CMPSCI 711: More Advanced Algorithms
Vectors 2: Sketching $F_0$ and $F_2$

Andrew McGregor
Hash Functions

Definition
A family $\mathcal{H}$ of functions from $A \rightarrow B$ is $k$-wise independent if for any distinct $x_1, \ldots, x_k \in A$ and $i_1, i_2, \ldots, i_k \in B$,

$$\mathbb{P}_{h \in \mathcal{H}}[h(x_1) = i_1, h(x_2) = i_2, \ldots, h(x_k) = i_k] = \frac{1}{|B|^k}$$

Example
Suppose $A \subset \{0, 1, 2, \ldots, p - 1\}$ and $B = \{0, 1, 2, \ldots, p - 1\}$. Then,

$$\mathcal{H} = \{ h(x) = \sum_{i=0}^{k-1} a_i x^i \mod p : 0 \leq a_0, a_1, \ldots, a_{k-1} \leq p - 1 \}$$

is a $k$-wise independent family of hash functions.

Note. If $|B|$ is not prime or $|A| > |B|$ more ideas are required.
Linear Sketches

- A sketch algorithm stores a random matrix \( Z \in \mathbb{R}^{k \times n} \) where \( k \ll n \) and computes projection \( Zf \) of the frequency vector.

- **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 \( Zf \):

\[
Zf' = Z(f + e_i) = Zf + Ze_i = Zf + (i\text{-th column of } Z)
\]

- **Useful?** Need to choose random matrices such that relevant properties of \( f \) can be estimated with high probability from \( Zf \).
Outline

$F_2$ Estimation

Distinct Elements
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: $\mathbb{E} \left[ s^2 \right] = F_2$
- Lemma: $\mathbb{V} \left[ s^2 \right] \leq 4F_2^2$
Expectation Lemma

- $s = z \cdot f$ where $z_i \in_R \{-1, 1\}$ are 4-wise independent.
- Then

$$
E[s^2] = E \left[ \sum_{i,j \in [n]} z_i z_j f_i f_j \right] = \sum_{i,j \in [n]} f_i f_j E[z_i z_j] = \sum_{i \in [n]} f_i^2
$$

since $E[z_i z_j] = 0$ unless $i = j$, 

Variance Lemma

- $\mathbb{E}[z_i z_j z_k z_l] = 0$ unless $(i, k) = (j, l), (i, j) = (k, l)$ or $(i, j) = (l, k)$

- Then

\[
\nabla [s^2] = \mathbb{E}[s^4] - \mathbb{E}[s^2]^2 = \sum_i f_i^4 + 6 \sum_{i<j} f_i^2 f_j^2 - \left( \sum_{i \in [n]} f_i^2 \right)^2 \\
= 4 \sum_{i<j} f_i^2 f_j^2 \\
\leq 4 F_2^2
\]

$F_2$
Averaging “ Appropriately”

- Group entries of the sketch into \( a = O(\log \delta^{-1}) \) groups of \( b = 12\varepsilon^{-2} \)
- Let \( Y_1, Y_2, \ldots, Y_a \) be the average of squared entries in each group.

\[
E[Y_i] = F_2 \\
\text{Var}[Y_i] \leq 4F_2^2/b
\]

- By Chebychev, \( \mathbb{P}[|Y_i - F_2| \geq \varepsilon F_2] \leq \frac{4F_2^2}{b(\varepsilon F_2)^2} = 1/3 \)
- By Chernoff, \( \text{median}(Y_1, \ldots, Y_a) \) is a \((\varepsilon, \delta)\) approximation of \( F_2 \).
Extension to Estimating $\ell_p$

- 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.
Outline

$F_2$ Estimation

Distinct Elements
Distinct Elements

- **Problem**: Construct an \((\epsilon, \delta)\) approximation for \(F_0 = \sum_i f_i^0\)
- **Simpler problem**: For given \(T > 0\), with probability \(1 - \delta\) distinguish between \(F_0 > (1 + \epsilon)T\) and \(F_0 < (1 - \epsilon)T\)
- If we can solve simpler problem, can solve original problem by trying \(O(\epsilon^{-1} \log n)\) values of \(T\)

\[
T = 1, (1 + \epsilon), (1 + \epsilon)^2, \ldots, n
\]

- **Algorithm**:
  - Choose random sets \(S_1, S_2, \ldots, S_k \subset [n]\) where \(\mathbb{P}[i \in S_j] = 1/T\)
  - Compute \(s_j = \sum_{i \in S_j} f_i\)
  - If at least \(k/e\) of the \(s_j\) are zero, output \(F_0 < (1 - \epsilon)T\)
- **Analysis**:
  - If \(F_0 > (1 + \epsilon)T\), \(\mathbb{P}[s_j = 0] < 1/e - \epsilon/3\)
  - If \(F_0 < (1 - \epsilon)T\), \(\mathbb{P}[s_j = 0] > 1/e + \epsilon/3\)
  - Chernoff: \(k = O(\epsilon^{-2} \log \delta^{-1})\) ensures correctness with prob. \(1 - \delta\).
Suppose $T$ is large and $\epsilon$ is small:

$$\mathbb{P}[s_j = 0] = (1 - 1/T)^{F_0} \approx e^{-F_0/T}$$

- If $F_0 > (1 + \epsilon)T$,
  $$e^{-F_0/T} \leq e^{-(1+\epsilon)} \leq e^{-1} - \epsilon/3$$

- If $F_0 < (1 - \epsilon)T$,
  $$e^{-F_0/T} \geq e^{-(1-\epsilon)} \geq e^{-1} + \epsilon/3$$