1 Probability Recap

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

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

\[
P(\left| \sum_i X_i - \mu \right| > \epsilon \mu) < 2e^{-\epsilon^2 \mu/3}
\]

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 \( \tilde{t} \) s.t. \( P(|t - \tilde{t}| > \epsilon t) < \delta \). The original solution was by Flajolet and Martin [2].
3.3 Idealized algorithm

1. Pick random function \( h : [n] \rightarrow [0, 1] \) (idealized, since we can’t actually nicely store this)

2. Maintain counter \( X = \min_{i \in \text{stream}} h(i) \)

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. \( E X = \frac{1}{t+1} \).

Proof.

\[
E X = \int_0^\infty P(X > \lambda) d\lambda \\
= \int_0^\infty P(\forall i \in \text{str}, h(i) > \lambda) d\lambda \\
= \int_0^\infty \prod_{r=1}^{t} P(h(i_r) > \lambda) d\lambda \\
= \int_0^1 (1 - \lambda)^t d\lambda \\
= \frac{1}{t+1}
\]

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

Proof.

\[
E X^2 = \int_0^1 P(X^2 > \lambda) d\lambda \\
= \int_0^1 P(X > \sqrt{\lambda}) d\lambda \\
= \int_0^1 (1 - \sqrt{\lambda})^t d\lambda \\
= 2 \int_0^1 u^t(1-u) du \\
= \frac{2}{(t+1)(t+2)}
\]

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

We average together multiple estimates from the idealized algorithm FM.

1. Instantiate \( q = \frac{1}{\epsilon^2} \) FMs independently
2. Let \( X_i \) come from FM\(_i\).
3. Output \( \frac{1}{Z} - 1 \), where \( Z = \frac{1}{q} \sum_i X_i \).

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

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

**Proof.** Chebyshev.

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

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

**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 \( \hat{t} \) of \( \{1/Z_j - 1\} \)\(_{j=1}^s\) where \( Z_j \) is from the \( j \)th copy of FM++.

**Claim 5.** \( P(\left| \hat{t} - t \right| > \epsilon t) < \delta \)

**Proof.** Let

\[
Y_j = \begin{cases} 
1 & \text{if } \left| \frac{1}{Z_j} - 1 - t \right| > \epsilon t \\
0 & \text{else}
\end{cases}
\]

We have \( \mathbb{E}Y_j = P(Y_j = 1) < 1/3 \) from the choice of \( \eta \). 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
\[ P(\sum Y_j > s/2) = P(\sum Y_j - s/3 > s/6) \quad (1) \]

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,
\[ < e^{-\left(\frac{1}{2}\right)^2 s/3} < \delta \]
as desired.

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

6  \textit{k-wise independent functions}

\textbf{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 \( i_1, \ldots, i_k \in [a] \),
\[ P_{h \in \mathcal{H}}(h(i_1) = j_1 \land \ldots \land h(i_k) = j_k) = \frac{1}{b^k} \]

\textbf{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 \( a \log b \) bits.

\textbf{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 \( k \log p \) bits.

\textbf{Claim 7.} \( \mathcal{H}_{poly} \) is \( k \)-wise independent.

\textit{Proof.} Interpolation.

7  \textit{Non-idealized FM}

First, we get an \( O(1) \)-approximation in \( O(\log n) \) bits, i.e. our estimate \( \tilde{t} \) satisfies \( t/C \leq \tilde{t} \leq Ct \) 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 $\text{lsb}(h(i)) = j$. Let $Z_{> j}$ be the number of $i$ with $\text{lsb}(h(i)) > j$.

Let

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

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

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

and also

$$\text{Var}[Z_j] = \text{Var}\left[\sum_{i} Y_i \right] = E\left(\sum_{i} Y_i^2\right) - (E\sum_{i} Y_i)^2 = \sum_{i_1, i_2} E(Y_{i_1} Y_{i_2})$$

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

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

Now for $j^* = \lfloor \lg t - 5 \rfloor$, we have

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

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

by Chebyshev.

For $j = \lfloor \lg t + 5 \rfloor$

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

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

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 $C/\epsilon^2$ distinct elements. This is correct if $t \leq C/\epsilon^2$. 

**Algorithm.**

1. Instantiate $\text{TS}_0, \ldots, \text{TS}_{\lg n}$
2. Pick $g : [n] \rightarrow [n]$ from 2-wise family
3. Feed $i$ to $\text{TS}_{\text{lsb}(g(i))}$
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: $\frac{C}{\epsilon^2} (\log n)^2 = O\left(\frac{1}{\epsilon^2} \log^2 n\right)$ 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


