Hashing, Bloom Filters, and Morris

Alperen Ergur

Contents

1 Moment Generating Functions, Bounds, and Morris

1.1 MGFs

- $M_X(t) := \mathbb{E}[e^{tX}]$

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}, \quad M_X(t) = pe^t + (1-p)$$

Theorem:

(a) $M_X^{(n)}(0) = \mathbb{E}[X^n]$

(b) If $M_X(t) = M_Y(t)$ $\forall t \in (-\delta, \delta)$ then $X \stackrel{d}{=} Y$

(c) If X, Y independent: $M_{X+Y}(t) = M_X(t)M_Y(t)$

1.2 MGF Example

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables.

Each X_i is distributed as:

$$X_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

Define the sum:

$$X = X_1 + X_2 + \cdots + X_n$$

Moment generating function of X_1:

$$\mathbb{E}[e^{tX_1}] = \frac{e^t + e^{-t}}{2}$$

Using the inequality:

$$\frac{e^t + e^{-t}}{2} \leq e^{\frac{t^2}{2}}$$

Moment generating function of X:

$$M_X(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \cdots \cdot M_{X_n}(t)$$

Since the X_i are i.i.d., this simplifies to:

$$M_X(t) = (M_{X_1}(t))^n \leq e^{\frac{t^2 n}{2}}$$

Therefore:

$$M_{X_1}(t) \leq e^{\frac{t^2}{2}}$$

1.3 Tail Bounds

$$X_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

$X = X_1 + X_2 + \cdots + X_n$

$P(X \geq a) \leq\ ?$

1.4 Chernoff Idea

For any $t > 0$:

$P(X \geq a) = P(tX \geq ta) = P(e^{tX} \geq e^{ta})$

$P(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}} \quad \rightarrow$ Markov Inequality

$P(X > \alpha) \leq \frac{E[X]}{\alpha} \quad$ for non-negative α

$P(X \geq a) \leq \frac{e^{\frac{t^2 \cdot n}{2}}}{e^{ta}}$

Pick $t: \quad t = \frac{a}{n}$

$P(X \geq a) \leq e^{-\frac{a^2}{2n}}$

1.5 Textbook Version

Theorem: $X = X_1 + \cdots + X_n$, X_i i.i.d.

$\mathbb{E}[X_i] = 0, \quad \mathbb{E}[X_i^2] = \sigma^2$

$|\mathbb{E}[X_i^k]| \leq \sigma^2 \cdot k! \quad$ for $k = 3, \ldots, \lfloor \frac{a^2}{4n\sigma^2} \rfloor$

Where $0 \leq a \leq \sigma^2 \cdot n \cdot \sqrt{2}$, we have

$\mathbb{P}(|X| > a) \leq 3 \cdot e^{-\frac{a^2}{12n\sigma^2}}$

For $a = t \cdot \sigma^2 n$

$\mathbb{P}(|X| > t \cdot \sigma^2 n) \leq 3 \cdot e^{-\frac{t^2 \cdot n \cdot \sigma^2}{12}}$

1.6 Morris

Goal: An algorithm that counts a stream with a lot less than $O(\log n)$ space.

- If this is $\tilde{n}$ we want:

  $P(|\tilde{n} - n| \geq \varepsilon \cdot n) \leq \delta$

  for small δ, ε.

The algorithm of Morris provides such an estimator for some ε, δ that we will analyze shortly. The algorithm works as follows:

1. Initialize $X \leftarrow 0$.
2. For each update, increment X with probability $\frac{1}{2^X}$.
3. For a query, output $\tilde{n} = 2^X - 1$.

Let X_n denote X in Morris' algorithm after n updates.

Claim 2.1.1. $E[2^{X_n}] = n + 1$.

Proof. We prove by induction on n. The base case is clear, so we now show the inductive step. We have:

$$\begin{aligned} E[2^{X_{n+1}}] &= \textstyle\sum_{j=0}^{\infty} P(X_n = j) \cdot E(2^{X_{n+1}} \mid X_n = j) \\ &= \textstyle\sum_{j=0}^{\infty} P(X_n = j) \cdot \left(2^j \left(1 - \frac{1}{2^j}\right) + \frac{1}{2^j} \cdot 2^{j+1}\right) \\ &= \textstyle\sum_{j=0}^{\infty} P(X_n = j) \cdot 2^j + \sum_j P(X_n = j) \\ &= E[2^{X_n}] + 1 \\ &= (n+1) + 1 \end{aligned} \tag{2.2}$$

$\mathbb{P}(|\tilde{n} - n| > \epsilon n) < \frac{1}{\epsilon^2 n^2} \cdot \mathbb{E}[(\tilde{n} - n)^2] = \frac{1}{\epsilon^2 n^2} \cdot \mathbb{E}[(2^{2X} - 1 - n)^2]$

$\hookrightarrow$ Chebyshev inequality: $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$

Claim 2.1.2.

$\mathbb{E}[2^{2X_n}] = \frac{3}{2}n^2 + \frac{3}{2}n + 1$

This implies

$\mathbb{E}[(\tilde{n} - n)^2] = \frac{1}{2}n^2 - \frac{1}{2}n - 1 < \frac{1}{2}n^2$

and thus

$\mathbb{P}(|\tilde{n} - n| > \epsilon n) < \frac{1}{\epsilon^2 n^2} \cdot \frac{n^2}{2} = \frac{1}{2\epsilon^2}$

Morris +

Do Morris s times and output

$\tilde{n} = \frac{1}{s} \sum_{i=1}^{s} n_i$

$\mathbb{P}(|\tilde{n} - n| > \epsilon n) < \frac{1}{2s\epsilon^2} < \delta$

for $s > \frac{1}{2\epsilon^2\delta} = \Theta\left(\frac{1}{\epsilon^2\delta}\right)$

Too much dependency on δ and ϵ.

Ideally $\ln(1/\epsilon^2)$ and $\ln(1/\delta)$.

Morris ++

Pick s so that

$\mathbb{P}(|\hat{n} - n| > \varepsilon n) \leq \frac{1}{2s\varepsilon^2} < \frac{1}{3}$

$s \sim \frac{3}{2\varepsilon^2}$

Repeat this t times and take the median

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ trial was a success} \\ 0 & \text{otherwise} \end{cases}$$

$X = X_1 + X_2 + \cdots + X_t$

$\mathbb{P}(X_i > 0) < \frac{1}{3}, \quad \mathbb{E}[X_i] > \frac{2}{3}$

$\mathbb{E}[X] > \frac{2t}{3}$

$\mathbb{P}\left(X \leq \frac{t}{2}\right) \leq \mathbb{P}\left(|X - \mathbb{E}[X]| > \frac{t}{6}\right)$

Textbook tail bound:

$\leq e^{-t/18} < \delta$

$t \sim \ln\left(18 \cdot \frac{1}{\delta}\right)$ suffices.

In summary,

One Morris+ is $\frac{3}{2\epsilon^2}$ Morris trials.

Then we do $O\left(\ln\left(\frac{1}{\delta}\right)\right)$ repeats

to get Morris++.

In total, $O\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right)$ trials.

Space Complexity

What is the probability of storing something bigger than $\ln\left(\frac{\epsilon t n}{\delta}\right)$?

Less than δ!

So with probability $1 - \delta$, $O\left(\ln\left(\ln\left(\frac{3tkn}{\delta}\right)\right)\right)$ span

$t \sim O\left(\frac{1}{q^n}\right), \quad s \sim O\left(\ln\left(\frac{1}{\delta}\right)\right)$

$O\left(\ln\left(\ln\left(n, \frac{\ln(1/\delta)}{\delta}\right), \frac{1}{q^n}\right)\right))$

For say $\delta = 0.01, \quad \rho = 0.01$

$O(\ln(\ln n))$

2 Streaming

$a_1, a_2, \dots, a_n$ from $\{1, 2, \dots, m\}$

2.1 Frequency

$s \in \{1, 2, \dots, m\}$

$f(s) :=$ # of s in $a_1, a_2, \dots, a_n$

$\sum_{i=1}^m f(s) = n \qquad \mathbb{E}[f(s)] = \frac{1}{m} \sum_{i=1}^m f(s) = \frac{n}{m}$

$\sum_{i=1}^m f(s)^0 =$ # of distinct elements

($0^0 = 0$ by convention)

$\sum_{i=1}^m f(s)^2 \quad ? \qquad s \in \{1, 2, \dots, m\}$

$Var(f(s)) = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m f(s)^2 - (\mathbb{E}[f(s)])^2\right]$

$Var(f(s)) = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m f(s)^2 - \frac{n^2}{m^2}\right]$

2.2 Memory Limits and Subset Detection

Number of distinct elements.

$a_1, a_2, \dots, a_n$

$n > m$, and we use m bits for memory.

m bits $\rightarrow 2^m - 1$ many different numbers.

Goal: Detect the subset of $\{1, 2, \dots, n\}$ that corresponds to distinct elements.

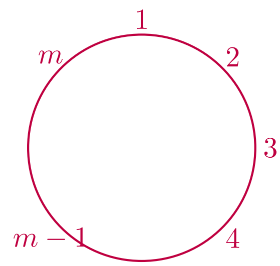
Size of the goal: Search through $2^n - 1$ many subsets.

Fact: With $2^m - 1$ numbers, some different subsets S_1, S_2 will map to same memory output since $2^m - 1 < 2^n - 1$.

2.3 Randomized Algorithms and Expectation

$n = |S|$ we choose $a_1, a_2, \dots, a_n$ randomly in $\{1, 2, \dots, m\}$

Smallest element?



$k := \#$ of distinct elements in S.

k balls into m bins, what is the smallest one in expectation?

$\frac{m}{k+1} \sim min \qquad \frac{m}{min} - 1 \sim k$

2.4 Hashing and Pairwise Independence

$h : \{1, 2, \ldots, m\} \to \{0, 1, 2, \ldots, M-1\}$

A set of hash maps

$H := \{h \mid h : \{1, 2, \ldots, m\} \to \{0, 1, \ldots, M-1\}\}$

is pairwise independent if a random element $h \in H$ satisfies

$\boxed{x \neq y, \quad x, y \in \{1, 2, \ldots, m\}}$

- then $h(x)$, $h(y)$ independent uniform distributions on $\{1, 2, \ldots, M\}$

A family of hash?

- M prime, $M > m$
- for $a, b \in \{0, 1, 2, \ldots, M-1\}$
- $h_{ab} := ax + b \mod M$
- $H := \{h_{ab} : a, b \in \{0, 1, 2, \ldots, M-1\}\}$
- $h(x) = u \quad h(y) = v$
- $\begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} \mod M$
- $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} \mod M$

$\mathbb{P}(h(x) = u, h(y) = v) = \mathbb{P}\left(\binom{a}{b} = \alpha\right) = \frac{1}{M^2}$

2.5 Counting Distinct Elements with Hashing

$h : \{a_1, a_2, \ldots, a_n\} \to \{0, 1, 2, \ldots, M-1\}$

$S :=$ image of stream after hashing
$k :=$ #distinct elements in S
$a_{\min} :=$ smallest element of S

$\frac{M}{k+1} \sim a_{\min} \qquad \frac{M}{a_{\min}} \sim k+1$

Output $\frac{M}{a_{\min}}$

Theorem: Let $k = \#$ distinct elements in S, then with probability at least $\frac{2}{3} - \frac{k}{M}$,
$\frac{M}{6k} \leq \min \leq \frac{6M}{k} \qquad \frac{k}{6} \leq \frac{M}{\min} \leq 6k$

Proof:

$P(\min \leq X) = P(\exists k : h(a_k) \leq X)$
$\leq \sum_{i=1}^{k} P(h(a_i) \leq X) = k \cdot \frac{\lceil X \rceil}{M}$

$P(\min \geq X) = P(\nexists k : h(a_k) \leq X)$

$$y_i = \begin{cases} 1 & \text{if } h(a_i) < X \\ 0 & \text{if } h(a_i) \geq X \end{cases}$$

$y = y_1 + y_2 + \cdots + y_k$

$P(y = 0) =?$

$\mathbb{E}[Y] = k \cdot \mathbb{E}[Y_1] = k \cdot \mathbb{P}(h(a_i) \leq X)$
$\mathrm{Var}(Y) = \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_k) = k \cdot \mathrm{Var}(Y_1)$
$\mathrm{Var}(Y) = k \cdot \left(\mathbb{E}[Y_1^2] - (\mathbb{E}[Y_1])^2\right)$

$\mathbb{P}(Y \geq 0) \leq \mathbb{P}(|Y - \mathbb{E}[Y]| \geq \mathbb{E}[Y])$
$\mathbb{P}(Y \geq 0) \leq \frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2} \leq \frac{1}{|\mathbb{E}[Y]|}$

for $X = \frac{M}{6k}$, $\mathbb{P}(Y \geq 0) \leq \frac{1}{6} + \frac{k}{M}$
for $X = \frac{6M}{k}$, $\mathbb{P}(Y \geq 0) \leq \frac{1}{6}$

In total, $\mathbb{P}(Y \geq 0) \leq \frac{1}{3} + \frac{k}{M}$

3 Bloom Filters

3.1 Bit Array and Space Usage

m elements, k bits per element

$m \cdot k$ bits

(1) $A[1] \quad A[2] \quad \dots \quad A[n]$

0 or 1 in every location

3.2 Hashing Mechanism

(2) k hash functions

$h_1, h_2, \dots, h_k$

$x \in S \to A[h_1(x)], \quad A[h_2(x)], \quad \dots, \quad A[h_k(x)]$

3.3 Query and Search

Total bits:

Search Time?

$y =$ query $\to$ Check $A[h_1(y)], \dots, A[h_k(y)]$

If all 1, then $y \in S$ (possible false positive).

If any of $A[h_1(y)], \dots, A[h_k(y)]$ is 0, then $y \notin S$.

Example: $y =$ password $\quad \Rightarrow A[h_1(y)], \quad A[h_2(y)], \quad \dots, \quad A[h_k(y)]$

$y \in S$

3.4 False Positive Probability

Balls and Bins: $S = \{s_1, \dots, s_n\}$

Every $s_i \to A[h_1(s_i)], \dots, A[h_k(s_i)]$

$h :$ input $\to \{0, 1, 2, \dots, n-1\}$ $\quad$ k balls per item

n bins: $A[0], A[1], \dots, A[n-1]$

m elements, total of mk balls

If m and n are fixed:

False positive: $(1-p)^k = 1 - e^{-mk/n} = g(k)$

$\min_{k \in \mathbb{Z}} g(k) \qquad \min_{x \in \mathbb{R}} g(x) \qquad g'(x) = 0$

$\xrightarrow{\text{optimal}} k = \ln 2 \cdot \frac{n}{m}$

$(1-p)^k \sim 2^{-k} = 2^{-\ln 2 \cdot \frac{n}{m}} \sim (0.618)^{\frac{n}{m}}$

False Positive $\sim 2^{-k}$

$k = \ln 2 \cdot \frac{n}{m}$ bits

If $n \sim m$, then $k \sim 4$ or 5 bits

3.5 Trade-off

Space vs. False Positive Rate

Number of bits per item.