Michael Forbes
miforbes@illinois.edu
2019-09-19.4 → 2019-09-24.1
2019-09-24.2
cs473

cs 473 Algorithms: Lecture 9

logistics: - pset3 due w10
- midterm 1 Oct 7

last time: - quicksort
- concentration bounds
- limited independence

today: - hashing

goal: - develop randomized algo w/ $\max_{|x|=n} \mathbb{E}[T(x)] =: \bar{T}(n)$ small [expected runtime]
- " w/ $\Pr[T(x) >> \bar{T}(n)] \le$ small [concentration bound] [stronger result often needed]

thm (Chernoff): $X_1, \ldots, X_n \in [0,1]$ indep. $X = \sum X_i$, $\varepsilon \ge 0$
$$\Pr[X \ge 2 \cdot \mathbb{E}[X]] \le e^{-\mathbb{E}[X]/3}$$ [among other forms]

rmk: - sums of independent random variables are concentrated [(weak) law of large numbers]
- fully independent random variables can be expensive
  - randomness
  - space [will see]
  - algorithmically

Q: concentration bounds without full independence? [somewhat independence?]

def. $X_1, \ldots, X_n \in \Sigma$ are independent if $\forall \sigma_1, \ldots, \sigma_n$
$$\Pr[X_1 = \sigma_1 \wedge \cdots \wedge X_n = \sigma_n] = \Pr[X_1 = \sigma_1] \cdots \Pr[X_n = \sigma_n]$$
are uniform if $\forall i, \sigma$ $\Pr[X_i = \sigma] = \frac{1}{|\Sigma|}$

$X_1, \ldots, X_n$ are pairwise independent (and uniform) if $\forall i \ne j, \sigma, \sigma'$ $\Pr[X_i = \sigma \wedge X_j = \sigma'] = \Pr[X_i = \sigma] \cdot \Pr[X_j = \sigma']$

ex: $X, Y \in \{0,1\}$ independent uniform
$\Rightarrow X, Y, X \oplus Y$ [mod 2] are pairwise independent and uniform [not independent]

eg: $\Pr[X = 1, X \oplus Y = 0] = \Pr[X \oplus Y = 0 \mid X = 1] \cdot \underbrace{\Pr[X = 1]}_{=1/2} = 1/4$
$= \Pr[1 \oplus Y = 0]$
$= \Pr[Y = 1] = 1/2$

def. $h: U \times \{0,1\}^{\ell}$ [seed] $\to T$ is a pairwise independent hash function
if $\forall u_1 \ne u_2 \in U$, $t_1, t_2 \in T$ $\Pr_{s \in \{0,1\}^\ell}[h(u_1, s) = t_1 \wedge h(u_2, s) = t_2] = \frac{1}{|T|^2}$
ℓ = seed length

rmk: ↳ $\equiv$ the random vars $(h(u,s))_{u \in U}$ are pairwise independent and uniform
- fully random function takes $\ell = |U| \cdot \lg |T|$ [seed stores entire truth table] [do better?]

lem: $\mathbb{Z}_2$ = arithmetic modulo 2
$$h: \mathbb{Z}_2^n \times (\mathbb{Z}_2^{k \times n} \times \mathbb{Z}_2^k) \to \mathbb{Z}_2^k$$
$$h: (x, (A, b)) \mapsto A \cdot x + b$$
$k \left[\overbrace{A}^{n} \right] \left[x \right] n + \left[b \right] k = \left[\ \right] k$

is pairwise independent hash function, seed length $\ell = k \cdot (n+1)$

pf: $x \ne y \in \mathbb{Z}_2^n$, $c, d \in \mathbb{Z}_2^k$: $\Pr_{A,b}[\underbrace{Ax + b = c \wedge Ay + b = d}]$
$\Leftarrow$ $Ax + b = c$
$A(x-y) = c - d$

Michael Forbes
miforbes@illinois.edu
2019-09-24.2 ← 2019-09-24.1
→ 2019-09-24.3
cs473

$= \Pr_{A,b}[\underbrace{Ax+b=c}_{\Leftrightarrow\, b = c - Ax,\ b \in \mathbb{Z}_2^k} \mid \underbrace{A(x-y)=c-d}_{\text{indep of } b}] \cdot \Pr_A[A\underbrace{(x-y)}_{\neq 0} = c-d]$

$= \frac{1}{2^k}$

exercise: $k\left[\text{random}\right]^{n}$ (nonzero) $= [\ ]^k$ random

$= \frac{1}{2^k}$

$= (1/2^k)^2$ ∎

Rmk: - seed length $\ell = k\cdot(n+1) = O(k\cdot n) \ll k \cdot 2^n$

- can achieve $O(k+n)$

(Chebyshev)

prop: $X_1, \ldots, X_n \in [0,1]$ pairwise indep. $X = \frac{\sum X_i}{n}$ $\varepsilon > 0$

$\Pr[\,|X - \mathbb{E}[X]| \geq \varepsilon\,] \leq \frac{1}{\varepsilon^2 n}$

Pf Sketch: exercise: $X_1, \ldots, X_n$ pairwise indep $\Rightarrow \mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$

prop: $h: \mathbb{Z}_p \times (\mathbb{Z}_p \times \mathbb{Z}_p) \longrightarrow \mathbb{Z}_p$ p prime

$h: x \times (a,b) \longmapsto ax+b$ is pairwise independent hash family

rmk: - seed length $\ell \approx 2\lceil \lg p \rceil \ll (\lg p)^2$

- requires finding prime numbers

def: a field $\mathbb{F}$ is a set s.t.
- $(\mathbb{F}, +)$ commutative group [identity] [inverses] [associativity]
- $(\mathbb{F}\setminus\{0\}, \times)$ commutative group
- distributivity $\alpha(\beta+\gamma) = \alpha\cdot\beta + \alpha\gamma$

eg: - $\mathbb{R}, \mathbb{C}, \mathbb{Q}$

- finite fields have $|\mathbb{F}| < \infty$.

Prop: $\mathbb{Z}_p = \{0, 1, \ldots, p-1\} =: \mathbb{F}_p$ is a field w/ arithmetic modulo p

Pf: $(\mathbb{F}_p, +)$: exercise

distributivity: exercise

$(\mathbb{F}_p - \{0\}, \times)$: identity: 1

associativity: exercise

inverses?

lem: p prime. $0 \neq x \in \mathbb{F}_p$. $\exists!\ y$ st $xy = 1 \bmod p$

Pf: define $\mu_x: \mathbb{F}_p \to \mathbb{F}_p$
$y \mapsto xy$

Clm: μ_x injective

pf: $y, z \in \mathbb{F}_p$ $\mu_x(y) = \mu_x(z) \equiv xy = xz\ (p)$

$x(y-z) = 0\ (p)$

$p \mid x(y-z)$ but $p \nmid x$

$p \mid (y-z)$

$y = z\ (p) \Rightarrow y = z$ in $\mathbb{F}_p$

Clm: μ_x bijection

Pf: $\mathbb{F}_p \xrightarrow{\mu_x \text{ injection}} \mathbb{F}_p$ $\underset{\text{cardinality}}{\Longrightarrow}$ bijection

Michael Forbes
mforbes@illinois.edu
2019-09-24.2 →
2019-09-24.4 ←
2019-09-24.3
CS473

$\Rightarrow (\mu_x)^{-1}: \mathbb{F}^p \to \mathbb{F}^p$ is bijection

$\Rightarrow (\mu_x)^{-1}(1) = y$ is unique y st $\underbrace{\mu_x(y)}_{=xy} = 1$

rmk: - any prime p, $k \in \mathbb{Z}$, exists unique $\mathbb{F}$ w/ $|\mathbb{F}| = p^k$ [did $k=1$]

Prop - $h: \mathbb{F}_p \times (\mathbb{F}_p \times \mathbb{F}_p) \to \mathbb{F}_p$ is pairwise indep

$x \times (a,b) \longmapsto ax+b$

Pf: $x \neq y \in \mathbb{F}_p$, $c,d \in \mathbb{F}_p$ $\quad \Pr_{a,b}[\underbrace{ax+b=c \wedge ay+c=d}_{\substack{ax+b=c \\ a(x-y)=c-d}}]$

$$= \underbrace{\Pr_{a,b}[\underbrace{ax+b=c}_{b=c-ax} \mid \underbrace{a(x-y)=c-d}_{\text{indep of } b}]}_{=1/p} \cdot \underbrace{\Pr_a[a\underbrace{(x-y)}_{\neq 0}=c-d]}_{=\Pr[a=\frac{c-d}{x-y}]=1/p}$$

$$=1/p^2$$

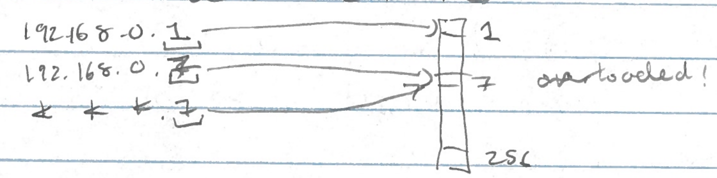
$\boxed{}$

Rmk - can define t-wise independence

- $h: \mathbb{F}_p \times (\mathbb{F}_p)^{t+1} \to \mathbb{F}_p$ defined by $x \mapsto a_0 + a_1x + a_2x^2 + \cdots + a_t x^t$ is t-wise indep

- concentration bounds for t-wise indep get stronger as $t \to \infty$

Q: how to load balance on the internet?



goal: minimize the maximum load

def: A dictionary is a data structure that maintains $S \subseteq \mathcal{U}$. [keys that we store]

w/ lookup(x): is $x \in S$?

insert(x): add x to S [no op if $x \in S$]

delete(x): remove x from S [no op if $x \notin S$]

is static: all insertions before lookup [no delete]

dynamic: intermingle insert/lookup/delete

ex: - sorted array - static - $O(\lg |S|)$ lookup

- self-balancing binary search tree - dynamic - $O(\lg |S|)$ operations

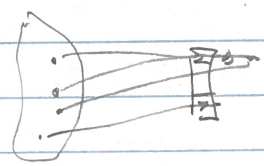
Q: do "better"?

A: relax: $|S|$ space $\mapsto O(|S|)$ space

deterministic $\longmapsto$ monte carlo randomization [some probability of error]

Michael Forbes
miforbes@illinois.edu
2019-09-24.4 $\leftrightarrows$ 2019-09-24.3
cs473

def: A **hash table w/ chaining** is a dictionary defined by hash function $h: U \to T$, $|T| = m$. $x \in S$ is stored at $h(x)$ ← linked list

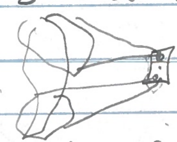


rmk: - other collision strategies exist
- dictionary operations take $\le \underbrace{\max_{x \in S} |T[h(x)]|}$ goal: min

ex: $h: U \to U$ identity
- max load $= 1$
- no space savings if $|S| \ll |U|$

lem: $h: U \to T$ fixed. $\exists S \subseteq U$ w/ $h(S) = \{t\}$, $|S| \ge \frac{|U|}{|T|}$

pf: pigeonhole



$\Rightarrow$ in worst case cannot use single fixed function [in practice can hope this does not happen OR use crypto to make worst case behavior hard to find]

Lem. $S \subseteq U$ [static] $h: U \to T$ random
any $x \in S$: $\mathbb{E}[|T[h(x)]|] \le 1 + \frac{|S|-1}{|T|}$

pf: $= \underbrace{\Pr[x \in T[h(x)]]}_{=1} + \sum_{y \ne x} \underbrace{\Pr[y \in T[h(x)]]}_{=\frac{1}{|T|}} = 1 + \frac{|S|-1}{|T|}$ [linearity, indicator vars]

Rmk: - $n = m \Rightarrow O(1)$ [load, op time]
- storing h takes $|U| \cdot \lg|T|$ bits $\gg |S|$

lem: $\Rightarrow$ if $h: U \times \{0,1\}^\ell \to T$ pairwise independent

pf. $\Pr_s[y \in T[h(x,s)]] = \Pr_s[h(y,s) = h(x,s)] = \frac{1}{|T|}$

Rmk: - can use $h(x) = Ax + b$ [matrices] or $h(x) = ax + b$ [$\mathbb{F}_p$]
- fast evaluation
- cheap to store h
- only need $\Pr[h(x,s) = h(y,s)] \le \frac{1}{|T|}$ "universal" hash function
- can extend to dynamic hashing [w/ limitations]

fact: $|S| = |T| = n$:
h pairwise indep $\Rightarrow \mathbb{E} \max_{t \in T} |T[t]| \in O(\sqrt{n})$ [tight]
h random: $O\left(\frac{\lg n}{\lg \lg n}\right)$
h $O\left(\frac{\lg n}{\lg \lg n}\right)$ indep: $O\left(\frac{\lg n}{\lg \lg n}\right)$
still cheaper than

logistics: - pset3 due W10
- midterm 1 Oct 7

today: - limited independence
- hashing

next time: - randomized algo