcal{H} = \{ h_{a,b} \mid 1 \le a \le p-1,\ 0 \le b \le p-1 \}$$

$$h_{a,b}(x) = ((ax+b) \bmod p) \bmod n$$

$$|\mathcal{H}| = \Theta(p^2)$$

$\Pr_{h \in_r \mathcal{H}} [h(x) = h(y)] \le 1/n$

$$ax + b = z_1 \pmod p$$
$$ay + b = z_2 \pmod p$$
$$z_1 = z_2 \pmod n$$

30/8/23

Thm: $\mathcal{H}$ is 2-universal

Pf: To show: $\Pr_{h \in \mathcal{H}} [h(x) = h(y)] \le 1/n$

$$ax + b = z_1 \pmod p$$
$$ay + b = z_2 \pmod p$$
$$z_1 = z_2 \pmod n$$

For every $z_1 \neq z_2$, $\exists$ unique soln. for a, b

$$|\{ z_2 \mid z_2 \neq z_1,\ z_1 = z_2 \pmod n \}| \le \frac{p}{n} - 1$$

$$\Rightarrow \Pr_{h \in \mathcal{H}} [h(x) = h(y)] \le \frac{p(\frac{p}{n} - 1)}{p(p-1)} = \frac{1}{n} \cdot \frac{(p-n)}{(p-1)}$$

$$\le 1/n$$

[same idea as pairwise indep. strings]

Pairwise-indep hash family

$|U| = 2^m \longrightarrow 2^n$

Consider the field $GF(2^m)$

$$\mathcal{H} = \{ h_{a,b} \mid a, b \in GF(2^m) \}$$

$h_{a,b}(x) = (ax + b)$ in $GF(2^m)$ → The "pairwise" comes from here. Fix y ($k-1$) $\Rightarrow$ k-wise indep.

m-bit string $\Longrightarrow$ truncate to n bits

Exercise: verify this is pairwise indep.

Perfect hash family [static dicts again]

A family $\mathcal{H} = \{h : \mathcal{U} \to [n]\}$ is perfect for sets of size $\leq n$ if $\forall S \subseteq \mathcal{U}, |S| \leq n$,

$\exists h \in \mathcal{H}$ s.t. $\forall x \neq y, h(x) \neq h(y)$

[Any n-sized set in $\mathcal{U}$ can be perfectly hashed by a fn. in $\mathcal{H}$]

$\Rightarrow$ static dictionary has an $O(n)$ data structure with $O(1)$ query time

↳ Is $A[h(x)] = x$?

Each cell of the table = $O(\log|\mathcal{U}|)$

we also want to store each h in a constant no. of cells & to index $\mathcal{H}$, we need $\log|\mathcal{H}|$ bits

$\Rightarrow \log|\mathcal{H}| = O(1) \cdot \log|\mathcal{U}| \Rightarrow |\mathcal{H}| = |\mathcal{U}|^{O(1)}$

* If $\mathcal{H} = \{h : \mathcal{U} \to [n]\}$ is a perfect hash family for sets of size n, then

$$|\mathcal{H}| = 2^{\Omega(n)}$$

$|\mathcal{U}| = n^2 \to [n]$ cannot be perfectly has

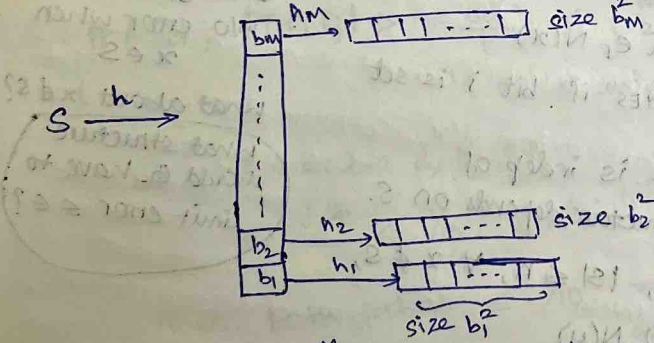
FKS hashing (Fredman - Komlos - Szemeredi, '84) [Cell probe model]

$\mathcal{U} \to [n]$, sets S, $|S| = m$

Choose $h \in \mathcal{H}$ (2-universal)

s.t. no. of collisions $\leq m$

Handle collisions using another level of hashing (2-universal again)



(Further collisions are chained, since only constant collisions now)

$$\text{Size} = m + \sum_{i=1}^{m} b_i^2$$

$$= O(m)$$

$$\sum_{i=1}^{m} \binom{b_i}{2} \leq m$$

(No. of collisions)

Bit probe model

Size, query are now in terms of bits rather than cells

FKS: $m \log|\mathcal{U}|$ $\log|\mathcal{U}|$

What should be n if we limit ourself to 1 bit query?

Buhrman, Miltersen, Radhakrishnan, Venkatesh, '00

$\Rightarrow$ size $= O\left(\frac{M}{\epsilon} \log |\mathcal{U}|\right)$ $\longrightarrow$ comparable to FKS result

1-bit query w/ 2-sided error of $\leq \epsilon$

If we restrict 1-sided error of $\leq \epsilon$, we show this

$$\text{size} = O\left(\frac{M^2}{\epsilon^2} \log |\mathcal{U}|\right)$$

↳ can show existence but can't obtain the data structure

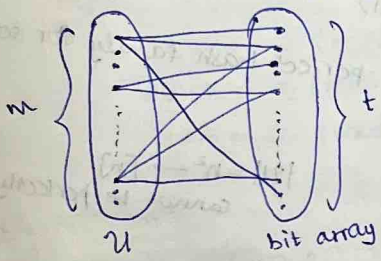
If we want to efficiently find the structure,

$$\text{size} = O\left(\frac{M^2}{\epsilon^2} \log^2 |\mathcal{U}|\right)$$

31/8/23

Change of notation for the proof: $|\mathcal{U}| = M$, $|S| = n$

1-query algo. $\longrightarrow$ bipartite graph



$x \overset{?}{\in} S$: Randomly choose $i \in_R [t]$ using the graph say, [edge wts. like proba. of choosing]

Given $G(\mathcal{U}, [t], E)$,

$S \subseteq \mathcal{U}$, $|S| \leq n$,

define $\forall x$, $N(x) = \{i \mid (x,i) \in E\}$

set all the bits in $\bigcup_{x \in S} N(x)$

query: $x \overset{?}{\in} S$

- choose $i \in_R N(x)$
- Answer "YES" if bit i is set

clearly, this works w/o error when $x \in S$

What about $x \notin S$?

What structure should G have to limit error $\leq \epsilon$?

Note that the graphs is indep. of S. Only the data structure depends on S.

We want: $\forall S \subseteq \mathcal{U}$, $|S| \leq n$, $\forall x \notin S$,

for $N(S) = \bigcup_{y \in S} N(y)$,

$$|N(x) \cap N(S)| \leq \epsilon |N(x)|$$

We show a stronger result:

Such graphs exist

$$\begin{cases} \forall x, |N(x)| = d & [d\text{-regular over } \mathcal{U}] \\ \forall x \neq y, |N(x) \cap N(y)| \leq \frac{\epsilon d}{n} & [\text{this implies the above}] \end{cases}$$

An (m, t, d, r)-combinatorial design is a family of sets

$T_1, T_2, \ldots, T_m$ s.t. $\forall i \; \forall j, i \neq j$

① $T_i \subseteq [t]$

② $|T_i| = d$

③ $|T_i \cap T_j| \leq r$

In our case,

$T_i = N(x_i)$, $x_1, x_2, \ldots, x_m \in \{0,1\}^n$

$r = \frac{\epsilon d}{n}$

We consider a random collection of sets & show that w.p. >0, the properties are satisfied, implying existence.

Pf: $\forall j \in [m]$, $\forall i \in [t]$,
put i in T_j w.p. $\frac{2d}{t}$ independently

$\Rightarrow E(|T_j|) = 2d$

$Pr(|T_j| < d) \leq e^{-2d(\frac{1}{2})^2/2} = e^{-d/4}$ [chernoff bound]

$\Rightarrow Pr(\exists j \; |T_j| < d) \leq m e^{-d/4}$ [Union bound]

< 1 [we want this, otherwise no set can have size d]

$E(|T_i \cap T_j|) = t\left(\frac{2d}{t}\right)^2$ $\left[\sum_{k \in [t]} E(|T_i \cap T_j| \mid k \in T_i) \cdot Pr(k \in T_i)\right]$ (check)

$= \frac{4d^2}{t}$

$Pr\left(|T_i \cap T_j| > \frac{8d^2}{t}\right) \leq e^{-\frac{4d^2}{3t}}$ [chernoff bound].

$Pr\left(\exists i,j \; |T_i \cap T_j| > r = \frac{8d^2}{t}\right) \leq \binom{m}{2} e^{-4d^2/3t}$

$\leq m^2 e^{-4d^2/3t}$

We find suitable d, t s.t.

$m e^{-d/4} \leq 1/10$ & $m^2 e^{-4d^2/3t} \leq 1/10$

Then, both probas $\leq 1/10$ $\Rightarrow Pr(\text{not satisfied}) \leq 1/5$

$\Rightarrow Pr(\text{satisfied}) \geq 4/5 > 0$

$r = \frac{\epsilon d}{n} = \frac{8d^2}{t} \Rightarrow t = \frac{8dn}{\epsilon}$

This is satisfied when

$d = \Theta\left(\frac{n}{\epsilon} \log m\right)$ & $t = \Theta\left(\frac{n^2}{\epsilon^2} \log m\right)$

Hence, proved

$r = \Theta(\log m)$

1/9/23

PS1 discussion

(1) Generate n distinct nos. & sort them

Assign each i the sorted posn. of k_i

$k_1, k_2, \ldots, k_n$

choose using $3\log n$ bits each

$\Rightarrow \Pr(k_i = k_j) \leq \frac{1}{n^3}$

$\therefore \Rightarrow \Theta(n \log n)$

$\Rightarrow \Pr(\exists i,j \text{ s.t. } k_i = k_j) \leq \frac{1}{n}$

(4) Intuitively, X increments every 2^x times, so in a sense, it appears to be counting the no. of bits in the representation of n

Exercise: Find variance, $\text{Var}(y)$ & show

n can be rep. ~~using~~ as $(1 \pm \epsilon)n$

using $\frac{1}{\epsilon^2} \log\log n$ bits

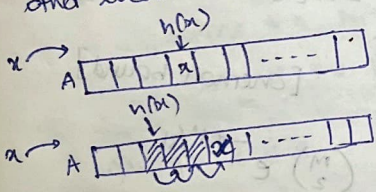
(Morris counter)

↳ way to ~~store~~ count n times using $\log\log n$ bits

4/9/23

Open addressing (Linear probing)

Our previous hashing schemes (such as FKS) handled collisions using an auxiliary data structure. This scheme maps collisions into some other location of the primary data structure.



If $A[h(x)]$ is empty

If $A[h(x)]$ contains some other element, probe linearly for an empty position

~~Addition~~: Insertion: Starting from $h(x)$, insert x at the first empty slot

Membership: Starting from $h(x)$, continue until you find x or an empty slot

[doesn't work when there are deletions, in which case we set flags/tombstones at posns. of deleted elements]

Works very well in practice.

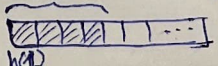
Knuth '63: If h is completely random, then expected time is $O(1)$.

2010-11: 5-wise independence is necessary & sufficient for expected $O(1)$ time.

"Run" A run is a maximal contiguous seq. of cells that are filled

Given $i \in S$, what is the length of the run containing $h(i)$?

Say, 1



Consider a k-wise indep. hash family $\mathcal{H}$, $h \in_r \mathcal{H}$

~~For the given i,~~ consider the interval of posⁿs in the hash table.

$$I_j = \left[h(1) - (2^j - 1),\ h(1) + (2^j - 1) \right] \quad \text{(dyadic intervals)}$$

$|S| = n$, hash table has size t

Expected no. of elements in $I_j = \frac{|I_j| \cdot n}{t}$

$$X_i = \begin{cases} 1, & \text{if } i \in I_j \\ 0, & \text{o.w.} \end{cases}$$

$$X = \sum_{i \in S \setminus \{1\}} X_i \quad \Rightarrow E(X) = \frac{n|I_j|}{t}$$

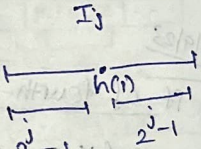
If $t = 8n$, $|I_j| = 2^{j+1} - 1 \Rightarrow E(X) \le 2^{j-2}$

Consider R, a run containing 1

$$E(|R|) = \sum_{l=0}^{n} l \cdot \Pr(|R| = l)$$

$$\le \sum_{j=0}^{\log t} 2^j \cdot \Pr\left(\underbrace{2^{j-1} \le |R| \le 2^j}_{\text{For these, } l \le 2^j} \right)$$

If $|R| \ge 2^{j-1}$, no. of elements in I_j is $\ge 2^{j-1}$



① Assume h is random $\Rightarrow X$ is a sum of Bernoulli r.v.s

$$\Pr(|R| \ge 2^{j-1}) \le \Pr\left(X \ge \underbrace{2^{j-1}}_{2E(X)}\right)$$

$$\le e^{-2^{j-2}/2}$$

$$\Rightarrow E(|R|) \le \sum_{j=0}^{\log t} \frac{2^j}{e^{2^j}} = O(1)$$

② Suppose h is picked from a 3-wise indep. hash family

Given $h(1)$, X_i's are pairwise indep.

$$\Rightarrow \text{var}(X) = \sum_{i \in S \setminus \{1\}} \text{Var}(X_i) = \sum_{i \in S \setminus \{1\}} \underbrace{E(X_i^2)}_{= E(X_i)} - [E(X_i)]^2 \quad \text{(Bernoulli)}$$

$$\le \sum_{i \in S \setminus \{1\}} E(X_i)$$

$$\Pr(|R| \ge 2^{j-1}) \le \Pr\left(X \ge \underbrace{2^{j-1}}_{2E(X) \le 2\text{var}(X)} \right) \le \frac{1}{\text{Var}(X)}$$

$$\le \frac{1}{2^{j-2}}$$

$$\Rightarrow E(|R|) \le \sum_{j=0}^{\log t} \frac{2^j}{2^{j-2}} = O(\log n)$$

Fourth-moment bound

Let $X_1, X_2, \ldots, X_n$ be 4-wise indep. s.t. ~~$E(X_i) = p$~~ $X_i \sim \text{Ber}(p)$

$$X = \sum_{i=1}^{n} X_i \Rightarrow E(X) = np = \mu$$

For $\mu \geq 1$, $\beta > \mu$

$$\Pr(X > \mu + \beta) \leq \frac{4\mu^2}{\beta^4}$$

---

(3) h is picked from a 5-wise indep. family

$\Rightarrow$ Given $h(1)$, X_i's are 4-wise indep

$$\Pr(|R| \leq 2^{j-1}) \leq \Pr(X \geq 2^{j-1})$$

$$= \Pr(X \geq 2^{j-2} + 2^{j-2})$$

$$\leq \frac{4}{\mu^2} = \frac{32}{2^{2j}}$$

$$E(|R|) \leq \sum_{j=0}^{\log t} \frac{2^j \cdot 32}{2^{2j}} = 32 \sum_{j=0}^{\log t} \frac{1}{2^j} \leq = \Theta(1)$$

6/9/23

Pf. of fourth-moment bound:

$$\Pr(X \geq \mu + \beta) = \Pr(X - \mu \geq \beta) = \Pr((X-\mu)^4 \geq \beta^4)$$

$$\leq \frac{E[(X-\mu)^4]}{\beta^4}$$

Define $Y_i = X_i - p$ Since X_i's are 4-wise indep., so are Y_i's

$$E(Y_i) = 0$$

$$E[(X-\mu)^4] = E\left[\left(\sum_{i=1}^{n}(X_i - p)\right)^4\right] = E\left[\left(\sum_{i=1}^{n} Y_i\right)^4\right]$$

$$= E\left[\sum_{i=1}^{n} Y_i^4 + \binom{4}{2}\sum_{i<j} Y_i^2 Y_j^2 + \binom{4}{3}\sum_{i\neq j} Y_i Y_j^3 + \sum_{i\neq j\neq k\neq l} Y_i Y_j Y_k Y_l + \binom{4}{2}\binom{2}{1}\sum_{i\neq j\neq k} Y_i Y_j Y_k^2\right]$$

$$= \sum_{i=1}^{n} E(Y_i^4) + \binom{4}{2}\sum_{i<j} E(Y_i^2)E(Y_j^2)$$ [others split & go to 0 due to 4-wise indep.]

$$E(Y_i^4) = p(1-p)^4 + (1-p)(0-p)^4$$

$$E(Y_i^2) = p(1-p)^2 + (1-p)p^2$$

$$E[(X-\mu)^4] \leq 4np^2 = 4\mu^2 \quad \text{(verify)}$$ Hence, proved.

Cuckoo hashing (Pagh, Rödler – '01)

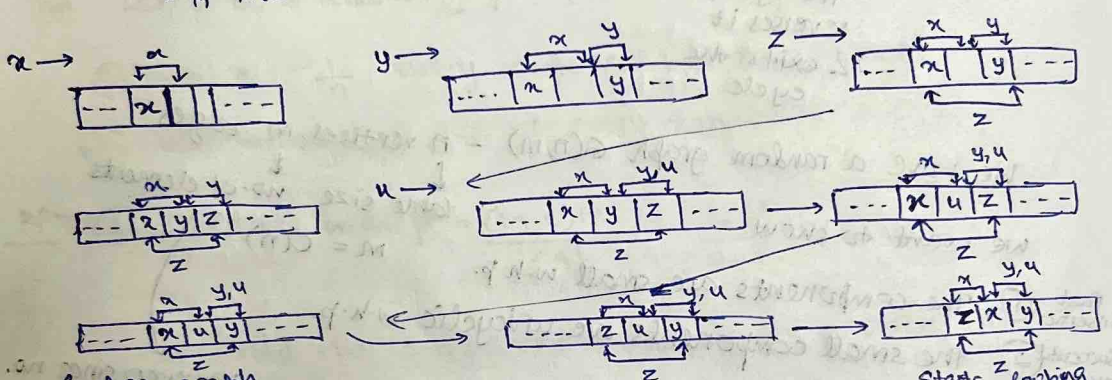
T

2 hash fns. h_1, h_2 (completely random)

* Every x is hashed to either $h_1(x)$ or $h_2(x)$
  ⇒ search for x in $O(1)$ – worst case

Insertions are slightly harder:

Insert (x):
- If one of $h_1(x), h_2(x)$ is empty, then insert in T at the empty loc.
- If both $h_1(x)$ & $h_2(x)$ are occupied,
  then $h_i(y) = h_1$ for some $y \neq x$, $i \in \{1,2\}$,
  place x in $h_1(x)$ & recursively try to insert y
- If this continues indefinitely, choose new h_1', h_2' & rehash the table



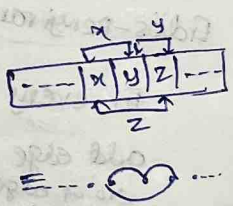
Cuckoo graph

$G(V, E)$

$V \rightarrow$ set of posns. in the hash table

$\forall x$ in the table, add edge $(h_1(x), h_2(x))$

- self loops & multi edges are possible

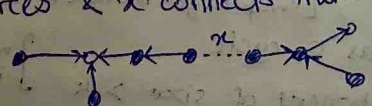


Obs: Let C be a connected component in the cuckoo graph, and x is being inserted. If the new component has >1 cycle, then insertion is impossible. [without rehashing]

Lemma: If after insertion, the component is unicyclic (at most 1 cycle), then insertion is possible in time $O(|C|)$
↳ new size of the component

Possible cases:

① 2 trees & x connects them



Directions show potential next posns. for hashed elements

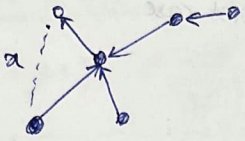
Each vertex has out-degree ≤ 1

Results in a DAG, which must have a sink ⇒ terminates

② x lies within a component C

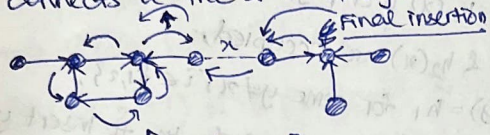
$\Rightarrow$ C must be a tree

~~Once again~~, once again, the tree becomes a DAG, & the sink node ~~gets~~ filled



7/9/23

③ x connects a tree & a ~~unicycle~~ cycle



Eventually it reaches the cycle, reverses it & exits the cycle

Insertion here we have seen

We have a random graph $G(n, m)$ – n vertices, m edges

↓ table size ↓ no. of elements

$m = O(n)$

we want to show,

Fast insertion ① The components are small w.h.p.

Successful insertion ② The small components are unicyclic w.h.p.

However, since no. of edges is fixed, existence of edges is dependent, making analysis difficult

Instead, we analyse

Erdös–Renyi random graph: $G(n, p)$

For every pair of vertices (v_1, v_2), add edge (v_1, v_2) w.p. p indep.

$\Rightarrow \mathbb{E}(\text{No. of edges}) = \binom{n}{2} p = m$, when $p = \frac{m}{\binom{n}{2}}$

Good h.p. bound too, since it is a binomial r.v.

Monotone graph property

A property $\mathcal{P}$ is monotone increasing if

$G \in \mathcal{P} \wedge G \subseteq G' \Rightarrow G' \in \mathcal{P}$

G has the property; G' is a supergraph of G; G' has the property

[Swap to get monotone decreasing]

(E.g.): $\mathcal{P}_1$ – graph has a cycle

$\mathcal{P}_2$ – If a component of size $> k$

} monotone increasing

$\mathcal{P}$ - max. indep. set is of size $\geq k$
- bipartiteness
- planarity

} monotone decreasing

Lemma: Let $\mathcal{P}$ be any monotone increasing property.

Denote by $P(n,M) = \Pr_{G \sim G(n,M)}(G \in \mathcal{P})$

& $P(n,p) = \Pr_{G \sim G(n,p)}(G \in \mathcal{P})$

If $p^+ = \frac{(1+\epsilon)M}{\binom{n}{2}}$, $p^- = \frac{(1-\epsilon)M}{\binom{n}{2}}$, then

$$P(n,p^-) - e^{-\Theta(n)} \leq P(n,M) \leq P(n,p^+) + e^{-\Theta(n)}$$

Thm: Let G be a cuckoo graph w/ $m = (1-\epsilon)\frac{n}{2}$, then

① w.p. $\geq 1 - \frac{1}{n}$, every component has size $\Theta(\log n)$

② ~~w.h.p.~~ the expected size of every component is $O(1)$

Pf: ① $G(n,p^+)$ with $p^+ = \frac{(1+\epsilon)m}{\binom{n}{2}} = \frac{1-\epsilon^2}{n-1}$

To bound the component size, we fix a vertex $v \in V$ & start a BFS

For any vertex v_i, $N_i = |N(v_i)|$ (num. neighbours) $\sim Bin(n-1, p^+)$

$v_1, v_2, \dots, v_{i-1}, v_i \xrightarrow{N_i} Bin(n-i, p^+)$

$\Pr\left(\sum_{i=1}^{k} N_i > k\right)$ ⟶ once again dep.

$B_i \sim Bin(n-1, p^+) \Rightarrow \Pr\left(\sum_{i=1}^{k} B_i > k\right)$ (Branching process)

more flexible, so can bound this to bound our target

$B = \sum_{i=1}^{k} B_i \sim Bin(k(n-i), p^+)$

$\mathbb{E}(B) = k(n-i)p^+ = k(1-\epsilon^2)$

$$\Pr(B > k) = \Pr\left(B \geq \frac{\mathbb{E}(B)}{1-\epsilon^2}\right) \leq \Pr(B \leq (1+\epsilon^2)\mathbb{E}(B))$$

8/9/23

① If $m = (1-\epsilon)\frac{n}{2}$, then the expected size of any connected component in the cuckoo graph is $O(1)$.

② w.h.p. all components of size $\Theta(\log n)$ are unicyclic

(1) Fix $v \in V$. Let S be the size of the component containing v in $G(n, M)$.

$$\mathbb{E}(S) = \sum_{k \geq 0} k \cdot \Pr(S = k)$$

$$= \Pr(S=1) + \Pr(S=2) + \Pr(S=2) + \Pr(S=3) + \Pr(S=3) + \Pr(S=3) + \cdots$$

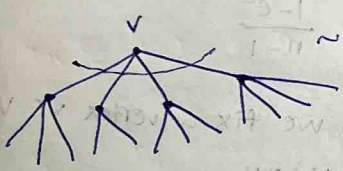
$$= \sum_{k \geq 0} \Pr(S \geq k) \rightarrow \text{monotone increasing}$$

$$\leq \sum_{k \geq 0} \Pr(S' \geq k) + ne^{-\Theta(n)} \rightarrow O(1)$$

S': component size in $G(n, p^+)$

outside summation, $k \leq n$

Branching process



$\sim Bin(n-1, p^+)$

$$p^+ = \frac{2(1+\epsilon)n(1-\epsilon)}{2n(n-1)} = \frac{1-\epsilon^2}{n-1}$$

Y_i = No. of nodes at level i

$$Y = \sum_{i \geq 0} Y_i$$

$$\mathbb{E}(Y_i) = \sum_{k \geq 0} \mathbb{E}(Y_i \mid Y_{i-1} = k) \cdot \Pr(Y_{i-1} = k)$$

$$= \sum_{k \geq 0} k \cdot (n-1)p^+ \cdot \Pr(Y_{i-1} = k) = (n-1)p^+ \mathbb{E}(Y_{i-1})$$

$$= ((n-1)p^+)^i = (1-\epsilon^2)^i$$

$$\mathbb{E}(Y) = \sum_{i \geq 0} (1-\epsilon^2)^i = \frac{1}{\epsilon^2} = O(1)$$

$$\Rightarrow \mathbb{E}(S) = O(1)$$

(2) In $G(n, m)$, how can a component not be unicyclic?

fix k vertices - $\binom{n}{k}$

Cayley's formula: There are k^{k-2} labelled trees on k vertices

Edges are placed into the tree w.p. $\binom{m-1}{k-1} \cdot (k-1)! \cdot \left(\frac{1}{n^2}\right)^{k-1}$ (choosing that edge in $G(n,m)$)

Add 2 edges $\binom{m-k+1}{2} \cdot 2! \underbrace{(k^2)^2 \left(\frac{1}{n^2}\right)^2}_{\text{within the tree}}$

We don't want any more edges to the tree (from other vertices)

$\Rightarrow \left(1 - \frac{k(n-k)}{n^2}\right)^{m-k-1}$ → can add more cycles inside k-vertex-set, which double counts ⇒ lower bound

⇒ Proba. that a component of size k isn't unicyclic:

$\leq \binom{n}{k} k^{k-2} \cdot \binom{m}{k-1} \cdot (k-1)! \cdot \binom{m-k+1}{2} \cdot 2! \cdot k^4 \left(\frac{1}{n^2}\right)^{k+1} \left(1 - \frac{k(n-k)}{n^2}\right)^{m-k}$

$\leq \frac{1}{n}$ when $k = \Theta(\log n)$, after heavy calculation

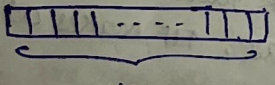
---

Midsem portions complete

Online algorithms

- Data arrives as a stream & not all at once
- At each step, we make irrevocable decisions
- We want to know how good the decisions (made with incomplete information) are

Paging problem

Cache 

k

Seq. of page reqs. $r_1, r_2, \ldots, r_m$

If r_i does not exist in the cache, ⇒ cache miss / page fault
evict some item & bring r_i in

Goal: Minimize page faults

How do we know if some strategy is good?
Optimal offline strategy
Farthest-in-the-future
⇒ compare against this

Deterministic online strategies:

1. LRU - Least recently used
2. FIFO - First-in first-out
3. LFU - Least frequently used

An adversary can design the requests to forcefully trigger faults at every step
⇒ what does it mean to minimize page faults?

Competitive ratio

An algo. A is α-competitive if $A(\sigma) \leq \alpha \cdot OPT(\sigma) + c$ (indep. of m)