Estimating F_2 in Turnstile streams

Recall $F_2(\sigma) = \sum_i f_i^2 = \|f\|_2^2$. In general, $F_k(\sigma) = \sum_i f_i^k$.

Recall we saw "AMS Sampling", which could (ϵ,δ)-approximate F_k using space $O\left(\frac{n^{1-1/k}\cdot\log(1/\delta)\cdot\log(mn)}{\epsilon}\right)$ in insertion-only streams.

Here, we will see an algorithm that can (ϵ,δ)-approximate F_2 in space $O\left(\frac{(\log m + \log n)\log(1/\delta)}{\epsilon^2}\right)$ in turnstile streams (due to AMS99, called the <u>Tug-of-War Sketch</u>).

Basic Estimator.

- Choose a random hash function $h : [n] \to \{-1,1\}$ from a 4-wise independent hash family.
- Initialize $x \leftarrow 0$.
- When processing update (a_j, δ_j): $x \leftarrow x + \delta_j \cdot h(a_j)$
- Output x^2.

← Same as Count-sketch with a single counter

Note: h takes $O(\log n)$ bits to store, and x takes $O(\log(m))$ bits.

Let X denote the random variable given by the value of x at the end of the stream. Let $Y_i \leftarrow h(i)$.

Then $X = \sum_{i=1}^n f_i \cdot Y_i$, and the returned estimate is X^2.

Claim: $\mathbb{E}[X^2] = F_2(\sigma)$.

Proof: $\mathbb{E}[X^2] = \mathbb{E}\left[\left(\sum_i f_i Y_i\right)^2\right] = \mathbb{E}\left[\sum_i f_i^2 Y_i^2 + \sum_i \sum_{j\neq i} f_i f_j Y_i Y_j\right]$

linearity of expectation → $= \sum_i \mathbb{E}[f_i^2 \underbrace{Y_i^2}_{1}] + \sum_i \sum_{j\neq i} \mathbb{E}[\boxed{f_i f_j} Y_i Y_j] \overset{\text{Pairwise independence of } h}{=} \underbrace{\sum_i f_i^2}_{=F_2} + \sum_i \sum_{j \neq i} f_i f_j \underbrace{\mathbb{E}[Y_i]}_{0}\underbrace{\mathbb{E}[Y_j]}_{0}$

Claim: $\mathrm{Var}[X^2] \le 2F_2^2$.

Proof: $\mathrm{Var}[X^2] = \mathbb{E}[X^4] - \mathbb{E}[X^2]^2 = \mathbb{E}[X^4] - F_2^2$. (*)

$$\mathbb{E}[X^4] = \mathbb{E}\left[\left(\sum_i f_i Y_i\right)^4\right] = \mathbb{E}\left[\sum_i \sum_j \sum_k \sum_l f_i f_j f_k f_l Y_i Y_j Y_k Y_l\right]$$

linearity of expectation

$$= \sum_i \sum_j \sum_k \sum_l f_i f_j f_k f_l \, \mathbb{E}[Y_i Y_j Y_k Y_l]. \quad (**)$$

Any term in (*) with ~~one~~ of the indices (i,j,k,l) appearing exactly once in the 4-tuple is 0.

E.g. if $i \notin \{j,k,l\}$, then $\mathbb{E}[Y_i Y_j Y_k Y_l] \overset{\text{4-wise independent}}{=} \mathbb{E}[Y_i]\,\mathbb{E}[Y_j Y_k Y_l] = 0$

So ~~the~~ only non-zero terms are of the form

- one index occurring 4 times or
- two indices occurring twice each

$$\text{So } (**) = \sum_i f_i^4 \underbrace{\mathbb{E}[Y_i^4]}_{1} + \underbrace{6}_{\binom{4}{2} \text{ permutations of } (i,i,j,j)} \sum_{i<j} f_i^2 f_j^2 \underbrace{\mathbb{E}[Y_i^2 Y_j^2]}_{1}$$

purpose of the rest of the calculation is to relate the expression to F_2^2.

$$= \sum_i f_i^4 + 6 \boxed{\sum_{i<j} f_i^2 f_j^2} = F_4 + 3(F_2^2 - F_4) = 3F_2^2 - 2F_4 \le 3F_2^2$$

$$\sum_{i<j} f_i^2 f_j^2 = \frac{1}{2}\left(\underbrace{\left(\sum_{i=1}^n f_i^2\right)^2}_{F_2^2} - \underbrace{\sum_{i=1}^n f_i^4}_{F_4}\right)$$

So $\mathrm{Var}[X^2] = \mathbb{E}[X^4] - F_2^2 \leq 3F_2^2 - F_2^2 = 2 \cdot F_2^2$ ■

Do the Median-of-Means trick to obtain the final estimator. That is, average $O(\frac{1}{\epsilon^2})$ copies of the basic estimator to drop its variance below (say) $\frac{\epsilon^2 F_2^2}{4}$.

Then Chebyshev's inequality says that

$$\Pr\left[\begin{array}{l}\text{estimate is off from expectation} \\ \text{by more than } \epsilon \cdot F_2\end{array}\right] \leq \frac{1}{4}.$$

So a median of $O(\log(\frac{1}{\delta}))$ copies of the above is an (ϵ, δ)-approximation algorithm.