$\{$ Csc 80030 | Lecture 4 $\}$

Probabilistic Analysis & Randomized Algorithms

Hoeteck Wee · hoeteck@cs.qc.edu

PART 0

- HOMEWORK: HW2 is out; review HW1
- TODAY: randomized median finding & balls n' bins

PART 1 | randomized median finding

Median Finding

MEDIAN FINDING Problem

Input: a set S of n values from some totally ordered universe

Goal: output the median element m of S

- WHAT'S KNOWN: "easier" than sorting – there is a deterministic linear-time algorithm.
- TODAY: a randomized linear-time algorithm based on random sampling and the fact that we can sort a set of size $O(n^{3/4})$ in $O(n)$ time.

Randomized Median Finding

RandMedian Algorithm

Input: A list S of n distinct values

1. Sample ℓ from S such that $\text{rank}_S(\ell)/n \in [\frac{1}{2} - 2\delta, \frac{1}{2}]$;
2. Sample u from S such that $\text{rank}_S(u)/n \in [\frac{1}{2}, \frac{1}{2} + 2\delta]$;
3. By comparing with each value in S, compute $C = \{y \in S \mid \ell \leq y \leq u\}$;
4. Output the $(\frac{1}{2}n - \text{rank}_S(\ell) + 1)$'th smallest element in the sorted set C.

- CORRECTNESS: $\text{rank}_S(\text{output}) = \text{rank}_C(\text{output}) + (\text{rank}_S(\ell) - 1)$.
- FACT: $|C| \leq 4\delta n + 1$, can sort in $O(\delta n \log n)$ time; set $\delta = 1/n^{1/4}$.
- GOAL: implement steps 1, 2 in $O(n)$ time.

Randomized Median Finding

RandMedian Subroutine

1. Sample ℓ from S such that $\mathrm{rank}_S(\ell)/n \in [\frac{1}{2} - 2\delta, \frac{1}{2}]$;
   1.1 Pick a set R of $n^{3/4}$ elements in S with replacement.
   1.2 Output the element x in R whose rank is $(\frac{1}{2} - \delta)|R|$.

- INTUITION: $\mathrm{rank}_S(x)/n \approx \mathrm{rank}_R(x)/|R| = \frac{1}{2} - \delta$
- CLAIM: $\Pr_R[\mathrm{rank}_S(x)/n \notin [\frac{1}{2} - 2\delta, \frac{1}{2}]$ is tiny.
- SUB-CLAIM 1: $\Pr_R[\mathrm{rank}_S(x)/n > \frac{1}{2}]$ is tiny.
  - $\mathrm{rank}_S(x)/n > \frac{1}{2} \Rightarrow \#\{\text{elements in } R > \mathrm{median}(S)\} \geq (\frac{1}{2} + \delta)|R|$
  - X: $\#\{\text{elements in } R > \mathrm{median}(S)\}$, so $X \sim B(|R|, \frac{1}{2})$
  - Show $\Pr\left[X \geq (\frac{1}{2} + \delta)|R|\right]$ is tiny.

Analysis of RANDMEDIAN

- SUB-CLAIM 1: $\Pr_R[\mathrm{rank}_S(x)/n > \frac{1}{2}]$ is tiny.
  - $X \sim B(|R|, \frac{1}{2})$, will show $\Pr\left[X \geq (\frac{1}{2} + \delta)|R|\right]$ is tiny.

- PROOF:
  - $\mathrm{E}[X] = \frac{1}{2}|R|$ and $\mathrm{Var}[X] = \frac{1}{4}|R|$.
  - Applying Chebyshev's,

$$\begin{aligned}\Pr\left[X \geq (\tfrac{1}{2} + \delta)|R|\right] &\leq \Pr\left[|X - \mathrm{E}[X]| \geq \delta|R|\right] \\ &\leq \frac{\mathrm{Var}[X]}{(\delta|R|)^2} = \frac{1}{4n^{1/4}}\end{aligned}$$

Analysis of RandMedian

- SUB-CLAIM 2: $\Pr_R[\mathrm{rank}_S(x)/n < \frac{1}{2} - 2\delta]$ is tiny.
  - $\mathrm{rank}_S(x)/n < \frac{1}{2} - 2\delta \Rightarrow$
    $\#\{\text{elements in } R \text{ with } \mathrm{rank}_S < (\frac{1}{2} - 2\delta)n\} \geq (\frac{1}{2} - \delta)|R|$
  - Y: $\#\{\text{elements in } R \text{ with } \mathrm{rank}_S < (\frac{1}{2} - 2\delta)n\}$, so $Y \sim B(|R|, \frac{1}{2} - 2\delta)$
  - Show $\Pr\left[Y \geq (\frac{1}{2} - \delta)|R|\right]$ is tiny.

- PROOF:
  - $\mathrm{E}[Y] = (\frac{1}{2} - 2\delta)|R|$ and $\mathrm{Var}[Y] = (\frac{1}{4} - 4\delta^2)|R| \leq \frac{1}{4}|R|$.
  - Applying Chebyshev's,

$$\begin{aligned}\Pr\left[Y \geq (\tfrac{1}{2} - \delta)|R|\right] &\leq \Pr\left[|Y - \mathrm{E}[Y]| \geq \delta|R|\right] \\ &\leq \frac{\mathrm{Var}[Y]}{(\delta|R|)^2} \leq \frac{1}{4n^{1/4}}\end{aligned}$$

Analysis of RANDMEDIAN

RANDMEDIAN Algorithm

Input: A list S of n distinct values

1. Sample ℓ from S such that $\text{rank}_S(\ell)/n \in [\frac{1}{2} - 2\delta, \frac{1}{2}]$;
2. Sample u from S such that $\text{rank}_S(u)/n \in [\frac{1}{2}, \frac{1}{2} + 2\delta]$;
3. ...

- $\Pr\left[\text{rank}_S(\ell)/n \notin [\frac{1}{2} - 2\delta, \frac{1}{2}]\right] \leq \frac{1}{2n^{1/4}}$
- similarly, $\Pr\left[\text{rank}_S(u)/n \notin [\frac{1}{2}, \frac{1}{2} + 2\delta]\right] \leq \frac{1}{2n^{1/4}}$
- $\Pr[\text{RANDMEDIAN outputs incorrect answer}] \leq \frac{1}{n^{1/4}}$

PART 2 | balls ’n bins

Birthday "Paradox"

QUESTION. What is the probability that amongst 30 people in a room, two share the same birthday?

MODEL. Everyone's birthday is independently and uniformly chosen at random amongst 365 days.

ANALYSIS. $\Pr[\text{all birthdays are distinct}]$ is

$(1-\frac{1}{365})\cdot(1-\frac{2}{365})\cdot(1-\frac{3}{365})\cdots(1-\frac{29}{365}) \approx 0.2937$

MORE GENERALLY... For m people and n "birthdays", it's

$$\begin{aligned} & (1-\tfrac{1}{n})\cdot(1-\tfrac{2}{n})\cdot(1-\tfrac{3}{n})\cdots(1-\tfrac{m-1}{n}) \\ \approx \; & \prod_{j=1}^{m-1} e^{-j/n} = e^{-m(m-1)/2n} \approx e^{-m^2/2n} \end{aligned}$$

$\Rightarrow$ constant prob of "collision" whenever $m \gtrsim \sqrt{2n \ln 2}$

Balls-and-Bins Model

- m balls thrown into n bins
  - location of each ball independent and random
- Example: job scheduling
  - balls = tasks, bins = processors
- Quantities of interest
  - average load = expected number of balls in each bin
  - maximum load = number of balls in fullest bin
  - number of empty bins (= number of idle processors)

Average/Maximum Load

- L_i be r.v. for # balls in Bin i
  - $L_i \sim B(m, \frac{1}{n})$, so $\mathrm{E}[L_i] = \frac{m}{n}$, $\mathrm{Var}[L_i] = \frac{m}{n}(1 - \frac{1}{n})$

Chernoff Bound

Let $X_1, \ldots, X_n$ be *independent* $\{0, 1\}$-r.v.'s. Let $X = X_1 + \cdots + X_n$ and $\mu = \mathrm{E}[X]$.
Then, for all $\delta > 0$, $\Pr[X \geq (1 + \delta)\mu] \leq e^{-\frac{\mu\delta^2}{2+\delta}}$

- APPLICATION. bounding $\Pr[L_i \geq 2\ln n + 1]$ for $m = n$
  - set $\mu = 1, \delta = 2\ln n$, so $\frac{\mu\delta^2}{2+\delta} \geq 2\ln n$
    $\Rightarrow \Pr[L_i \geq 2\ln n + 1] \leq e^{-2\ln n} = \frac{1}{n^2}$
  - By union bound, $\Pr[\bigvee_{i=1}^{n}(L_i \geq 2\ln n + 1)] \leq \frac{1}{n}$
  - Hence, $\Pr[\text{maximum load} \leq 2\ln n + 1] \geq 1 - \frac{1}{n}$.
  - e.g. $n = 1$ million, max load is at most 30 w.h.p.

Maximum Load for $m = n$

- BETTER ANALYSIS.

$$\begin{aligned}
\Pr[L_i \geq k] &= \Pr[\exists \text{ subset of } k \text{ balls all of which fall into bin } i] \\
&\leq \binom{n}{k} \cdot (1/n)^k \\
&\leq (ne/k)^k \cdot (1/n)^k = (e/k)^k \\
&\leq 1/n^2 \qquad \text{for } k \geq \tfrac{3 \ln n}{\ln \ln n}
\end{aligned}$$

- BETTER BOUND.
  - obtain a bound of $O(\frac{\log n}{\log \log n})$ instead of $O(\log n)$ for the maximum load.
  - e.g. $n = 1$ million, max load is at most 16 w.h.p.

Empty Bins

- Let X be random variable for # empty bins.
- Let X_i be r.v. indicating whether Bin i is empty.
- $\Pr[X_i = 1] = (1 - \frac{1}{n})^m$ and $\mathrm{E}[X] = n(1 - \frac{1}{n})^m$.
- NOTE. X_i and X_j are *not* independent, e.g.
  $\Pr[X_i = 1 \wedge X_j = 1] = (1 - \frac{2}{n})^m \neq \Pr[X_i = 1] \cdot \Pr[X_j = 1]$

Empty Bins: Variance

- Recall $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$
  - $\mathrm{E}[X^2] = \mathrm{E}[(X_1 + \cdots + X_n)^2] = \sum_{i=1}^{n} \mathrm{E}[X_i^2] + \sum_{i \neq j} \mathrm{E}[X_i X_j]$
  - If $X_i \in \{0, 1\}$, then $\mathrm{E}[X_i^2] = \mathrm{E}[X_i]$
- Computing $\mathrm{E}[X_i X_j]$
  - $\mathrm{E}[X_i X_j] = \Pr[X_i X_j = 1] = \Pr[X_i = 1 \wedge X_j = 1] = (1 - \frac{2}{n})^m$
- Computing $\mathrm{Var}[X]$
  - $\mathrm{E}[X^2] = n(1 - \frac{1}{n})^m + n(n-1)(1 - \frac{2}{n})^m$
  - $\mathrm{Var}[X] = n(1 - \frac{1}{n})^m + n(n-1)(1 - \frac{2}{n})^m - n^2(1 - \frac{1}{n})^{2m}$

THE END | HW2 due next week