Probabilistic Counting and Counting Distinct Elements

Christian Wulff-Nilsen
Algorithmic Techniques for Modern Data Models
DTU

October 31, 2025

Overview for today

- 2-universal hash functions
- Distinct Elements:
  - Tidemark algorithm
  - Analysis: expected output, concentration bounds
- Approximate Counting:
  - Morris Counter
  - Analysis: expected output, concentration bounds
- Law of Total Expectation with proof (if time allows)

Properties of 2-Universal Hash Functions

- Notation: for any positive integer a, $[a] = \{1, 2, \ldots, a\}$
- A hash function $h : [m] \to [n]$ is 2-*universal* if

$$\mathrm{P}[h(x_1) = y_1 \wedge h(x_2) = y_2] = \frac{1}{n^2}$$

  for all distinct $x_1, x_2 \in [m]$ and for all $y_1, y_2 \in [n]$
- Useful properties of a 2-universal hash function h:
  - h is uniform:

$$\mathrm{P}[h(x) = y] = \frac{1}{n} \text{ for all } x \in [m], y \in [n]$$

  - h hashes any two distinct values x_1, x_2 independently:

$$\mathrm{P}[h(x_1) = y_1 \wedge h(x_2) = y_2] = \mathrm{P}[h(x_1) = y_1] \cdot \mathrm{P}[h(x_2) = y_2]$$

The Distinct Elements Problem

- Given stream $\sigma = \langle a_1, \ldots, a_m \rangle$ with each $a_i \in [n]$
- This defines a frequency vector $\mathbf{f} = (f_1, \ldots, f_n)$
- Example with $n = 4$ and $m = 10$:

$$\sigma = \langle 4, 2, 4, 1, 4, 2, 4, 4, 1, 2 \rangle$$
$$\mathbf{f} = (2, 3, 0, 5)$$

- Let $d = |\{j \mid f_j > 0\}|$ be the number of distinct elements
- In the example above, $d = |\{1, 2, 4\}| = 3$
- Algorithm output: an (ϵ, δ)-estimate $\hat{d}$ of d
- This means that $\hat{d}$ should satisfy:

$$\mathrm{P}\left[\left|\frac{\hat{d}}{d} - 1\right| > \epsilon\right] \leq \delta$$

Zeros of an Integer

- For an integer $p \geq 0$, $\text{zeros}(p)$ is the number of zeros that p ends with in its binary representation
- Examples:

$$\begin{aligned} \text{zeros}(2) &= 1 && (2 \text{ is } 10 \text{ in binary}) \\ \text{zeros}(3) &= 0 && (3 \text{ is } 11 \text{ in binary}) \\ \text{zeros}(16) &= 4 && (16 \text{ is } 10000 \text{ in binary}) \\ \text{zeros}(24) &= 3 && (24 \text{ is } 11000 \text{ in binary}) \end{aligned}$$

- We can also write $\text{zeros}(p)$ as

$$\text{zeros}(p) = \max\{i \mid 2^i \text{ divides } p\}$$

- Example: $\text{zeros}(24) = 3$ since $2^3 = 8$ is the largest power of 2 that divides 24

The Tidemark Algorithm (AMS Algorithm)

- Pseudo-code:

Tidemark Algorithm

Initialize:
Choose a 2-universal hash function $h : [n] \to [n]$
$z \leftarrow 0$

Process(token j):
$z \leftarrow \max\{z, \text{zeros}(h(j))\}$

Output: $2^{z+1/2}$

- Thus, for stream $\sigma = \langle a_1, a_2, \ldots, a_m \rangle$, the final z value is:

$$z = \max_{i \in [m]} \{\text{zeros}(h(a_i))\}$$

- We now analyze how good an estimate to d the algorithm obtains

Analysis: Intuition

- Every d'th value in $[n]$ ends with at least $\log_2 d$ zeros
- Example with $d = 4$ and numbers of $[19]$ written in binary:

$$1 \quad 10 \quad 11 \quad \mathbf{100} \quad 101 \quad 110 \quad 111 \quad \mathbf{1000} \quad 1001 \quad 1010 \quad 1011$$
$$\mathbf{1100} \quad 1101 \quad 1110 \quad 1111 \quad \mathbf{10000} \quad 10001 \quad 10010 \quad 10011$$

- Only few of these values have significantly more than $\log_2 d$ zeros
- d values are hashed to $[n]$ over the entire stream
- Since these values are hashed uniformly, z should be close to $\log_2 d$ at termination
- This gives output:

$$2^{z+1/2} \approx 2^{\log_2 d+1/2} \approx 2^{\log_2 d} = d$$

- We now prove this more formally

Random Variables for Analysis

- Consider a token $j \in [n]$ and any integer $r \geq 0$
- $X_{r,j}$: indicator variable for the event that $h(j)$ has at least r zeros:

$$X_{r,j} = 1 \Leftrightarrow \mathrm{zeros}(h(j)) \geq r$$

- Let random variable Y_r count the number of such tokens:

$$Y_r = \sum_{j:f_j>0} X_{r,j}$$

- Note: if token j occurs, e.g., $f_j = 10$ times in the stream, it only contributes with 0 or 1 to Y_r

Relating Random Variables to Final z Value

- In the following, let z_{out} be the value of z at termination
- We have $\mathrm{z}_{out} \geq r$ if and only if for at least one token j, $\mathrm{zeros}(h(j)) \geq r$
- Since Y_r counts the number of such tokens,

$$Y_r \geq 1 \Leftrightarrow \mathrm{z}_{out} \geq r$$

- Equivalently,

$$Y_r = 0 \Leftrightarrow \mathrm{z}_{out} \leq r - 1$$

Calculating Expectations

- Since $X_{r,j}$ is an indicator variable,

$$E[X_{r,j}] = \mathrm{P}[X_{r,j} = 1] = \mathrm{P}[\mathrm{zeros}(h(j)) \geq r]$$

- h is 2-universal $\Rightarrow$ h is uniform:

$$\mathrm{P}[h(x) = i] = \frac{1}{n} \text{ for each } i, x \in [n]$$

- How many $i \in [n]$ have $\mathrm{zeros}(i) \geq r$? Only a $1/2^r$ fraction
- Thus, $h(x)$ has only a $1/2^r$ chance of hitting one such i
- This gives:

$$E[X_{r,j}] = \mathrm{P}[\mathrm{zeros}(h(j)) \geq r] = \frac{1}{2^r}$$

- By linearity of expectation:

$$E[Y_r] = \sum_{j : f_j > 0} E[X_{r,j}] = \frac{d}{2^r}$$

Concentration Bounds

- Let $\hat{d} = 2^{z_{out} + 1/2}$ be the estimate of d by the algorithm
- We will bound the probability that it deviates too much from d:

$$\mathrm{P}[\hat{d} \geq 3d] \leq \frac{\sqrt{2}}{3} \approx 0.47 \qquad \mathrm{P}[\hat{d} \leq d/3] \leq \frac{\sqrt{2}}{3} \approx 0.47$$

Showing $\mathrm{P}[\hat{d} \geq 3d] \leq \sqrt{2}/3$

- Let a be the smallest integer with $2^{a+1/2} \geq 3d$
- a is the smallest z_{out} giving output $\hat{d} \geq 3d$, so:

$$\mathrm{P}[\hat{d} \geq 3d] = \mathrm{P}[2^{\mathrm{z}_{out}+1/2} \geq 3d] = \mathrm{P}[\mathrm{z}_{out} \geq a] = \mathrm{P}[Y_a \geq 1]$$

- By Markov's inequality,

$$\mathrm{P}[Y_a \geq 1] \leq \frac{E[Y_a]}{1} = \underbrace{E[Y_a] = \frac{d}{2^a}}_{\text{shown earlier}}$$

- We then get:

$$\mathrm{P}[\hat{d} \geq 3d] = \mathrm{P}[Y_a \geq 1] \leq \underbrace{\frac{d}{2^a} \leq \frac{2^{a+1/2}/3}{2^a}}_{\text{by definition of } a} = \frac{\sqrt{2}}{3}$$

Showing $\mathrm{P}[\hat{d} \leq d/3] \leq \sqrt{2}/3$

- Let b be the largest integer with $2^{b+1/2} \leq d/3$
- b is the largest z_{out} giving output $\hat{d} \leq d/3$, so:

$$\mathrm{P}[\hat{d} \leq d/3] = \mathrm{P}[2^{\mathrm{z}_{out}+1/2} \leq d/3] = \mathrm{P}[\mathrm{z}_{out} \leq b] = \mathrm{P}[Y_{b+1} = 0]$$

- We will use Chebyshev's inequality to show that for any r:

$$\mathrm{P}[Y_r = 0] \leq \frac{2^r}{d}$$

- Since $d \geq 3 \cdot 2^{b+1/2}$, we get:

$$\mathrm{P}[\hat{d} \leq d/3] = \mathrm{P}[Y_{b+1} = 0] \leq \frac{2^{b+1}}{d} \leq \frac{2^{b+1}}{3 \cdot 2^{b+1/2}} = \frac{\sqrt{2}}{3}$$

- To use Chebyshev, we need $\mathrm{Var}[Y_r]$

Calculating $\mathrm{Var}[Y_r]$

- Recall: 2-universality of $h \Rightarrow h$ hashes any two values independently
- Since the $X_{r,j}$-variables are functions of hash values, these variables are 2-independent (exercise)
- 2-independence allows us to use linearity of variance:

$$\mathrm{Var}[Y_r] = \mathrm{Var}[\sum_{j:f_j>0} X_{r,j}] = \sum_{j:f_j>0} \mathrm{Var}[X_{r,j}]$$

- Shown later: for any random variable X, $\mathrm{Var}[X] \leq E[X^2]$
- Since $X_{r,j}$ is an indicator variable, $X_{r,j}^2 = X_{r,j}$
- Since $E[X_{r,j}] = 1/2^r$, this gives:

$$\mathrm{Var}[Y_r] = \sum_{j:f_j>0} \mathrm{Var}[X_{r,j}] \leq \sum_{j:f_j>0} E[X_{r,j}^2]$$

$$= \sum_{j:f_j>0} E[X_{r,j}] = \frac{d}{2^r}$$

Showing $\mathrm{P}[Y_r = 0] \leq 2^r/d$

- Have shown:

$$E[Y_r] = \frac{d}{2^r} \qquad \mathrm{Var}[Y_r] \leq \frac{d}{2^r}$$

- We have the following implication between events:

$$Y_r = 0 \Rightarrow |Y_r - E[Y_r]| = |E[Y_r]| \geq \frac{d}{2^r}$$

- Thus, the left-hand side is not more likely that the right-hand side

$$\mathrm{P}[Y_r = 0] \leq \underbrace{P\left[|Y_r - E[Y_r]| \geq \frac{d}{2^r}\right]}_{\text{right form for Chebyshev}}$$

- Chebyshev:

$$\mathrm{P}[Y_r = 0] \leq P\left[|Y_r - E[Y_r]| \geq \frac{d}{2^r}\right] \leq \frac{\mathrm{Var}[Y_r]}{(d/2^r)^2} \leq \frac{d/2^r}{(d/2^r)^2} = \frac{2^r}{d}$$

Showing $\mathrm{Var}[X] \leq E[X^2]$ (Used Earlier)

- Lemma: For any random variable X, we have

$$\mathrm{Var}[X] = E[X^2] - E[X]^2$$

- Proof:

$$\begin{aligned}\mathrm{Var}[X] &\overset{\text{def}}{=} E[(X - E[X])^2] \\ &= E[X^2 + E[X]^2 - 2XE[X]] \\ &= E[X^2] + E[X]^2 - 2E[X]^2 \\ &= E[X^2] - E[X]^2\end{aligned}$$

- Corollary: For any random variable X, we have $\mathrm{Var}[X] \leq E[X^2]$
- Proof:

$$\mathrm{Var}[X] = E[X^2] - \overbrace{E[X]^2}^{\geq 0} \leq E[X^2]$$

Approximate Counting

- Problem:
  - Count the length n of the stream seen so far ($n \leq m$)
  - Use as few bits as possible for this
- Trivial with $O(\log m)$ bits (how?)
- This is in fact optimal
- We can do better if we only need an estimate of m:
  - We analyze the *Morris counter*
  - With slight modifications, it can obtain an (ϵ, δ)-estimate using only $O(\log \log m)$ bits (for constant ϵ and δ) (exercise)
  - Instead, we show that its output is an unbiased estimator of n

Estimating m: The Morris Counter

- Space-efficient version:

Morris Counter

Initialize: $x \leftarrow 0$
Process(token): with probability 2^{-x} update $x \leftarrow x + 1$
Output: $2^x - 1$

- Space-inefficient version:

Space-inefficient Morris Counter

Initialize: $c \leftarrow 1$
Process(token): with probability $1/c$ update $c \leftarrow 2c$
Output: $c - 1$

- The algorithms give the same output since in each iteration, $c = 2^x$
- We focus on the second version since it is easier to analyze

Unbiased Estimator

- Pseudo-code:

Space-inefficient Morris Counter

Initialize: $c \leftarrow 1$
Process(token): with probability $1/c$ update $c \leftarrow 2c$
Output: $c - 1$

- Let C_i be c after processing i tokens ($C_0 = 1$)
- The output after n tokens is $C_n - 1$
- Need to show that $C_n - 1$ is an *unbiased estimator* of n:

$$E[C_n - 1] = n$$

Indicator Variable Z_i

- Pseudo-code:

Space-inefficient Morris Counter

Initialize: $c \leftarrow 1$
Process(token): with probability $1/c$ update $c \leftarrow 2c$
Output: $c - 1$

- Z_i: indicates if c doubles when processing token $i + 1$
- Thus, Z_i is 1 if $C_{i+1} = 2C_i$ and 0 if $C_{i+1} = C_i$:

$$C_{i+1} = C_i(1 + Z_i)$$

- When processing token $i + 1$, the probability $1/c$ is $1/C_i$ (not $1/C_{i+1}$) since we update c to C_{i+1} *after* the random choice:

$$E[Z_i \mid C_i] = \mathrm{P}[Z_i = 1 \mid C_i] = 1/C_i$$

Relating $E[C_{i+1}]$ and $E[C_i]$

- Indicator variable Z_i: is 1 if $C_{i+1} = 2C_i$ and 0 if $C_{i+1} = C_i$

$$C_{i+1} = C_i(1 + Z_i) \qquad E[Z_i \mid C_i] = 1/C_i$$

- Law of total expectation: for any random variables X and Y:

$$E[X] = E[E[X \mid Y]]$$

- Applying this with $X = C_{i+1}$ and $Y = C_i$:

$$\begin{aligned} E[C_{i+1}] &= E[E[C_{i+1} \mid C_i]] \\ &= E[E[C_i(1 + Z_i) \mid C_i]] \\ &= E[C_i(1 + E[Z_i \mid C_i])] \\ &= E[C_i(1 + 1/C_i)] \\ &= 1 + E[C_i] \end{aligned}$$

Unbiased Estimator: showing $E[C_n - 1] = n$

- Have shown that for each i:

$$E[C_{i+1}] = 1 + E[C_i]$$

- Since $C_0 = 1$, we have:

$$\begin{aligned} E[C_1] &= 1 + E[C_0] = 1 + 1 = 2 \\ E[C_2] &= 1 + E[C_1] = 1 + 2 = 3 \\ &\dots \\ E[C_n] &= 1 + E[C_{n-1}] = n + 1 \end{aligned}$$

- Thus $E[C_n - 1] = n$
- In words, $C_n - 1$ is an unbiased estimator of n
- Next step: if possible, show that $\mathrm{Var}[C_n] = \mathrm{Var}[C_n - 1]$ is small in order to get a high concentration bound with Chebyshev

Bounding $\mathrm{Var}[C_n]$

- Our Lemma from earlier gives: $\mathrm{Var}[C_n] = E[C_n^2] - E[C_n]^2$
- We already showed $E[C_n] = n + 1$ so $E[C_n]^2 = (n+1)^2$
- We will show:

$$E[C_n^2] = 1 + \frac{3n(n+1)}{2}$$

- This will give us:

$$\begin{aligned}\mathrm{Var}[C_n] &= E[C_n^2] - E[C_n]^2 \\ &= 1 + \frac{3n(n+1)}{2} - (n+1)^2 \\ &= 1 + \frac{3}{2}n^2 + \frac{3}{2}n - n^2 - 1 - 2n \\ &= \frac{n(n-1)}{2}\end{aligned}$$

- This variance is too large for Chebyshev to be useful
- We deal with this in Exercise 4-1 (Streaming notes)

Bounding $E[C_{i+1}^2]$ in Terms of $E[C_i^2]$

- Using the law of total expectation:

$$\begin{aligned} E[C_{i+1}^2] &= E[E[C_{i+1}^2 \mid C_i]] \\ &= E[E[((1+Z_i)C_i)^2 \mid C_i]] \\ &= E[E[(Z_i^2 + 2Z_i + 1)C_i^2 \mid C_i]] \\ &= E[E[(3Z_i + 1)C_i^2 \mid C_i]] \\ &= E[3C_i^2 E[Z_i \mid C_i] + C_i^2] \\ &= E[3C_i^2 \cdot 1/C_i + C_i^2] \\ &= E[3C_i + C_i^2] \\ &= 3E[C_i] + E[C_i^2] \\ &= 3(i+1) + E[C_i^2] \end{aligned}$$

Showing $E[C_n^2] = 1 + 3n(n+1)/2$

- Have shown $E[C_{i+1}^2] = 3(i+1) + E[C_i^2]$ for $i \geq 0$
- This is equivalent to $E[C_i^2] = 3i + E[C_{i-1}^2]$ for $i \geq 1$
- We sum up all these contributions to obtain $E[C_n^2]$:

$$
\begin{aligned}
E[C_0^2] &= 1^2 = 1 \\
E[C_1^2] &= 3(0+1) + E[C_0^2] = 3 \cdot 1 + 1 \\
E[C_2^2] &= 3(1+1) + E[C_1^2] = 3 \cdot 2 + 3 \cdot 1 + 1 \\
E[C_3^2] &= 3(2+1) + E[C_2^2] = 3 \cdot 3 + 3 \cdot 2 + 3 \cdot 1 + 1 \\
&\dots \\
E[C_n^2] &= 1 + 3\sum_{i=1}^{n} i = 1 + \frac{3n(n+1)}{2}
\end{aligned}
$$

Law of Total Expectation with Proof

- For two random variables X and Y, $E[X] = E[E[X \mid Y]]$
- Proof, where $g(Y) = E[X \mid Y]$:

$$
\begin{aligned}
E[E[X \mid Y]] = E[g(Y)] &= \sum_y g(y) \cdot \mathrm{P}[Y = y] \\
&= \sum_y E[X \mid Y = y] \cdot \mathrm{P}[Y = y] \\
&= \sum_y \sum_x x \cdot \mathrm{P}[X = x \mid Y = y] \cdot \mathrm{P}[Y = y] \\
&= \sum_y \sum_x x \cdot \mathrm{P}[Y = y \mid X = x] \cdot \mathrm{P}[X = x] \\
&= \sum_x x \cdot \mathrm{P}[X = x] \cdot \sum_y \mathrm{P}[Y = y \mid X = x] \\
&= \sum_x x \cdot \mathrm{P}[X = x] \cdot 1 \\
&= E[X]
\end{aligned}
$$