Algorithms for (strict) turnstile streams $\sigma = \langle A_i, d_i \rangle_{i=1}^{m}$: 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

$$0 \leq \hat{f}_i - f_i \leq \epsilon \|f\|_1$$

$$\sum_i \hat{f}_i = M.$$ 

Subtle distinction in randomized case:

* "For-all error guarantee": with probability $\geq 1-\delta$, (*) hold simultaneously for all estimates $\hat{f}_i$ returned by the algorithm.

* "For-each error guarantee": For each $i$, (*) 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 $\delta/n$ using the median trick and then union bounding overall.
Recall: Misra-Gries achieves this goal for insert-only streams using \( O\left(\frac{\log n}{\epsilon}\right) \) bit 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 since

\[
\epsilon \cdot \sum_{j=1}^{m} \left| f_j \right| \text{ instead of } \epsilon \cdot \sum_{i=1}^{n} f_i.
\]

Run separate instances of Misra-Gries on positive increments and negative decrement updates. Let \( f_i \) 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 \( \epsilon \cdot \sum_{j=1}^{m} \left| f_j \right| \).

**Count-Min Sketch** [Cm05]

- \( \epsilon \cdot \frac{1}{\epsilon} \) counters
- \( t = \log \left( \frac{1}{\delta} \right) \) sets of counters
- Choose \( t \) hash functions \( h_1, \ldots, h_t : [n] \rightarrow [K] \) at random from a pairwise independent hash family.
- When processing update \((a_j, f_j)\):
  - For \( \ell = 1 \ldots t \)
    - \( C[\ell][h_\ell(a_j)] \leftarrow f_j \)
- On query output \( \hat{f}_i = \min_{1 \leq \ell \leq t} C[\ell][h_i] \)
In the transitive model, it is clear that \( f_i \) is always an overestimate of \( f_i \), since its counter's value is just the sum of the frequencies of all items that hash to it.

Claim: For any fixed \( i \in [n] \), \( 0 \leq \hat{f}_i - f_i \leq e \cdot M \) with probability \( \geq 1 - \delta \).

Proof: We analyze the "excess" mean counter \( C[i] \) that \( \hat{f}_i \) hashes to. Let \( X_e = C[i] - f_i \) denote this excess. For each \( j \neq i \), let

\[
Y_{e,j} = \begin{cases} 
1 & \text{if } h(e) = h(j), \\
0 & \text{otherwise} 
\end{cases}
\]

Then \( X_e = \sum_{j \in [n] \setminus i} f_j \cdot Y_{e,j} \).

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

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

Since each \( f_j \geq 0 \), \( X_e \) is a non-negative random variable. Hence, we can apply Markov's inequality to conclude

\[
\Pr[X_e \geq e \cdot M] \leq \frac{1}{e \cdot K} \leq \frac{\delta}{2}.
\]

Since the hash functions are mutually independent,

\[
\Pr[\hat{f}_i - f_i \geq e \cdot M] = \Pr[X_e \geq e \cdot M \text{ for all } e \in [n]]
\]

\[
\leq \frac{2}{\delta} \sum_{e=1}^{\delta} \Pr[X_e \geq e \cdot M] \leq \frac{1}{\delta} \leq \delta.
\]

Note: To get a formal error guarantee, need to increase space usage from

\[
O\left( \frac{1}{\delta} \log(\frac{1}{\delta}) \left( \log m + \log n \right) \right)
\]

to

\[
O\left( \frac{1}{\delta} \cdot \log(\frac{1}{\delta}) \left( \log m \log n \right) \right).
\]
Count sketch [CLP04].

For any fixed \( i \in [n] \), one can derive an estimate \( \hat{f}_i \) of \( f_i \) such that with probability \( \geq \frac{3}{4} \), \( |\hat{f}_i - f_i| \leq \epsilon_i \|f\|_2 \).

The sketch uses space \( O(n^{1/2} \log(n) \cdot (\log \log(n))^{-2}) \).

Note: For any vector \( f \), \( \|f\|_2 \leq \|f\|_1 \), so it is in comparable to CardOMP.

It uses a factor \( \frac{1}{\epsilon} \) more space, but its error might not be smaller.

Algorithm:

Full sketch runs \( \log(\frac{i}{\epsilon}) \) copies of \( \hat{f}_i \).

The median estimate.

Analysis of Basic Estimator: Fix \( i \in [n] \) and let \( \hat{X} = \hat{f}_i \). For each \( j \neq i \), let \( Y_j = \hat{f}_j \) if \( h(j) = h(i) \), otherwise \( Y_j = 0 \).

Then \( \hat{X} = g(i) \sum_{j=1}^{n} f_j - g(j) \), \( Y_j = f_j + \sum_{j \in [n], j \neq i} f_j \cdot g(i) \cdot g(j) \).

Since \( g, h \) are independent, we have:

\[
\sum_{j=1}^{n} g(i) \cdot g(j) \cdot Y_j = \sum_{j=1}^{n} g(i) \cdot Y_j = 0 \text{ for } \sum_{j=1}^{n} g(j) = 0.
\]

So by linearity of expectation,

\[
\mathbb{E}[\hat{X}] = \hat{f}_i + \sum_{j \in [n], j \neq i} \mathbb{E}[g(i) \cdot g(j) \cdot Y_j] = \hat{f}_i + \sum_{j \in [n], j \neq i} \mathbb{E}[g(j) \cdot Y_j] = \hat{f}_i + \hat{f}_j,
\]

Claim: \( \mathbb{E}[\hat{X}] = \hat{f}_i \).
Similar to (*) we also have:

\[ (*) \text{ For any } j \neq i, \quad \mathbb{E}[g(i) \cdot g(j) \cdot y_i \cdot y_j] = \mathbb{E}[g(i)] \mathbb{E}[g(j)] \cdot \mathbb{E}[y_i] \cdot \mathbb{E}[y_j] = 0. \]

In addition we have:

\[ (***) \text{ For } y_i \quad \mathbb{E}[y_i^2] = \mathbb{E}[y_i] = \Pr[h(i) = h(i)] = \frac{1}{k}. \]
Claim 2: \( \text{Var}[X] \leq \frac{11f_{11}^2}{K} \).

Proof: \( \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - f_c^2. \)

\[ \mathbb{E}[X^2] = \mathbb{E} \left[ \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot g(i,j) \cdot y_j \right]^2 \]

By the distributive law

\[ = \mathbb{E} \left[ f_c^2 + \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot g(i,j) \cdot y_j + \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot g(i,j) \cdot y_j \right] \]

Linearity of expectation

\[ = f_c^2 + f_c \cdot \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} \mathbb{E} \left[ g(i,j) \cdot y_j \right] + \mathbb{E} \left[ \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot g(i,j) \cdot y_j \right] \]

By (\star)

\[ = 0 \quad \text{by (\star)} \]

Thus

\[ = f_c^2 + \mathbb{E} \left( \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j^2 \cdot y_j^2 + \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot g(i,j) \cdot y_j \right) \]

Linearity of expectation

\[ = f_c^2 + \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j^2 \cdot \mathbb{E}[y_j^2] + \sum_{j \in \mathcal{E}_n \setminus \mathcal{E}_3} f_j \cdot f_c \cdot \mathbb{E}[g(i,j) \cdot y_j] \]

By (\star\star)

\[ = 0 \quad \text{by (\star\star)} \]
Hence $\text{Var}[X] = 4\mathbb{E}[X] - f_i^2 = \sum_{j \neq i} f_j^2 / k$. 

By Chebyshev, $\Pr[|X - \mathbb{E}[X]| \geq 6 \cdot 11F_{1/2}^2] \leq \frac{1}{(\text{Var}[X])^2} = \frac{1}{\frac{1}{k^2}} \leq \frac{1}{3}$. 

Since the basic estimator gives an estimate with error $\leq 6 \cdot 11F_{1/2}$ with probability $2/3$, the median of $O(\log(\frac{1}{\delta}))$ basic estimators satisfies the error bound $w.p. \geq 1-\delta$. 

$\text{Var}[X] = \frac{1}{k}$.