

L12: Streaming & Sampling

Sep 29, 2025

Data Mining

Jeff M. Phillips



Streaming

Input Data $A = \langle a_1, a_2, \dots, a_i, \dots, a_m \rangle$

each $a_i \in M = \{\mathbb{R}, [0]\}$

- items seen in order a_i before a_{i+1}
only get one look
- Memory/Space is very limited.

→ Goal Summary of $A \Rightarrow S$
Summary

size to store
 S is small.

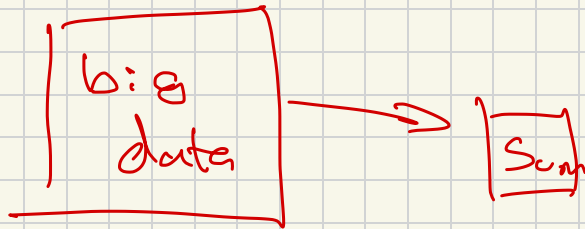
small $|S| = \text{polynomial}(\log(m), \log(v))$

counter

label

Where are streaming algos used?

- router
- busy website
- review stores of big data
- time series, online



Warm-Up

$$A = \underbrace{\langle a_1, a_2, \dots, a_i, \dots \rangle}_{A_i}$$

$a_i \in [1, 2, \dots, v]$

Goal: average (A)

$$\begin{aligned} - \text{count } C_i &= |\{a_1, a_2, \dots, a_i\}| = i \\ - \text{sum } S_i &= \sum_{j=1}^i a_j \end{aligned}$$

update

$$C_i = C_{i-1} + 1$$

$$S_i = S_{i-1} + a_i$$

$$\text{average}(A_i) = \frac{S_i}{C_i} = \frac{1}{i} \sum_{j=1}^i a_j$$

Sampling: Maintain random sample S of A
of size k .

- with replacement random sample

- can draw same item twice

$a_i, a_i \rightarrow S$

- iid $S \stackrel{iid}{\sim} A$

= without replacement random sample

- cannot draw same item twice

- not iid

- usually better estimate.

if $|S| \ll |A|$
very different
1.111

Reservoir Sampling $k=1$

$$S = \{s\}$$

reservoir

RS(A)

1. $s \leftarrow a_1$

for $i = 2$ to n

Generate $u \sim \text{Unif}[0, 1]$

if $(u < 1/i)$

$s \leftarrow a_i$

return s

$A = \langle \underline{a_1}, a_2, \dots, \textcircled{a_i}, \dots, a_n \rangle$

at each step i

each a_j ($j \leq i$)

equally likely $= 1/i$

Inductively say at $i-1$

each a_j $j = \{1, \dots, i-1\}$

s w.p. $\frac{1}{i-1}$

Still there w.p. $(1 - \frac{1}{i}) = \frac{i-1}{i}$

$$\frac{1}{i-1} \cdot \frac{i-1}{i} = \frac{1}{i}$$

With Replacement Reservoir Sampling $k > 1$

Run k independent copies of RS(A) Algo.

Without Replacement Reservoir Sampling

RS(A, k)

1. $S \leftarrow \{a_1, a_2, \dots, a_k\}$
2. for $i = k+1$ to n
 $v \leftarrow \text{Unif}(0, 1)$
 if $(v \leq k/i)$
 remove random $s' \in S$
 put a_i into S .

reservoir

$$S = \{s_1, s_2, \dots, s_k\}$$

Guarantee S random w/o replac
RS of A .

$$\Pr[a_i \in S] = \frac{k}{i} \quad \forall j \leq i$$

Weighted Sampling

Stream $\langle a_1, a_2, \dots, a_m \rangle$

w/ weights $w_1, w_2, \dots, w_m$

Select

a_j

w.p.

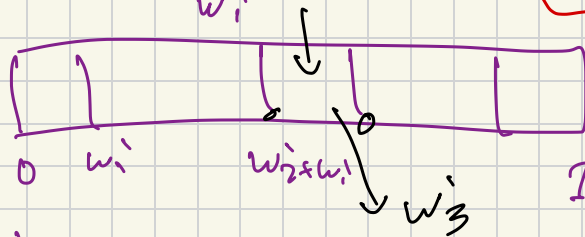
$$\frac{w_j}{W_j}$$

$$w_j = \frac{w_j}{W_j}$$

$$W_j = \sum_{i=1}^j w_i$$

↑ maintain
in stream

$k=1$



always know w_i for processing a_i in stream

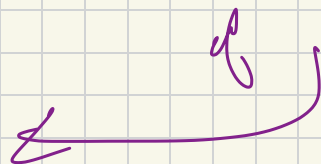
keep a_i w.p. $\frac{w_i}{W_j}$

Weighted Reservoir Sampling $b=1$

0. $W = 0$
1. $s \leftarrow a_i$ $W = w_i$
2. for $j = 2$ to m
 $W = W + w_j$
 $u \sim \text{Unif}(0, 1)$
 if $\left(u \leq \frac{w_j}{W} \right)$
 $s \leftarrow a_{j_i}$
3. Return s .

Weighted with
Replacement,
size k .

Run k instances



Weighted Withhood Replacement $k \geq 1$

$$S = \{a_1, \dots, a_k\}$$

$p_1, \dots, p_k$ priorities

$$p_j = \frac{1}{w_j} \ln(v_j) \quad v_j \sim \text{Unif}(0,1)$$

In stream

S set of k a_i with smallest p_i