CS 229r: Algorithms for Big Data Fall 2015

Lecture 2 — Sept. 8, 2015

Prof. Jelani Nelson *Scribe: Jeffrey Ling*

1 Probability Recap

Chebyshev: $P(|X - \mathbb{E}X| > \lambda) < \frac{Var[X]}{\lambda^2}$

Chernoff: For $X_1, \ldots, X_n$ independent in $[0,1]$, $\forall 0 < \epsilon < 1$, and $\mu = \mathbb{E}\sum_i X_i$,

$$P(|\sum_i X_i - \mu| > \epsilon\mu) < 2e^{-\epsilon^2\mu/3}$$

2 Today

- Distinct elements
- Norm estimation (if there's time)

3 Distinct elements (F_0)

Problem: Given a stream of integers $i_1, \ldots, i_m \in [n]$ where $[n] := \{1, 2, \ldots, n\}$, we want to output the number of distinct elements seen.

3.1 Straightforward algorithms

1. Keep a bit array of length n. Flip bit if a number is seen.
2. Store the whole stream. Takes $m \lg n$ bits.

We can solve with $O(\min(n, m \lg n))$ bits.

3.2 Randomized approximation

We can settle for outputting $\widetilde{t}$ s.t. $P(|t - \widetilde{t}| > \epsilon t) < \delta$. The original solution was by Flajolet and Martin [2].

3.3 Idealized algorithm

1. Pick random function $h : [n] \to [0, 1]$ (idealized, since we can't actually nicely store this)
2. Maintain counter $X = \min_{i \in stream} h(i)$
3. Output $1/X - 1$

Intuition. X is a random variable that's the minimum of t i.i.d $Unif(0, 1)$ r.v.s.

Claim 1. $\mathbb{E}X = \frac{1}{t+1}$.

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_0^\infty P(X > \lambda) d\lambda \\
&= \int_0^\infty P(\forall i \in str, h(i) > \lambda) d\lambda \\
&= \int_0^\infty \prod_{r=1}^{t} P(h(i_r) > \lambda) d\lambda \\
&= \int_0^1 (1 - \lambda)^t d\lambda \\
&= \frac{1}{t+1}
\end{aligned}$$

□

Claim 2. $\mathbb{E}X^2 = \frac{2}{(t+1)(t+2)}$

Proof.

$$\begin{aligned}
\mathbb{E}X^2 &= \int_0^1 P(X^2 > \lambda) d\lambda \\
&= \int_0^1 P(X > \sqrt{\lambda}) d\lambda \\
&= \int_0^1 (1 - \sqrt{\lambda})^t d\lambda \qquad\qquad u = 1 - \sqrt{\lambda} \\
&= 2 \int_0^1 u^t (1 - u) du \\
&= \frac{2}{(t+1)(t+2)}
\end{aligned}$$

□

This gives $Var[X] = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{t}{(t+1)^2(t+2)}$, and furthermore $Var[X] < \frac{1}{(t+1)^2} = (\mathbb{E}X)^2$.

4 FM+

We average together multiple estimates from the idealized algorithm FM.

1. Instantiate $q = 1/\epsilon^2\eta$ FMs independently
2. Let X_i come from FM_i.
3. Output $1/Z - 1$, where $Z = \frac{1}{q}\sum_i X_i$.

We have that $\mathbb{E}(Z) = \frac{1}{t+1}$, and $Var(Z) = \frac{1}{q}\frac{t}{(t+1)^2(t+2)} < \frac{1}{q(t+1)^2}$.

Claim 3. $P(|Z - \frac{1}{t+1}| > \frac{\epsilon}{t+1}) < \eta$

Proof. Chebyshev.

$$P(|Z - \frac{1}{t+1}| > \frac{\epsilon}{t+1}) < \frac{(t+1)^2}{\epsilon^2}\frac{1}{q(t+1)^2} = \eta$$

□

Claim 4. $P(|(\frac{1}{Z} - 1) - t| > O(\epsilon)t) < \eta$

Proof. By the previous claim, with probability $1 - \eta$ we have

$$\frac{1}{(1 \pm \epsilon)\frac{1}{t+1}} - 1 = (1 \pm O(\epsilon))(t+1) - 1 = (1 \pm O(\epsilon))t \pm O(\epsilon)$$

□

5 FM++

We take the median of multiple estimates from FM+.

1. Instantiate $s = \lceil 36\ln(2/\delta)\rceil$ independent copies of FM+ with $\eta = 1/3$.
2. Output the median $\widehat{t}$ of $\{1/Z_j - 1\}_{j=1}^{s}$ where Z_j is from the jth copy of FM+.

Claim 5. $P(|\widehat{t} - t| > \epsilon t) < \delta$

Proof. Let

$$Y_j = \begin{cases} 1 & \text{if } |(1/Z_j - 1) - t| > \epsilon t \\ 0 & \text{else} \end{cases}$$

We have $\mathbb{E}Y_j = P(Y_j = 1) < 1/3$ from the choice of η. The probability we seek to bound is equivalent to the probability that the median fails, i.e. at least half of the FM+ estimates have $Y_j = 1$. In other words,

$$\sum_{j=1}^{s} Y_j > s/2$$

We then get that

$$P(\sum Y_j > s/2) = P(\sum Y_j - s/3 > s/6) \tag{1}$$

Make the simplifying assumption that $\mathbb{E}Y_j = 1/3$ (this turns out to be stronger than $\mathbb{E}Y_j < 1/3$. Then equation 1 becomes

$$P(\sum Y_j - \mathbb{E}\sum Y_j > \frac{1}{2}\mathbb{E}\sum Y_j)$$

using Chernoff,

$$< e^{-\frac{(\frac{1}{2})^2 s/3}{3}} < \delta$$

as desired. □

The final space required, ignoring h, is $O(\frac{\lg(1/\delta)}{\epsilon^2})$ for $O(\lg(1/\delta))$ copies of FM+ that require $O(1/\epsilon^2)$ space each.

6 k-wise independent functions

Definition 6. *A family $\mathcal{H}$ of functions mapping $[a]$ to $[b]$ is k-wise independent if $\forall j_1, \ldots, j_k \in [b]$ and $\forall$ distinct $i_1, \ldots, i_k \in [a]$,*

$$P_{h \in \mathcal{H}}(h(i_1) = j_1 \wedge \ldots \wedge h(i_k) = j_k) = 1/b^k$$

Example. The set $\mathcal{H}$ of all functions $[a] \to [b]$ is k-wise independent for every k. $|\mathcal{H}| = b^a$ so h is representable in $a \lg b$ bits.

Example. Let $a = b = q$ for $q = p^r$ a prime power, then $\mathcal{H}_{poly}$, the set of degree $\leq k-1$ polynomials with coefficients in $\mathbb{F}_q$, the finite field of order q. $|\mathcal{H}_{poly}| = q^k$ so h is representable in $k \lg p$ bits.

Claim 7. *$\mathcal{H}_{poly}$ is k-wise independent.*

Proof. Interpolation. □

7 Non-idealized FM

First, we get an $O(1)$-approximation in $O(\lg n)$ bits, i.e. our estimate $\widetilde{t}$ satisfies $t/C \leq \widetilde{t} \leq Ct$ for some constant C.

1. Pick h from 2-wise family $[n] \to [n]$, for n a power of 2 (round up if necessary)
2. Maintain $X = \max_{i \in str} lsb(h(i))$ where lsb is the least significant bit of a number
3. Output 2^X

For fixed j, let Z_j be the number of i in stream with $lsb(h(i)) = j$. Let $Z_{>j}$ be the number of i with $lsb(h(i)) > j$.

Let

$$Y_i = \begin{cases} 1 & lsb(h(i)) = j \\ 0 & \text{else} \end{cases}$$

Then $Z_j = \sum_{i \in str} Y_i$. We can compute $\mathbb{E}Z_j = t/2^{j+1}$ and similarly

$$\mathbb{E}Z_{>j} = t(\frac{1}{2^{j+2}} + \frac{1}{2^{j+3}} + \ldots) < t/2^{j+1}$$

and also

$$Var[Z_j] = Var[\sum Y_i] = \mathbb{E}(\sum Y_i)^2 - (\mathbb{E} \sum Y_i)^2 = \sum_{i_1, i_2} \mathbb{E}(Y_{i_1} Y_{i_2})$$

Since h is from a 2-wise family, Y_i are pairwise independent, so $\mathbb{E}(Y_{i_1} Y_{i_2}) = \mathbb{E}(Y_{i_1})\mathbb{E}(Y_{i_2})$. We can then show

$$Var[Z_j] < t/2^{j+1}$$

Now for $j^* = \lceil \lg t - 5 \rceil$, we have

$$16 \leq \mathbb{E}Z_{j^*} \leq 32$$

$$P(Z_{j^*} = 0) \leq P(|Z_{j^*} - \mathbb{E}Z_{j^*}| \geq 16) < 1/5$$

by Chebyshev.

For $j = \lceil \lg t + 5 \rceil$

$$\mathbb{E}Z_{>j} \leq 1/16$$

$$P(Z_{>j} \geq 1) < 1/16$$

by Markov.

This means with good probability the max lsb will be above j^* but below j, in a constant range. This gives us a 32-approximation, i.e. constant approximation.

8 Refine to $1 + \epsilon$

Trivial solution. Algorithm TS stores first C/ϵ^2 distinct elements. This is correct if $t \leq C/\epsilon^2$.

Algorithm.

1. Instantiate $\text{TS}_0, \ldots, \text{TS}_{\lg n}$
2. Pick $g : [n] \to [n]$ from 2-wise family
3. Feed i to $\text{TS}_{lsb(g(i))}$
4. Output 2^{j+1} out where $t/2^{j+1} \approx 1/\epsilon^2$.

Let B_j be the number of distinct elements hashed by g to TS_j. Then $\mathbb{E}B_j = t/2^{j+1} = Q_j$. By Chebyshev $B_j = Q_j \pm O(\sqrt{Q_j})$ with good probability. This equals $(1 \pm O(\epsilon))Q_j$ if $Q_j \geq 1/\epsilon^2$.

Final space: $\frac{C}{\epsilon^2}(\lg n)^2 = O(\frac{1}{\epsilon^2}\lg^2 n)$ bits.

It is known that space $O(1/\epsilon^2 + \log n)$ is achievable [4], and furthermore this is optimal [1, 5] (also see [3]).

References

[1] Noga Alon, Yossi Matias, Mario Szegedy The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.* 58(1): 137–147, 1999.

[2] Philippe Flajolet, G. Nigel Martin Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[3] T. S. Jayram, Ravi Kumar, D. Sivakumar: The One-Way Communication Complexity of Hamming Distance. *Theory of Computing* 4(1): 129–135, 2008.

[4] Daniel M. Kane, Jelani Nelson, David P. Woodruff An optimal algorithm for the distinct elements problem. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 41–52, 2010.

[5] David P. Woodruff. Optimal space lower bounds for all frequency moments. In *SODA*, pages 167–175, 2004.