Algorithms for (strict) turnstile streams $\sigma = \langle (a_1, \Delta_1), \ldots, (a_m, \Delta_m) \rangle$

Point queries: Goal is to output a sketch from which one can derive, for any $i \in [n]$, an estimate $\hat{f}_i$ of f_i such that

$$(*) \quad 0 \le \hat{f}_i - f_i \le \epsilon \cdot \|f\|_1$$

$$\|f\|_1 = \sum_i f_i = M.$$

Subtle distinction in randomized case:

- "For-all error guarantee": with probability $\ge 1-\delta$, $(*)$ hold simultaneously for all estimates $\hat{f}_i$ returned by the algorithm.
- "For-each error guarantee": For each $i \in [n]$, $(*)$ holds with probability $1-\delta$.

One can turn any For-each sketch into a for-all sketch with an $O(\log n)$ blowup in space by reducing failure probability to $\frac{\delta}{n}$ using the median trick and then union bounding over all $i \in$

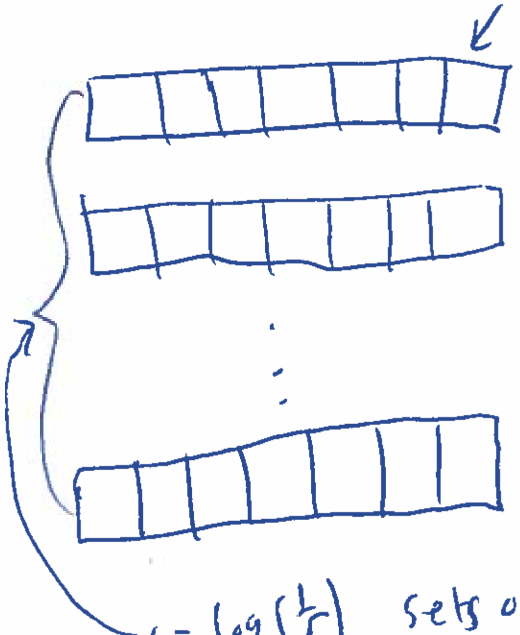
Recall: • Misra-Gries achieves this goal for insert-only streams using $O(\frac{\log n}{\epsilon})$ bits of space (since it is deterministic, it is automatically a for-all sketch).

• There is a trivial way to turn Misra-Gries into a turnstile streaming algorithm, but the error will grow like $\epsilon \cdot \sum_{j=1}^{m} |\delta_j|$ instead of $\epsilon \cdot \sum_{i=1}^{n} f_i$.

Run separate instances of Misra-Gries on positive increments and negative increment updates, Let $\hat{f}_i$ be the difference of the estimates returned for i by the two instances. Error is at most the sum of the errors in the two estimates so at most $\epsilon \cdot \sum_j |\delta_j|$.

Count-Min Sketch [CM05]

$k = \frac{2}{\epsilon}$ counters



$t = \log(\frac{1}{\delta})$ sets of counters

• choose t hash functions $h_1, \ldots h_t : [n] \to [k]$ at random from a pairwise independent hash family

• When processing update (a_j, δ_j):

· For $\ell = 1 \ldots t$

$$C[\ell][h_\ell(a_j)] \mathrel{+}= \delta_j$$

• On query i output $\hat{f}_i := \min_{1 \le \ell \le t} C[\ell][h_\ell(i)]$

In the strict turnstile model, it is clear each $\hat{f}_i$ is always an overestimate of f_i, since a counter's value is just the sum of the frequencies of all items that hash to it.

Claim: For any fixed $i \in [n]$, $0 \le \hat{f}_i - f_i \le \epsilon \cdot M$ with probability $\ge 1-\delta$.
This is a for-each error guarantee.

Proof: We analyze the "excess" in each counter $C[\ell][h_\ell(i)]$ that i hashes to. Let $X_\ell := C[\ell][h_\ell(i)] - f_i$ denote this excess. For each $j \neq i$, let

$$Y_{\ell,j} = \begin{cases} 1 & \text{if } h_\ell(i) = h_\ell(j) \\ 0 & \text{otherwise} \end{cases}. \quad \text{Then } X_\ell = \sum_{j \in [n] \setminus \{i\}} f_j \cdot Y_{\ell,j}.$$

By pairwise independence of the hash family, $\mathbb{E}[Y_{\ell,j}] = \frac{1}{K}$. Thus, by linearity of expectation,

$$\mathbb{E}[X_\ell] = \sum_{j \in [n] \setminus \{i\}} \frac{f_j}{K} = \frac{M - f_i}{K}.$$

Since each $f_j \ge 0$, X_ℓ is a non-negative random variable, so we can apply Markov's inequality to conclude $\Pr[X_\ell \ge \epsilon \cdot M] \le \frac{1}{\epsilon \cdot K} \overset{K = 2/\epsilon}{=} \frac{1}{2}$.

Since the t hash functions are mutually independent,

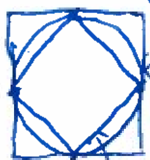
$$\Pr[\hat{f}_i - f_i \ge \epsilon \cdot M] = \Pr[X_\ell \ge \epsilon \cdot M \text{ for } \underline{\text{all}}\ \ell \in [t]]$$

$$= \prod_{\ell=1}^{t} \Pr[X_\ell \ge \epsilon M] \le \frac{1}{2^t} \le \delta. \blacksquare$$

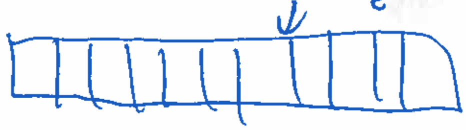
Note: To get a for-all error guarantee need to increase space usage from $O(\frac{1}{\epsilon} \log(\frac{1}{\delta})(\log m + \log n))$ to $O(\frac{1}{\epsilon} \cdot \log(\frac{n}{\delta})(\log m + \log n))$.

Count sketch [CCFC04].

Guarantee of this sketch: For each fixed $i \in [n]$, one can derive an estimate $\hat{f}_i$ of f_i such that with probability $1-\delta$, $|\hat{f}_i - f_i| \le \varepsilon \cdot \|f\|_2$.
The sketch uses space $O(\frac{1}{\varepsilon^2} \cdot \log(\frac{1}{\delta}) \cdot (\log m + \log n))$.
Note: For <u>any</u> vector f, $\|f\|_2 \le \|f\|_1$, so it is incomparable to Countmin:
it uses a factor $\frac{1}{\varepsilon}$ more space but its error might be smaller.



Algorithm:



Basic Estimator

- Choose a random hash function $h : [n] \to [k]$ from a pairwise independent hash family
- Choose a random hash function $g : [n] \to \{-1, 1\}$ from a pairwise independent hash family
- When processing update (a_j, δ_j):
  $C[h(a_j)] \pm \delta_j \cdot g(a_j)$.
- On query i, output $\hat{f}_i := g(i) \cdot C[h(i)]$

Full sketch runs $\log(\frac{1}{\delta})$ copies of basic estimator and outputs the median estimate.

Analysis of Basic Estimator: Fix $i \in [n]$ and let $X := \hat{f}_i$. Claim: $\mathbb{E}[X] = f_i$. For each $j \ne i$, let
$$Y_j := \begin{cases} 1 & \text{if } h(j) = h(i) \\ 0 & \text{otherwise} \end{cases}.$$
Then
$$X = g(i) \cdot \sum_{j=1}^{n} f_j \cdot g(j) \cdot Y_j = f_i + \sum_{j \in [n] \setminus \{i\}} f_j \cdot g(i) \cdot g(j) \cdot Y_j.$$

Since g, h are independent we have:
$$(*)\quad \mathbb{E}_{g,h}[g(j) Y_j] = \mathbb{E}[g(j)] \cdot \mathbb{E}[Y_j] = 0 \cdot \mathbb{E}[Y_j] = 0.$$
($= \mathbb{E}[g(i)] \cdot \mathbb{E}[g(j)] \cdot \mathbb{E}[Y_j]$ by pairwise indep. of g)

So by linearity of expectation,
$$\mathbb{E}[X] = f_i + \sum_{j \in [n] \setminus \{i\}} f_j \cdot \mathbb{E}_{g,h}[g(i) \cdot g(j) \cdot Y_j] \overset{\text{by } (*)}{=} f_i + 0 = f_i.$$

Similar to (*) we also have:

(**) For any $j \neq j'$, $\mathbb{E}[g(j) \cdot g(j') \cdot Y_j \cdot Y_{j'}] = \mathbb{E}[g(j)] \cdot \mathbb{E}[g(j')] \cdot \mathbb{E}[Y_j \cdot Y_{j'}]$

$= 0 \cdot 0 \cdot \mathbb{E}[Y_j Y_{j'}] = 0$

In addition we have:

(***): $\mathbb{E}[Y_j^2] = \mathbb{E}[Y_j] = \Pr[h(j) = h(i)] = \frac{1}{k}$.

Claim 2: $\mathrm{Var}[X] \leq \dfrac{\|f\|_2^2}{k}$,

Proof: $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - f_i^2$.

$$\mathbb{E}[X^2] = \mathbb{E}\Big[\Big(f_i + \sum_{j \in [n]\setminus\{i\}} f_j \cdot g(i) \cdot g(j) \cdot Y_j\Big)^2\Big]$$

distributive law

$$= \mathbb{E}\Big[f_i^2 + f_i \sum_{j \in [n]\setminus\{i\}} f_j \cdot g(i) \cdot g(j) \cdot Y_j + \Big(\sum_{j \in [n]\setminus\{i\}} f_j \cdot g(i) \cdot g(j) \cdot Y_j\Big)^2\Big]$$

linearity of expectation

$$= f_i^2 + \underbrace{f_i \cdot \sum_{j \in [n]\setminus\{i\}} f_j\, \mathbb{E}[g(i) \cdot g(j) \cdot Y_j]}_{=0 \text{ by } (*)} + \mathbb{E}\Big[\Big(\sum_{j \in [n]\setminus\{i\}} f_j\, g(i) \cdot g(j) \cdot Y_j\Big)^2\Big]$$

~~linearity of expectation~~

$$= f_i^2 + \mathbb{E}\Big(\sum_{j \in [n]\setminus\{i\}} f_j^2 \cdot \underset{=1}{g^2(i)} \cdot \underset{=1}{g^2(j)} \cdot Y_j^2 + \sum_{j \neq j' \in [n]\setminus\{i\}} f_j \cdot \underset{=1}{g(i) \cdot g(j) \cdot Y_j \cdot f_{j'} \cdot g(i)} \cdot Y_{j'}\Big)$$

$$= f_i^2 + \mathbb{E}\Big(\sum_{j \in [n]\setminus\{i\}} f_j^2 Y_j^2 + \sum_{j \neq j' \in [n]\setminus\{i\}} f_j \cdot f_{j'} \cdot g(j) \cdot g(j') \cdot Y_j \cdot Y_{j'}\Big)$$

linearity of expectation

$$= f_i^2 + \sum_{j \in [n]\setminus\{i\}} f_j^2\, \mathbb{E}[Y_j^2] + \underbrace{\sum_{j \neq j' \in [n]\setminus\{i\}} f_j \cdot f_{j'} \cdot \mathbb{E}[g(j) \cdot g(j') \cdot Y_j \cdot Y_{j'}]}_{=0 \text{ by } (**)}$$

by $(***)$

$$= f_i^2 + \sum_{j \in [n]\setminus\{i\}} f_j^2 / k.$$

Hence $\text{Var}[X] = \mathbb{E}[X^2] - f_i^2 = \cancel{f_i^2} + \sum_{j \neq i} f_j^2 / k - \cancel{f_i^2}$

$$\leq \frac{\|f\|_2^2}{k}.$$

By Chebyshev, $\Pr[|X - \mathbb{E}[X]| \geq \underbrace{\epsilon \cdot \|f\|_2}_{= \epsilon\sqrt{k} \cdot \sqrt{\text{Var}[X]}}] \leq \frac{1}{(\epsilon \cdot \sqrt{k})^2} = \frac{1}{\epsilon k^2} \overset{\text{choose } k}{\leq} \frac{1}{3}.$

Since the basic estimator gives an estimate with error $\leq \epsilon \cdot \|f\|_2$ with probability $\geq 2/3$, the median of $O(\log(\frac{1}{\delta}))$ basic estimators satisfies the error bound w.p. $\geq 1 - \delta$.