Simplest Data Stream Model

Definition
Input is a length $m$ list $\langle a_1, \ldots, a_m \rangle$ where $a_i \in [n]$. You can only access the elements sequentially and have memory that is sub-linear in $m$ and $n$.

Problem (Frequency-based Problems)

Let $f = (f_1, f_2, \ldots, f_n)$ encode element frequencies, i.e., $f_i = |\{j : a_j = i\}|$
1. Find all $i$ such that $f_i \geq \epsilon m$, i.e., the heavy hitters.
2. $1 + \epsilon$ approximate $F_0 = \sum_{i \in [n]} f_i^0$, i.e., the number of distinct items.
3. $1 + \epsilon$ approximate $F_2 = \sum_{i \in [n]} f_i^2$, i.e., the self-join size.

Can solve all these problems using $\text{poly}(1/\epsilon, \log n, \log m)$ space even in the model where the frequencies are incremented and decremented.
Linear Sketches

- **Recap:** A linear sketch algorithm stores a random matrix $Z \in \mathbb{R}^{k \times n}$ where $k \ll n$ and computes projection $Zf$ of the input $f$. $Z$ is chosen independently of $f$ but the distribution is chosen such that w.h.p. we can determine the relevant properties of any $f$.

- **Application to Streams:** Sketches can be computed incrementally!
  - Suppose we have sketch $Zf$ of current frequency vector $f$.
  - If we see an occurrence of $i$, the new frequency vector is $f' = f + e_i$.
  - Can update sketch be just adding $i$ column of $Z$ to $Zf$:
    \[
    Zf' = Z(f + e_i) = Zf + Ze_i = Zf + (i\text{-th column of } Z)
    \]
  - If the $i$th entry of $f$ was decremented then
    \[
    Zf' = Z(f - e_i) = Zf - Ze_i = Zf - (i\text{-th column of } Z)
    \]
Example: Dynamic Connectivity in Data Streams

Definition (Dynamic Graph Stream)

Input is a sequence of edge insertions and deletions that define a graph $G$. You can only access the elements sequentially and have memory that is sub-linear in the number of edges.

We designed sketches $\mathcal{A}_1, \ldots, \mathcal{A}_{\log n}$ and player $i$ sent messages $\mathcal{A}_1 x_i, \ldots, \mathcal{A}_{\log n} x_i$. Each message can be computed in $O(\text{polylog } n)$ bits in the data stream model. Total of $O(n \text{polylog } n)$ bits.

When edge $\{i, j\}$ is inserted where $j > i$:

$$
\mathcal{A}_t x_i \leftarrow \mathcal{A}_t x_i + \mathcal{A}_t e_{i,j}
$$

$$
\mathcal{A}_t x_j \leftarrow \mathcal{A}_t x_j - \mathcal{A}_t e_{i,j}
$$

where $e_{i,j}$ is the length $\binom{n}{2}$ binary vector whose only non-zero entry is in the $\{i,j\}$th entry and this entry is 1. When edge $\{i, j\}$ is deleted where $j > i$:

$$
\mathcal{A}_t x_i \leftarrow \mathcal{A}_t x_i - \mathcal{A}_t e_{i,j}
$$

$$
\mathcal{A}_t x_j \leftarrow \mathcal{A}_t x_j + \mathcal{A}_t e_{i,j}
$$
Problem: Construct an $(\epsilon, \delta)$ approximation for $F_2 = \sum_i f_i^2$

Algorithm:
- Let $Z \in \{-1, 1\}^{k \times n}$ where entries of each row are $4$-wise independent and rows are independent.
- Compute $Zf$ and average squared entries appropriately.

Analysis:
- Let $s = z.f$ be an entry of $Zf$ where $z$ is a row of $Z$.
- Lemma: $\mathbb{E}[s^2] = F_2$
- Lemma: $\mathbb{V}[s^2] \leq 4F_2^2$
Expectation Lemma

- \( s = z.f \) where \( z_i \in \{ -1, 1 \} \) are 4-wise independent.
- Then

\[
E[s^2] = E \left[ \sum_{i,j \in [n]} z_iz_jf_if_j \right] = \sum_{i,j \in [n]} f_if_jE[z_iz_j] = \sum_{i \in [n]} f_i^2 = F_2
\]

since \( E[z_iz_j] = 0 \) unless \( i = j \),
Variance Lemma

\[ \mathbb{E}[z_iz_jz_kz_l] = 0 \text{ unless } (i = j \text{ and } k = l) \text{ or } (i = k \text{ and } j = l) \text{ or } (i = l \text{ and } j = k) \]

Then

\[ \nabla [s^2] = \mathbb{E}[s^4] - \mathbb{E}[s^2]^2 = \sum_{i} f_i^4 + 3 \sum_{i \neq j} f_i^2 f_j^2 - \left( \sum_{i \in [n]} f_i^2 \right)^2 \]

\[ = 2 \sum_{i \neq j} f_i^2 f_j^2 \]

\[ \leq 2F_2^2 \]
Averaging “ Appropriately”

- Group entries of the sketch into $a = O(\log \delta^{-1})$ groups of $b = 6\epsilon^{-2}$
- Let $Y_1, Y_2, \ldots, Y_a$ be the average of squared entries in each group.

$$\mathbb{E}[Y_i] = F_2$$

$$\nabla [Y_i] \leq 2F_2^2 / b$$

- By Chebychev, $\mathbb{P}[|Y_i - F_2| \geq \epsilon F_2] \leq \frac{2F_2^2}{b(\epsilon F_2)^2} = 1/3$
- By Chernoff, $\text{median}(Y_1, \ldots, Y_a)$ is a $(\epsilon, \delta)$ approximation of $F_2$. 
Extension to Estimating $\ell_p$ Norms

- The $\ell_p$ norm is defined as $\ell_p(f) = (\sum_i |f_i|^p)^{1/p}$
- A distribution $D$ is $p$-stable if given $X, Y \sim D$ and $a, b \in \mathbb{R}$ then
  \[ aX + bY \sim (a^p + b^p)^{1/p} D \]
- E.g., Cauchy and Gaussian distributions are 1 and 2-stable:
  \[
  \text{Cauchy}(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} \quad \text{Gaussian}(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}
  \]
- If entries of matrix $z_{i,j} \sim D$ are $p$ stable, then projection entries:
  \[ s \sim \ell_p(f) D \]
- For $p \in (0, 2]$, can $(\epsilon, \delta)$ estimate $\ell_p$ in $O(\epsilon^{-2} \text{polylog}(n, m))$ space.
Relationship to Johnson Lindenstrauss

Algorithm

1. Let \( Y = \frac{1}{\|X\|_2} (x_1, \ldots, x_n) \) be such that each \( x_i \) are independent \( N(0, 1) \) variables where \( \|X\|_2 = (x_1^2 + \ldots + x_n^2)^{1/2} \)
2. Compute \( h_1(f) = Y \cdot f \)
3. Repeat \( d \) times to get \( h(f) = (h_1(f), \ldots, h_d(f)) \)

Theorem

For any \( p \) vectors in Euclidean space \( v_1, \ldots, v_p \in \mathbb{R}^n \), there exists a map \( h : \mathbb{R}^n \to \mathbb{R}^d \) such that

\[
(1 - \epsilon)\|h(v_i) - h(v_j)\|_2 \leq \|v_i - v_j\|_2 \leq (1 + \epsilon)\|h(v_i) - h(v_j)\|_2
\]

where \( d = O(\epsilon^{-2} \log n) \).
Thanks!