Universal sketches for the frequency negative moments

Vladimir Braverman
Johns Hopkins University

Stephen Chestnut
ETH Zürich

August 25, 2015

A stream of $m = 7$ items from $[n] = [4]$

$$4,\ \ 2,\ \ 2,\ \ 3,\ \ 4,\ \ 2,\ \ \ 2$$

$$f = \left[\ \ \right]$$

$$\sum \tfrac{1}{f_i} =$$

A stream of $m = 7$ items from $[n] = [4]$

$$4,\ \ 2,\ \ 2,\ \ 3,\ \ 4,\ \ 2,\ \ 2$$

$$f = \begin{bmatrix} \\ \\ \\ 1 \end{bmatrix}$$

$$\sum \frac{1}{f_i} = \quad 1$$

A stream of $m = 7$ items from $[n] = [4]$

$$2,\ \ 2,\ \ 3,\ \ 4,\ \ 2,\ \ \ 2$$

$$f = \begin{bmatrix} \\ 1 \\ \\ \\ 1 \end{bmatrix}$$

$$\sum \tfrac{1}{f_i} = \qquad 2$$

A stream of $m = 7$ items from $[n] = [4]$

$$2,\ 3,\ 4,\ 2,\ 2$$

$$f = \begin{bmatrix} 2 \\ \\ 1 \end{bmatrix}$$

$$\sum \frac{1}{f_i} = \frac{3}{2}$$

A stream of $m = 7$ items from $[n] = [4]$

$$3,\ \ 4,\ \ 2,\ \ 2$$

$$f = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

$$\sum \frac{1}{f_i} = \frac{5}{2}$$

A stream of $m = 7$ items from $[n] = [4]$

$$4, \quad 2, \quad 2$$

$$f = \begin{bmatrix} \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

$$\sum \frac{1}{f_i} = 2$$

A stream of $m = 7$ items from $[n] = [4]$

$$2, \quad 2$$

$$f = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

$$\sum \frac{1}{f_i} = \frac{11}{6}$$

A stream of $m = 7$ items from $[n] = [4]$

$$\begin{matrix} & 2 \\ f = & \begin{bmatrix} \\ 4 \\ 1 \\ 2 \end{bmatrix} \\ \sum \frac{1}{f_i} = & \frac{7}{4} \end{matrix}$$

The pth moment, $p > 0$, is

$$F_p = \sum_i f_i^p.$$

How much storage is necessary to get a $(1 \pm \epsilon)$-approximation to F_p?

The pth moment, $p > 0$, is

$$F_p = \sum_i f_i^p.$$

How much storage is necessary to get a $(1 \pm \epsilon)$-approximation to F_p?

$$O(\tfrac{\log m}{\epsilon^2}), \quad 0 < p \leq 2, \quad \text{[AMS96,I06,KNW10]}$$

$$O(n^{1-\frac{2}{p}} \tfrac{\log m}{\epsilon^2}), \quad 2 < p, \quad \text{[AMS96,IW05,G11,BKSV14]}$$

Lower bounds: [AMS96,CKS03,LW13]

The pth negative moment, $p > 0$, is

$$F_{-p} = \sum_{i \in \mathsf{supp}(f)} f_i^{-p}.$$

How much storage is necessary to get a $(1 \pm \epsilon)$-approximation to F_{-p}?

The pth negative moment, $p > 0$, is

$$F_{-p} = \sum_{i \in \mathsf{supp}(f)} f_i^{-p}.$$

How much storage is necessary to get a $(1 \pm \epsilon)$-approximation to F_{-p}?

Let $g(x) \geq g(y) \geq 0$, for all $0 < x < y$.
How much storage is necessary to approximate $g(f) := \sum g(f_i)$?

The pth negative moment, $p > 0$, is

$$F_{-p} = \sum_{i \in \mathrm{supp}(f)} f_i^{-p}.$$

How much storage is necessary to get a $(1 \pm \epsilon)$-approximation to F_{-p}?

Let $g(x) \geq g(y) \geq 0$, for all $0 < x < y$.
How much storage is necessary to approximate $g(f) := \sum g(f_i)$?

$$\begin{aligned} \sigma = & \max |\mathrm{supp}(f)| \\ & \epsilon g(f) \leq g(1), \\ & \textstyle\sum f_i \leq m, \\ & f \in \mathbb{Z}^n_{\geq 0} \end{aligned}$$

$$\sigma = \max |\mathrm{supp}(f)|$$
$$\begin{aligned} \epsilon g(f) &\leq g(1), \\ \sum f_i &\leq m, \\ f &\in \mathbb{Z}^n_{\geq 0} \end{aligned}$$

- For F_{-p} we get
$$\sigma \approx \min\{\left(\frac{m^p}{\epsilon}\right)^{\frac{1}{1+p}}, n\}$$
- When $p = 1$ and $m = O(n)$ this is
$$\sigma = O(\sqrt{n/\epsilon})$$
- When $p = 1$ and $m \geq n^2$ it is $\sigma = n$.

Outline

- $O(\frac{1}{\epsilon}\sigma \log m + \log^2 n)$ space $(1 \pm \epsilon)$-approximation algorithm
- $\Omega(\sigma)$ lower bound
- Evaluating σ

$(1 \pm \epsilon)$-**approximation for** $g(f)$

1. Compute $\sigma = \sigma(\epsilon, g, m, n)$ and let

$$q \geq \min\{\frac{9\sigma}{\epsilon|\operatorname{supp}(f)|}, 1\}$$

2. Sample $W \subseteq \operatorname{supp}(f)$ with $P(i \in W) = q$, pair-wise indep.
3. Compute f_i, for each $i \in W$
4. Output $\hat{G} = q^{-1} \sum_{i \in W} g(f_i)$

Lemma

For all $f \in \mathbb{Z}^n_{\geq 0}$ *with* $\sum_i f_i \leq m$

$$\frac{\sigma}{\epsilon |\operatorname{supp}(f)|} \geq \min\{1, \max_i \frac{g(f_i)}{\epsilon^2 g(f)}\}.$$

Since the samples are pairwise independent, with probability at least $8/9$

$$\hat{G} = (1 \pm \epsilon) g(f).$$

Storage requirements

- $O(\frac{1}{\epsilon}\sigma \log m)$ bits for counters
- $O(\log^2 n)$ random bits

1. Compute $\sigma = \sigma(\epsilon, g, m, n)$ and let

$$q \geq \min\{\frac{9\sigma}{\epsilon|\operatorname{supp}(f)|}, 1\}$$

2. Sample $W \subseteq \operatorname{supp}(f)$ with $P(i \in W) = q$, pair-wise indep.
3. Compute f_i, for each $i \in W$
4. Output $\hat{G} = q^{-1} \sum_{i \in W} g(f_i)$

Universality: *The same parameter works for every decreasing function of the same (or smaller) space complexity σ*

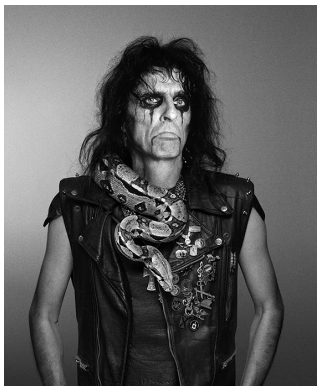
Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Alice: $A = \{1, 4, 7, 8\}$

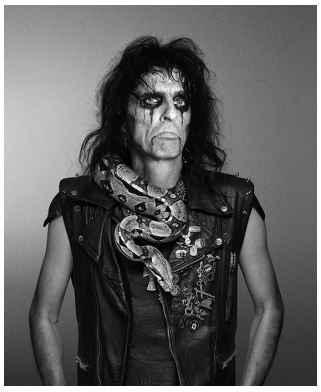


Bob: $b = 2$

Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Alice: $A = \{1, 4, 7, 8\}$



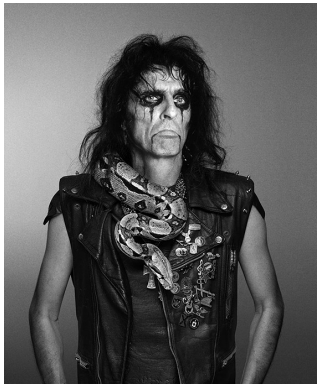
$1^{(f_1)}, 4^{(f_4)}, 7^{(f_7)}, 8^{(f_8)}$

Bob: $b = 2$

Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

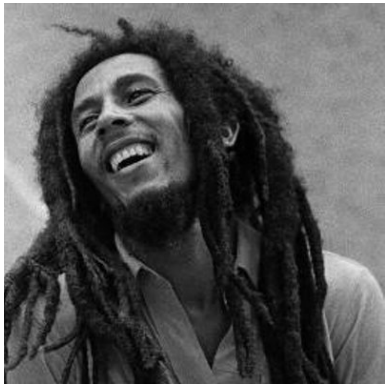
Alice: $A = \{1, 4, 7, 8\}$



memory

$1^{(f_1)}, 4^{(f_4)}, 7^{(f_7)}, 8^{(f_8)}$

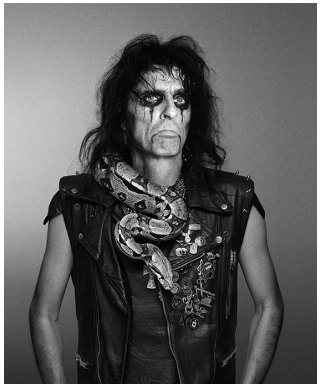
Bob: $b = 2$



Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Alice: $A = \{1, 4, 7, 8\}$

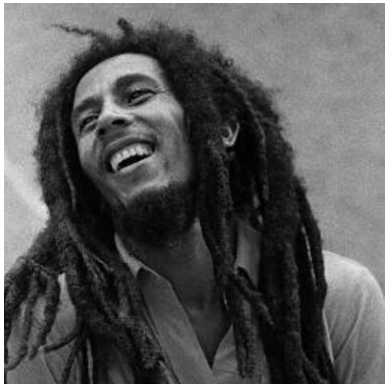


memory

$1^{(f_1)}, 4^{(f_4)}, 7^{(f_7)}, 8^{(f_8)}$

2

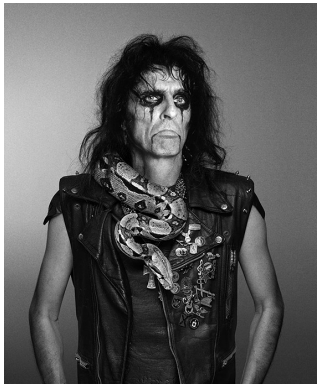
Bob: $b = 2$



Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Alice: $A = \{1, 4, 7, 8\}$

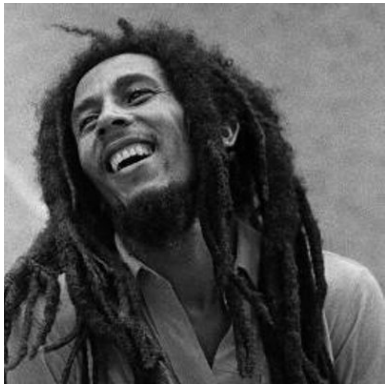


memory

$1^{(f_1)}, 4^{(f_4)}, 7^{(f_7)}, 8^{(f_8)}$

2

Bob: $b = 2$

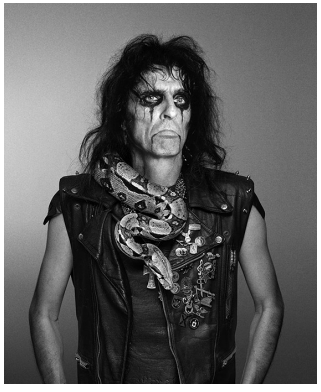


$\hat{G} > 2g(f)$
therefore
"$b \notin A$"

Let $f \in \mathbb{Z}^n_{\geq 0}$ satisfy $\sum f_i \leq m$ and $g(1) \geq 2g(f)$

Reduction from index for $\Omega(|\operatorname{supp}(f)|)$ storage lower bound

Alice: $A = \{1, 4, 7, 8\}$

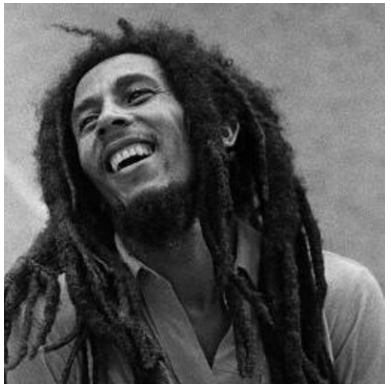


memory

$1^{(f_1)}, 4^{(f_4)}, 7^{(f_7)}, 8^{(f_8)}$

2

Bob: $b = 2$



$\hat{G} > 2g(f)$
therefore
"$b \notin A$"

Best possible lower bound here

$$\sigma = \max\{|\operatorname{supp}(f)| : g(1) \geq \epsilon g(f), \sum f_i \leq m, f \in \mathbb{Z}^n_{\geq 0}\}$$

$$\sigma = \max |\mathrm{supp}(f)|$$
$$\begin{aligned} \epsilon g(f) &\leq g(1), \\ \sum f_i &\leq m, \\ f &\in \mathbb{Z}^n_{\geq 0} \end{aligned}$$

4-approximation for σ

- Since g is decreasing, enforce that $\sum f_i = m$
- Take all frequencies f_i the same or 0
- Assume the (single) frequency is a power of two
- Can check each of $O(\log m)$ possibilities in with $O(\log n)$ space

Single stream f	$\Rightarrow$	$\lvert\mathrm{supp}(f)\rvert$ lower bound
Optimize over f	$\Rightarrow$	maximal lower bound
Optimality	$\Rightarrow$	algorithm correctness

I didn't tell you about

- Sampling $\mathrm{supp}(f)$ without additional space
- Handling deletions in $O(\frac{1}{\epsilon}\sigma \log m \log n)$ bits
- Multiple passes