15 → 17

written → (2022-03-22)

cs473 Algorithms: Lecture 16

pset 6 due F17

logistics:

last lecture: randomized algo — closest pair — $O(n\lg n)$ deterministic [lecture 2 divide and conquer]
— verification dist ≥ Δ [insert points, create local subspaces, use dictionary]
— randomly order points [low prob of many updates]

today: randomized algo

Q: what is complexity of randomized algo?
[we've been doing this]
[been using (expected) runtime]

runtime is a random variable ← has distribution [how to quantify?]

[see example]
recall quicksort

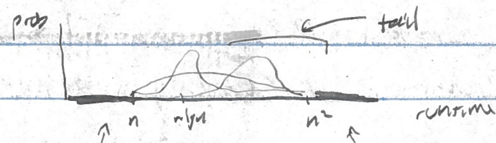
given $a = a_1, \ldots a_n$ integers, want to sort

$T(a)$ = run time of quicksort on a ← random variable [input fixed] [algo randomized] — is comparisons

[expected runtime] [sort]

$T(n) = \max_{|a|=n} \mathbb{E}[T(a)] \le O(n\lg n)$

distribution:



[distribution?]

always ≥ n — expectation — always ≤ $O(n^2)$

Q: $\Pr[T(a) \ge n^2/10]$ [really bad runtime for sorting] — $\ge 1/2$? [algo hard of time expensive] no [will see]
$\le 1/\sqrt{n}$? [small prob, acceptable] yes [will see]
$\le 1/2^{\sqrt{n}}$? [so small, essentially zero] yes [will see]

[need more refined notion of complexity]

def: a randomized algo runs in time $T(n)$ with probability $1-\delta(n)$ if
for all inputs of size n, the algorithm runs in time $T(n)$, and outputs the $\boxed{\text{correct}}$ answer
with probability $\ge 1-\delta(n)$

Algorithm runs in time $T(n)$ with high probability if $\delta(n) = 1/n$ [error → 0 quickly]
[two types of randomness] [correctness now random]

rmk: quicksort will always $\boxed{\text{(correctly)}}$ output the correct answer, the $\boxed{\text{runtime}}$ is random
next lecture will see algorithm whose runtime is $\boxed{\text{not}}$ random, correctness only with $\boxed{\text{high probability}}$

Q: $\Pr[X \le \mathbb{E}[X]] \ge 1$? [we've been using $\mathbb{E}[X]$ as measure of runtime]
[want to argue this is meaningful, as $X \le \mathbb{E}[X]$ in our situations]

lem [Markov's inequality]: $X \ge 0$ and rv [non-negative is key!]
for $a \ge 0$, $\Pr[X \ge a] \le \mathbb{E}[X]/a$

pf: $\mathbb{E}[X] = \underbrace{\mathbb{E}[X \mid X \ge a]}_{\ge a} \cdot \Pr[X \ge a] + \underbrace{\mathbb{E}[X \mid X < a]}_{\ge 0 \,\leftarrow\, X \ge 0} \cdot \underbrace{\Pr[X < a]}_{\ge 0}$ — $a \ne 0$.
[conditional] [what we want] [!] [prob ∈ [0,1]]

$\ge a \cdot \Pr[X \ge a]$ ∎

cor: $X \ge 0$, $k \ge 0$. $\Pr[X \ge k \cdot \mathbb{E}[X]] \le 1/k$ [cannot deviate too far from expectation]
[runtime is non-negative!] [quicksort]

cor: $\Pr[\underbrace{T(a) \ge k \cdot \Theta(n\lg n)}_{\ge n^2/10}] \le 1/k$
$\le \Theta\left(\frac{\lg n}{n}\right)$ [goes to zero quickly]

Q: do better?

algo: quicksort-restart: - while

- run b = quicksort(a) for 2·$\Theta(n\lg n)$ steps [keep clock] [abort if runs too long] — expected runtime [unspecified constant]

- if b sorted, output b [else restart]

• clock

no need to check if b sorted only if terminates

prop: quicksort-restart is always correct [clear]

prop: quicksort-restart runs in $O(k\, n\lg n)$ time with probability $1-\frac{1}{2^k}$ [vs $1-\frac{1}{k}$]

pf: clm: checking if b is sorted is $O(n)$ time [single pass, look for violations]

let L be number of loops

Clm: runtime is $L\cdot(2\cdot\Theta(k\, n\lg n)+\Theta(n))$ [clear] [how many loops?]

$X_i = \begin{cases} 1 & \text{ith attempt at quicksort succeeds in } 2\cdot\Theta(n\lg n) \text{ time} \\ 0 & \text{else} \end{cases}$

clm: $\Pr[X_i=0] = \Pr[\text{quicksort runtime} \geq 2\cdot \mathbb{E}[\text{quicksort runtime}]] \leq \frac{1}{2}$ ← Markov

stochastic dominance

cor: $\Pr[L>k] = \Pr[X_1=\dots=X_k=0] \leq \frac{1}{2^k}$

$\Rightarrow \Pr[\text{runtime} > k\cdot\Theta(n\lg n)] \leq \frac{1}{2^k}$ □

cor: quicksort-restart runs in $O(n\lg^2 n)$ time w/p $1-\frac{1}{n}$ [why]

[can generalize to] [generic way to turn expected runtimes into whp also] [slight penalty in runtime]

$O(n\lg^2 n)$ time w/p $1-\frac{1}{n^c}$, any $c=\Theta(1)$ [tradeoff runtime and success prob]

$\geq n^2/10$ time w/p $\frac{1}{2}\Theta(n/\lg n)$

requires always correct

Q: avoid penalty turning expected runtime into whp runtime? [better probability tools] [better analysis of algo]

•

fact: quicksort [original] runs in $O(n\lg n)$ time with probability $1-\frac{1}{n^c}$, any $c=\Theta(1)$

$\geq n^2/10$ time w.p. $\frac{1}{2}\Theta(n)$ ...

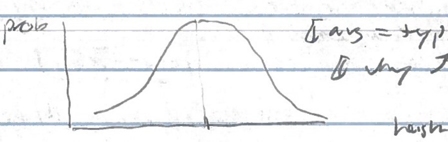
[better probabilistic tools]

Q: better tail bounds?

Q: what is expectation? [why should we expect expectation]

- dete

eg distribution of human height,



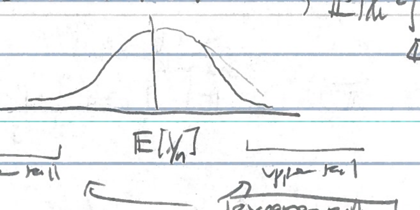
[avg = typical] [why ?]

[intuition: height is sum of many small independent environmental/genetic factors ⇒ deviations "cancel out"]

thm [central limit theorem]:

$X_1,\dots X_i,\dots$ random var over $\mathbb{R}$, identical and independent, $\mathbb{E}[X_i^2]<\infty$

then $Y_n = \frac{X_1+\dots+X_n}{\sqrt{n}} \xrightarrow[n\to\infty]{} \text{gaussian distribution}$ [not too pathological]



Q: quantitative version? [needed for algo analysis]

[exponentially] small [much better than Markov]

[[family of bounds]]

thm [Chernoff bound]:

$X_1, \dots, X_n \in \{0,1\}$ [[binary]] independent, $\mathbb{E}[X_i] = p_i$ [[may not be identical]] $\mu \geq \mathbb{E}[X]$

$X = X_1 + \dots + X_n$ $\quad \delta \geq 0$, $\Pr[X \geq (1+\delta)\mathbb{E}[X]] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}$ [[messy...]]

$\varepsilon \geq 0$, $\Pr[|X - \mathbb{E}[X]| \geq \varepsilon \cdot n] \leq 2 \cdot e^{-\varepsilon^2 n/4}$ [[simpler, less general]] $\quad \varepsilon \sim \delta$

[[do applications, then prove, rest on pset]]

eg (polling)

$S \subseteq \Omega$ estimate $\frac{|S|}{|\Omega|}$ [[fraction of voters for candidate]]

algo: pick $\sigma_1, \dots, \sigma_t \in \Omega$ at random [[t parameter]]

return $\sum_i \frac{\mathbb{1}[\sigma_i \in S]}{t}$

complexity: $O(t)$ [[what is t?]]

correctness: $X_i = \mathbb{1}[\sigma_i \in S]$ $\quad \mathbb{E}[X_i] = \frac{|S|}{|\Omega|} = \mu$

$X = \sum X_i$ $\quad \mathbb{E}[X] = \mu \cdot t$

$Y = \frac{1}{t} X$ [[normalize]] [[also output]] $\quad \mathbb{E}[Y] = \mu$

$\Pr[\underbrace{|Y - \mu|}_{= \frac{1}{t}|X - \mathbb{E}[X]|} \geq \varepsilon] \leq 2e^{-\varepsilon^2 t/4} \leq .05$ if $\varepsilon = .05$ [[error]] [[chernoff]]

if $t \geq \underbrace{8}_{\text{can remove}} \cdot 738$ [[polling eg from before]]

eg (balls into bins)

m balls, n bins, each ball thrown into random bin. take $\boxed{m=n}$

Q: what is the max load of bin j? [[interesting case]]

$X_{ij} = \mathbb{1}[\text{ball } i \to \text{bin } j]$ $\quad \mathbb{E}[X_{ij}] = 1/n$

$Y_j = \sum_i X_{ij}$ = # balls in bin j $\quad \mathbb{E}[Y_j] \overset{\text{[[linearity]]}}{=} n \cdot 1/n = 1$ [[same as in hashing w/ rand. func.]]

[[tail bounds]]

$\Pr[Y_j \geq \underbrace{c}_{= c \cdot \mathbb{E}[Y_j]}] \leq \left(\frac{e^{c-1}}{c^c}\right)^1 = e^{-\Theta(c \ln c)}$ [[small!]]

[[worst case notion]]

Q: what is the max load over $\boxed{\text{all}}$ bins?

$\Pr[\max_j Y_j \geq c] = \Pr[\exists j \; Y_j \geq c]$

$\overset{\text{[[union]]}}{\leq} \sum_j \Pr[Y_j \geq c]$

$= n \cdot \underbrace{e^{-\Theta(c \ln c)}}$

choose $c = k \cdot \frac{\ln n}{\ln \ln n}$

$= n \cdot \frac{1}{n^{\Theta(k)}} = \frac{1}{n^{\Theta(k)}}$ [[whp max $\leq O(\frac{\ln n}{\ln \ln n})$]]

$\mathbb{E}[\underbrace{\max_j Y_j}_{=: Y}] = \underbrace{\mathbb{E}[Y \mid Y < k \frac{\ln n}{\ln \ln n}]}_{\leq k \cdot \frac{\ln n}{\ln \ln n}} \cdot \underbrace{\Pr[\text{''}]}_{\leq 1} + \underbrace{\mathbb{E}[Y \mid Y \geq k \frac{\ln n}{\ln \ln n}]}_{\leq n \text{ [[n balls]]}} \cdot \underbrace{\Pr[-]}_{\leq \frac{1}{n^{\Theta(k)}}}$

$= O(\frac{\ln n}{\ln \ln n})$ for $k = \Theta(1)$

[[each bin has $\mathbb{E} = 1$ individually, but max load is expected to be higher]] thm [Chernoff]. $X_1, \ldots, X_n \in \{0,1\}$ random variables, independent, $\mathbb{E}[X_i] = p_i \in [0,1]$

$X = X_1 + \cdots + X_n$, $\delta \geq 0$. $\Pr[X \geq (1+\delta)\mathbb{E}[X]] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}$ [it was weak before]

pf: idea = Markov's inequality to [transformed] variable

$t > 0$ parameter. $\Rightarrow$ $x \mapsto e^{tx}$ is strictly increasing function

$$\Pr[X \geq (1+\delta)\underbrace{\mu}_{\mathbb{E}[X]}] = \Pr[e^{tX} \geq e^{t(1+\delta)\mu}] \overset{[\text{Markov}]}{\leq} \mathbb{E}[e^{tX}] / e^{t(1+\delta)\mu}$$

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(X_1 + \cdots + X_n)}] = \mathbb{E}[e^{tX_1 + \cdots + tX_n}]$$

$$= \mathbb{E}[e^{tX_1} \cdots e^{tX_n}] \quad [\text{exp rules}]$$

$$\overset{\text{indep}}{=} \mathbb{E}[e^{tX_1}] \cdots \mathbb{E}[e^{tX_n}]$$

$$\mathbb{E}[e^{tX_i}] \overset{[X_i \text{ binary}]}{=} e^t \cdot \underbrace{\Pr[X_i = 1]}_{p_i} + \underbrace{e^0}_{1} \cdot \underbrace{\Pr[X_i = 0]}_{1-p_i} = 1 + p_i(e^t - 1) \leq e^{p_i(e^t - 1)}$$

$1 + z \leq e^z$ [lecture 12]

$$\leq \prod_i e^{p_i(e^t - 1)} = e^{\sum p_i (e^t - 1)} = e^{\mu(e^t - 1)}$$

$$\leq e^{\mu(e^t - 1)} / e^{t(1+\delta)\mu}$$

choose $t = \ln(1+\delta)$

$$= e^{\mu((1+\delta) - 1)} / (1+\delta)^{(1+\delta)\mu} = \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \quad \square$$

---

today: randomized algo — complexity of randomized algo
- expectation [previously]
- distribution [concentration]
- tails [more details]

— tail bounds
- Markov
- quicksort - restart [restart if taking too long]

— tail bounds
- CLT
- Chernoff
- polling
- balls-n-bins

next lecture: randomized algo

logistics: pset 6 due F 17