LECTURE 4

Data stream $A = (a_1, a_2, \ldots, a_m)$

$a_i \in N = \{1, 2, \ldots, n\}$

Let $M_i = \{ j \mid a_j = i \}$

$m_i = |M_i|$ = # copies of value i in stream

$\vec{m} = (m_1, m_2, \ldots, m_n)$ FREQUENCY VECTOR

The kth FREQUENCY MOMENT F_k of A is

$$F_k = \begin{cases} |\{ i \mid m_i \neq 0 \}| & \text{if } k = 0 \\ \max_i m_i & \text{if } k = \infty \\ \|m\|_k^k = \sum_{i=1}^{n} m_i^k & \text{otherwise} \end{cases}$$

In particular F_0 = # distinct elements

F_1 = length of sequence ($= m$)

F_2 = REPEAT RATE or GINI's INDEX OF HOMOGENEITY

Motivation for Alon-Matias-Szegedy STOC 99

FM important for database applications

Claim: Essentially all query optimization methods in relational DB systems requires F_0

F_k, $k \geq 2$ indicate degree of SKEW

Important in many parallel DB applications

Can determine selection of algorithm.

(See intro of AMS for more sales talk...)

Thus, beneficial to maintain estimates for F_k

1) Needs to be done online as elements inserted into database, i.e., small time

2) Needs to stay in primary memory to be fast
large memory requirements ⇒ store data on disk ⇒ 10^6 access and update times.
i.e., small memory

STREAMING ALGORITHMS (simplest version)

- One pass over data
- Sublinear memory (measured in input size)

This area of research essentially started in AMS 99

Influential paper — Gödel Prize 2005

Today whole area of research.

This and next lecture

- Algorithms streaming for frequency moments (upper bounds)
- Impossibility proofs (lower bounds; uses CC)

But first quick recap of probability theory

(SELECTIVE) RECAP OF PROBABILITY THEORY

Focus on discrete random variables X

Can take values x_i, $i = 1, 2, \ldots, n$

$$\Pr[X = x_i] = p_i \geq 0 \quad \left(\sum_{i=1}^{n} p_i = 1\right)$$

"Events" are subsets S, T of possible outcomes

INCLUSION-EXCLUSION PRINCIPLE

$$\Pr[S \cup T] = \Pr[S] + \Pr[T] - \Pr[S \cap T]$$

UNION BOUND

For any events $S_1, S_2, \ldots, S_m$, the probability that one or more of these occurs is upper-bounded by

$$\Pr\left[\bigcup_{i=1}^{m} S_i\right] \leq \sum_{i=1}^{n} \Pr[S_i]$$

CONDITIONAL PROBABILITY

Probability of event S given that T is known to have occurred

$$\Pr[S \mid T] = \frac{\Pr[S \cap T]}{\Pr[T]}$$

INDEPENDENCE

Two events S and T are independent if

$$\Pr[S \cap T] = \Pr[S] \cdot \Pr[T]$$

E.g. for discrete random variables X, Y they are independent if

$$\Pr[X = x_i, Y = y_j] = \Pr[X = x_i] \cdot \Pr[Y = y_j]$$

EXPECTATION $\mathcal{E}[X] = \sum_{i=1}^{n} p_i \cdot x_i$

"expected value" or "mean" $= \sum_{i=1}^{n} Pr[X = x_i] \cdot x_i$

Linearity of expectation

$$\mathcal{E}[\alpha X + \beta Y] = \alpha \mathcal{E}[X] + \beta \mathcal{E}[Y]$$

VARIANCE $Var(X) = \mathcal{E}[(X - \mathcal{E}[X])^2]$

$= \mathcal{E}[X^2] - (\mathcal{E}[X])^2$

$Var(c \cdot X) = c^2 \, Var(X)$

For <u>independent</u> random variables $X_1, \ldots, X_s$

$$Var\left(\sum_{i=1}^{s} X_i\right) = \sum_{i=1}^{s} Var(X_i)$$

Let all X_i be identically distributed and independent (i.e., repetitions of same experiment).

Let $Y = \sum_{i=1}^{s} X_i / s$

Then $Var(Y) = Var(X_i)/s$

STANDARD DEVIATION $= \sqrt{\text{variance}}$

MARKOV'S INEQUALITY for $X \geq 0$

$$Pr[X \geq a] \leq \mathcal{E}[X]/a$$

<u>Proof</u> Define indicator random variable $I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$

Always $I \leq X/a$. Take expectation on both sides.

Useful reformulation of Markov's inequality

$$Pr[\ X \geq c \cdot \mathbb{E}[X]\ ] \leq 1/c$$

"The probability that X exceeds its expectation by factor c is at most $1/c$."

Not very strong bound, but holds for <u>any X</u> (≥ 0).

CHEBYSHEV'S INEQUALITY

"In any data sample (in particular random samples from probability distributions), nearly all values are close to the mean"

More precisely: "No more than $1/s^2$ fraction of distributions values can be more than s standard deviations away from mean."

Suppose X has mean μ and variance σ^2 (both $< \infty$). Then

$$Pr[\ |X - \mu| \geq s\sigma\ ] \leq 1/s^2$$

<u>Proof</u> Consider $Y = (X - \mathbb{E}[X])^2$ and apply Markov's inequality with $a = (s\sigma)^2$.

$$Pr\left[(X - \mathbb{E}[X])^2 \geq s^2\sigma^2\right] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{s^2\sigma^2} = 1/s^2.$$

$$\| \qquad Pr[\ |X - \mathbb{E}[X]| \geq s\sigma\ ]$$

CHERNOFF BOUND

In fact several flavours of bounds/inequalities

"If r.v. is sampled many times, then the average value converges (quickly) to the expected value."

$x_1, \ldots, x_n$ independent 0-1 r.v.

$$Pr[x_i = 1] = p$$
$$Pr[x_i = 0] = 1-p$$

$\sum_{i=1}^{n} x_i / n$ should be close to $E[x_i] = p$

$$Pr\left[\left|\frac{\sum_{i=1}^{n} x_i}{n} - p\right| \geq \delta\right]$$

$$\leq 2 \cdot \exp\left(-\frac{\delta^2 n}{2p(1-p)}\right)$$

$$= \exp(-\Omega(n))$$

exponential decrease in n = # repeated experiments.

* * * * *

Now back to frequency moments of data streams and how to estimate them...

Each moment F_k can be computed exactly and deterministically with $O(n \log m)$ bits.

Simply compute all numbers m_i

Use randomness to do better!

Example 1 (not very good) (and not very related)

Say we have a stream $A = (a_1, a_2, \ldots, a_m)$

Want to compute average $\frac{1}{m} \sum_{i=1}^{m} a_i$

Pick uniformly and randomly $i \in [m] = \{1, \ldots, m\}$

Store a_i

Pros: Correct in expectation
Uses very little memory

Cons: Can vary wildly (variance large)
How do we know we're even close to being right

Want to define random variable X s.t.

a) $\mathbb{E}[X] = F_k$
b) X can be computed in small space
c) Variance of X small

This is our goal.

We will prove (or at least sketch proof) of following thm

Thm 2 For every $k \geq 1$, every $\lambda > 0$, every $\varepsilon > 0$, there is a randomized algorithm that given $A = (a_1, \ldots, a_m)$, $a_i \in N - \{1, \ldots, n\}$, makes one pass over A, computes Y s.t.

$$P_r\left[\, |Y - F_k| > \lambda F_k \,\right] \leq \varepsilon$$

and uses

$$O\left(\frac{k \log(1/\varepsilon)}{\lambda^2} \, n^{1-1/k} \, (\log n + \log m) \right)$$

bits of memory

{Asking for closer approximation or lower error probability will increase memory, but only by constant factors}.

First try [assume for simplicity m known, can be fixed later]

Let $p \in \{1, \ldots m\}$ uniformly random.

Suppose $a_p = \ell$

Let $r = |\{q \mid q \geq p, a_q = \ell\}|$

= number of occurrences of $a_p = \ell$ starting from position p. Define

$$X = m\left(r^k - (r-1)^k \right)$$

To compute X, need $\log n$ bits for a_p
+ $\log m$ bits for r

What is $E[X]$?

(Mentally) sort A by occurrences/copies of distinct elements. Note that all are equally likely to be chosen, so divide by $1/m$ to get expectation.

$$\underbrace{m\left(m_1^k-(m_1-1)^k\right)}_{\text{1st copy of 1}} + \underbrace{m\left((m_1-1)^k-(m_1-2)^k\right)}_{\text{2nd copy of 1}} + \cdots + \underbrace{m\left(2^k-1^k\right)}_{\text{next to last copy of 1}} + \underbrace{m\,1^k}_{\text{last copy of 1}}$$

$$\underbrace{m\left(m_2^k-(m_2-1)^k\right)}_{\text{1st copy of 2}} + \underbrace{m\left((m_2-1)^k-(m_2-2)^k\right)}_{\text{2nd copy of 2}} + \cdots + \underbrace{m\left(2^k-1^k\right)}_{\text{next to last}} + \underbrace{m\,1^k}_{\text{last}}$$

$$\vdots$$

$$\underbrace{m\left(m_n^k-(m_n-1)^k\right)}_{\text{1st copy of } n} + \underbrace{m\left((m_n-1)^k-(m_n-2)^k\right)}_{\text{2nd copy of } n} + \cdots + \underbrace{m\left(2^k-1^k\right)}_{\text{next to last}} + \underbrace{m\,1^k}_{\text{last}}$$

Sum all and divide by m — telescoping sum

$$\frac{m}{m}\left(m_1^k + m_2^k + \cdots + m_n^k\right) = F_k$$

So expectation correct!

What about variance?

Use inequality for $a > b > 0$

$$a^k - b^k = (a-b)\left(a^{k-1} + a^{k-2}b + \dots + ab^{k-2} + b^{k-1}\right)$$
$$\leq (a-b)\, k\, a^{k-1}$$

To estimate variance, bound $E[X^2] =$

$$\frac{m^2}{m}\Big[\left(m_1^k - (m_1-1)^k\right)^2 + \dots + \left(2^k - 1^k\right)^2 + 1^{2k}$$
$$+ \left(m_2^k - (m_2-1)^k\right)^2 + \dots + \left(2^k - 1^k\right)^2 + 1^{2k} +$$
$$\dots$$
$$+ \left(m_n^k - (m_n-1)^k\right)^2 + \dots + \left(2^k - 1^k\right)^2 + 1^{2k}\Big]$$

$\leq \dots$ [see AMS '99 page 139]

$$\leq m\left[k\, m_1^{2k-1} + k\, m_2^{2k-1} + \dots k\, m_n^{2k-1}\right]$$

$$= k\, m\, F_{2k-1} = k\, F_1\, F_{2k-1} = *$$

Would like sth in terms of F_k, though...

Fact: For any $m_i > 0$

$$\left(\sum_{i=1}^n m_i\right)\left(\sum_{i=1}^n m_i^{2k-1}\right) \leq n^{1-1/k}\left(\sum_{i=1}^n m_i^k\right)^2$$

[Again see AMS '99 page 139]

$$\mathrm{Var}(X) \leq E[X^2] \leq k \cdot n^{1-1/k}\, F_k^2$$

This variance is too high... How to decrease it? Take many independent samples and average!

Namely $s_1 = 8 k n^{1-1/k} / \lambda^2$ samples

Let $Y = \sum_{i=1}^{s_1} X_i / s_1$

$$Var(Y) \le \frac{k \cdot n^{1-1/k} F_k^2}{8 k n^{1-1/k} / \lambda^2} = \frac{\lambda^2}{8} \cdot F_k^2$$

How close should we expect that Y gets to F_k?

Use Chebyshev's inequality

$$Pr\left[|Y - F_k| > \lambda F_k\right] \le Pr\left[|Y - F_k| > \sqrt{8 \cdot Var(Y)}\right]$$

$$\le 1/8 \qquad (\dagger)$$

So with constant probability we are close. Not good enough! Want ε - small error probability!

Do many independent samples of Y, and upper-bound probability that half of these samples fails to get within λF_k of correct answer.

Do $s_2 = 4 \log(1/\varepsilon)$ samples $Y_1, \ldots, Y_{s_2}$ of Y and pick median. This median is within λF_k of correct answer unless (†) fails to hold for 1/2 of the experiments, but should only fail for $\approx 1/8$ fraction.

$$\Pr[\text{half of } s_2 \text{ } Y\text{'s wrong}]$$

$$= \Pr\left[\text{3/8 more than expected } Y\text{'s wrong}\right]$$

$$\leq \Pr\left[\left| \frac{\sum_{i=1}^{s_2} Y_i}{s_2} - \frac{1}{8} \right| \geq \frac{3}{8}\right]$$

$$\leq 2 \exp\left(- \frac{\frac{3}{8}\cdot\frac{3}{8}\, s_2}{2 \cdot \frac{1}{8}\cdot\frac{7}{8}} \right)$$

$$\leq 2 \exp\left(-2 \log(1/\varepsilon)\right)$$

$$= 2 \exp\left(2 \log \varepsilon\right) = 2\varepsilon^2 < \varepsilon.$$

So the median of $4 \log(1/\varepsilon)$ Y's will get us within λF_k of right answer with prob $> 1-\varepsilon$.

Need a total of $4 \log(1/\varepsilon) \cdot 8 k n^{1-1/k} / \lambda^2$ to compute

X_{ij} samples, each requiring space $\approx \log n + \log m$.

QED.

Except we don't know m = length of A!

What to do?

Suppose we have run algorithm for upto m elements. Now $(m+1)$st element comes along

We have computed all x_{ij} so far for

$i = 1, \ldots, s_1$, $j = 1, \ldots, s_2$

Every x_{ij} has its a_p (uniform over $a_1, \ldots, a_m$) and r

With probability $1/(m+1)$ (independently for every x_{ij}), replace a_p by a_{m+1} and update $r = 1$

With prob $1 - 1/(m+1)$, keep a_p and increment r if $a_{m+1} = a_p$

Then after this step all $a_1, a_2, \ldots, a_{m+1}$ are equally likely to have been chosen, and r is correct.

This completes the proof (modulo computations that we skipped).

In particular, can compute F_2 (or at least approximate closely with arbitrarily high probability) in

$$O\left(\sqrt{n}\,(\log n + \log m)\right)$$ memory

(fixing closeness λ and probability $1-\varepsilon$).

But AMS '99 shows possible to do <u>much better</u>

<u>Thm 3</u> For every $\lambda > 0$, $\varepsilon > 0$ $\exists$ randomized algorithm that given $A = (a_1, \ldots, a_m)$ $a_i \in N = \{1, 2, \ldots, n\}$, makes <u>one pass</u> over A, computes Y s.t.

$$Pr\left[\,|Y - F_2| > \lambda F_2\,\right] \leq \varepsilon$$

and uses $O\left(\frac{\log(1/\varepsilon)}{\lambda^2}\,(\log n + \log m)\right)$ bits of memory.

Proof uses same general idea, but needs some neat ideas about somewhat-but-not-quite independent ~~distributions~~ random variables and some nontrivial abstract algebra.

OK, nice algorithms. But let's not ignore the disadvantages:

(1) Uses lots of memory for large k

(2) Doesn't give correct answer, only approximation

(3) And even this only with high probability over random choices. But computers are deterministic, so shouldn't there be a deterministic algorithm.

Next lecture Show that limitations (1) - (3) are inherent.

It is impossible to get around them!