L14: Streaming : Frequent Items and Quantiles

Jeff M. Phillips

February 25, 2025

Data is Big? Too big to fit on computer or in memory

1. Parallelism More computers

MapReduce

2. Sampling very large X data

$S \sim X$ $S = \{s_1, s_2, \ldots s_n\}$

$n << |X|$

3. Streaming $X = \langle x_1, x_2, x_3, \ldots x_i, \ldots x_n \rangle$

Read each x_i once, but not store

Maintain small space summary.

Data $\quad A = \langle a_1, a_2, a_3, \ldots a_i], \ldots a_n \rangle$

ex. $a_i \in \mathbb{R}$ $\qquad A_i = \langle a_1, a_2, \ldots a_i \rangle$

$a_1 \; a_2 \; \ldots \; a_i \quad a_{i+1}$

store $\boxed{\begin{matrix} V_i \\ i \end{matrix}} \Rightarrow V_{i+1}$

$$\frac{(V_i \cdot i) + a_{i+1}}{i+1}$$

mean (A_i)

$$\frac{1}{i} \sum_{j=1}^{i} a_i = V_i$$

sum $\boxed{\begin{matrix} S_i = \sum_{j=1}^{i} a_i \\ i = \# \text{ steps} \end{matrix}}$ $\oplus a_{i+1} \rightarrow$ $\boxed{\begin{matrix} S_{i+1} = S_i + a_{i+1} \\ i+1 = i+1 \end{matrix}}$

$$V_i = \frac{S_i}{i}$$

$$\frac{1}{i} \sum_{j=1}^{i} (a_j - E[A_i])^2 = \mathrm{Var}[A_i]$$

$$V_{i+1} = \frac{S_{i+1}}{i+1}$$

Variance $[A_i]$

$$\frac{1}{i}\sum_{j=1}^{i}\left(a_i - E[A_i]\right)^2 = \frac{1}{i}\sum_{j=1}^{i} a_i^2 - \left(\frac{1}{i}\sum_{j=1}^{i} a_i\right)^2$$

$$= E[a_i^2] - \left(E[a_i]\right)^2$$

store

$$\boxed{\begin{array}{c} i \\ S_i \\ Q_i \end{array}}$$

$$S_i = \sum_{j=1}^{i} a_j$$

$$Q_i = \sum_{j=1}^{i} (a_j)^2$$

$$\mathrm{Var}[A_i] = \frac{Q_i}{i} - \left(\frac{S_i}{i}\right)^2$$

Resevoir Sampling

$\langle a_1, a_2, \dots a_n \rangle$ stream

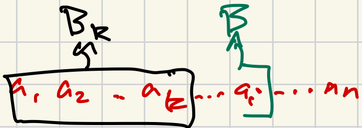
Goal: Maintain Random Sample $B \subset A$
without replacement $|B| = k$

1. Keep first k item $B = \{a_1, a_2, \dots a_k\}$
2. for $j = k+1 \dots$ to n

   keep a_j in B w/p $\frac{k}{j}$

   ↳ boot a random $b \in B$ from B.

   o.w. keep B the same



Stream $A = \langle a_1, a_2, \ldots a_i \ldots a_n \rangle$

$a_i \in [m]$

$[m] = $ IP addresses

$[m] = $ k-grams

both m and n too big.

label $j \in [m]$ $\log m$ bits

counter $i \in [n]$ $\log n$ bits

approx freq

$\hat{f}_j$

$|f_j - \hat{f}_j| \leq \varepsilon n$

frequency $f_j = |\{a_i \in A \mid a_i = j\}|$
$j \in [m]$

Heavy-Hitter j s.t. $f_j > \phi n$

$\phi = 0.1$
$\varepsilon = 0.01$

ok $f_j > \phi n - \varepsilon n$

MAJORITY

$A = (a_1 \ldots a_n)$

$a_i \in [m]$

if (some $f_j > \frac{m}{2}$) $\rightarrow$ output j

else output anything $j \in [m]$

[counter $\in [n]$ | label $j \in [m]$

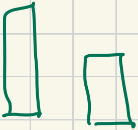
c $\qquad$ l

if ($a_i = l$) $\quad c = c + 1$

else $\quad c = c - 1$

if ($c < 0$)

$l = a_i$; $c = 1$



l_1 $\quad$ l_2

Majority

Majority(A)

Set $c = 0$ and $\ell = \emptyset$
for $i = 1$ **to** $\cancel{m}\ n$ **do**
  if $(a_i = \ell)$ **then**
    $c = c + 1$
  else
    $c = c - 1$
  if $(c < 0)$ **then**
    $c = 1,\ \ell = a_i$
return ℓ

Frequency Approximation — Misra-Gries

$k-1$ counters, $k-1$ labels

$\forall_{j \in [n]}$ $\hat{f}_j$ s.t. $f_j - \frac{n}{k} \le \boxed{\hat{f}_j} \le f_j$ apx freq. $k = \frac{1}{\varepsilon}$

$\boxed{k = \frac{1}{\varepsilon} \quad \frac{n}{k} = \varepsilon n}$

Labels $l_1, l_2, \ldots l_{k-1} = L$

counters $c_1, c_2, \ldots c_{k-1}$

for $(a_i \in A)$
  if $(a_i \in L)$ $a_i = l_j$
    $c_j = c_j + 1$
  else $(a_i \notin L)$
    if (some $c_j \le 0$) $\to$ else { decrease all counter by 1 $\forall_j$ $c_j = c_j - 1$
      $l_j = a_i$
      $c_j = 1$

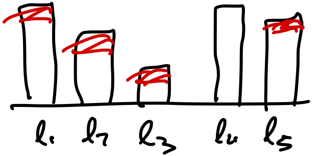
no labels unused.

Misra-Gries

counter array C : $C[1], C[2], \ldots, C[k-1]$
location array L : $L[1], L[2], \ldots, L[k-1]$

<u>Misra-Gries(A)</u>

  Set all $C[i] = 0$ and all $L[i] = \emptyset$
  for $i = 1$ **to** $\cancel{m}\ n$ **do**
    if $(a_i = L[j])$ **then**
      $C[j] = C[j] + 1$
    else
      if (some $C[j] = 0$) **then**
        Set $L[j] = a_i$ & $C[j] = 1$
      else
        for $j \in [k-1]$ **do** $C[j] = C[j] - 1$
  return C, L



how many total decrements?

$\hookrightarrow \dfrac{n}{k}$

Frugal Median

Frugal Median(A)

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  if $(a_i > \ell)$ **then**
    $\ell \leftarrow \ell + 1$.
  if $(a_i < \ell)$ **then**
    $\ell \leftarrow \ell - 1$.
return ℓ.

Frugal Quantile

Frugal Quantile(A, ϕ) e.g. $\phi = 0.75$

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  $r = \mathrm{Unif}(0, 1)$ (at random)
  if ($a_i > \ell$ **and** $r > 1 - \phi$) **then**
    $\ell \leftarrow \ell + 1$.
  if ($a_i < \ell$ **and** $r > \phi$) **then**
    $\ell \leftarrow \ell - 1$.
return ℓ.

Frequent Itemsets : Apriori

$$
\begin{aligned}
T_1 &= \{1, 2, 3, 4, 5\} \\
T_2 &= \{2, 6, 7, 9\} \\
T_3 &= \{1, 3, 5, 6\} \\
T_4 &= \{2, 6, 9\} \\
T_5 &= \{7, 8\} \\
T_6 &= \{1, 2, 6\} \\
T_7 &= \{0, 3, 5, 6\} \\
T_8 &= \{0, 2, 4\} \\
T_9 &= \{2, 4\} \\
T_{10} &= \{6, 7, 9\} \\
T_{11} &= \{3, 6, 9\} \\
T_{12} &= \{6, 7, 8\}
\end{aligned}
$$