Stream: $\sigma =< a_1, a_2, ..., a_m >$, token $a_i \in U = [n]$.
Goal: process σ using space s st $s << m \wedge s << n$, $s = O(\min(m, n))$, pref $s = O(\log m + \log n)$.
Sometimes best $s = polylog(\min(m, n))$, $polylog = O((\log(g(n))^c)$ for $c > 0$.
Streams can only be access in sequence.
Multiplicative condition: Let $A(\sigma)$ be random stram alg A on σ: Let ϕ be target function. A is an ϵ, δ-approx alg of ϕ if

$$\mathbf{Pr}[|\frac{A(\sigma)}{\phi(\sigma)} - 1| > \epsilon] \leq \delta$$

To strong a condition when $\phi(\sigma)$ has values close to zero.

Additive-approx: (ϵ, δ)-additative-approx-alg A of ϕ if

$$\mathbf{Pr}[|A(\sigma) - \phi(\sigma)|] \leq \sigma$$

Often instrested in statistical properties of multiset in σ. Let vector $\vec{F} = (f_1, ..f_n)^T$, $f_j = |\{i : a_i = j\}|$. So $\phi(\vec{F})$ target. $\vec{F}$ frequency vector.
Turnstile model: $\sigma : [n] \times \{-L, ..., L\}$ so $f_j \sum_{|\{i:a_i=j\}|} l_i$. Redefined m to max number tokens in multiset $\|\vec{F}\|_1 \leq m$.
Strict turnstile model: $\vec{F} \geq 0$
Cash register model: $\forall i.l_i > 0$

1 Frequency Problem

Majority problem: if $\exists j : f_j > m/2$, then output j else null. Frequency problem: For some k, output $\{j : f_j > m/k\}$
Frequency-estimateion problem: For stream σ produce structure that can estimate $\hat{f}_a$ for freq f_a for $a \in [n]$.
Misra-Gries Alg: Takes param k (same as freq problem). Maintain associative array, fx bbt

```
Process j:
  If j in keys(A) then
    A[j] <- A[j]+1
  else if |keys(A)| < k -1 then
    A[j] <- 1
  else
    foreach l in keys(A)
      A[l] <- A[l] - 1
      if A[l] = 0 then remove l
  output: f^_a = A[a] if keys(a) else 0
```

Thrm The Misra-Gries alg. with param k uses one pass and $O(k(\log m + \log n))$ bits of space and fives estimate $\hat{f}_j$ satisfying

$$\max(0, f_j - \frac{m}{k}) \leq \hat{f}_j \leq f_j$$

Proof: When $A[j]$ decr then we decr k-1 other counters $\Rightarrow$ decr witnessed (applied to) k tokens $\Rightarrow \leq m/k$ decrs as $|\sigma| = m$, therefore $\hat{f}_j \leq f_j - m/k$.

2 The median trick

Event $X = \sum_1^t x_i$, X is bad if $|X - f_x| \geq \epsilon \|f_{-x}\| \geq \gamma$
$X_{t/2}$ (median) bad implies $X_{<t/2}$ bad events.
$B_i \equiv X_i$ bad, $B = \sum B_i$
$X_{t/2} \Rightarrow B \geq t/2$, $\mathbf{Pr}[B_i] \leq 1/3 \Rightarrow E[B] \leq t/3$
$\mathbf{Pr}[X_{t/2} bad] \leq exp(\Omega(t))$

3 Sketchs (Ran freq algs)

Sketch: Datastructure d st. $d(\sigma_1 \cdot \sigma_2)$, $\cdot$ is string concat, can be computed from $d(\sigma_1)$ and $d(\sigma_2)$ using space efficient comb alg C $C(d(\sigma_1), d(\sigma_2)) = d(\sigma_1 \cdot \sigma_2)$.
Linear Sketch: Sketching alg. A st $\forall \sigma \subseteq [n]$ $A(\sigma)$ takes value in vector space of dim $l = l(n)$ and $A(\sigma)$ is linear function of $\vec{F}(\sigma)$. l is the dim of linear sketch. Works under turnstile and strict turnstile model.

3.1 Count Sketch

Basic sketch parameter ϵ (desired accuracy)

```
Init
  C[1,..,k] <- 0,.,0
  h: [n] -> [k] 2-unviersal hash function
  g: [k] -> {-1,1} 2-unviersal hash function

Process(j,c)
  C[h(j)] += cg(j)

Query(a in [N])
  f^_a = g(a)C[h(a)]
```

Analysis of Basic Sketch Estimate:
Fix a, consider $X = \hat{f}_a$, let $Y_j = 1$ iff $h(j) = h(a)$, j contributes to counter $C[h(a)] \Leftrightarrow h(j) = h(a)$, contributes sign $g(j)$ and freq f_j.
Therefore,

$$X = g(a) \sum_{1}^{n} f_j g(j) Y_j = f_a + \sum_{j \neq a} f_j g(a) g(j) Y_j$$

$$E[g(j)Y_j] = E[g(j)]E[Y_j] = (1/2 - 1/2)E[Y_j] = 0$$

first equality uses g is independent of h, above implies

$$E[X] = f_a + \sum f_j g(a) E[g(j)Y_j] = f_a$$

f_j and $g(a)$ are constants. Hence $X = \hat{f}_a$ is an unbiased estimator for f_a.
Second moment of Basic sketch:
By 2-unviersality of $h \in \mho(h)$ we have $\forall j \in [n] - a$

$$E[Y_j^2] = E[Y_j] = \mathbf{Pr}[h(j) = h(a)] = 1/k$$

By 2-unviersality of g and independence of g and h $\forall i, j \in [n], i \neq j$

$$E[g(j)g(i)Y_iY_j] = E[g(j)]E[g(i)]E[Y_iY_j] = 0 * 0 * E[Y_iY_j] = 0$$

This implies that

$$\begin{aligned}
var[X] &= 0 + g(a)^2 Var[\sum_{j \neq a} f_j g(j) Y_j && (1)\\
&= E[\sum_{j \neq a} f_j^2 Y_j^2 + \sum_{i,j \neq a, i<j} f_i f_j g(i) g(j) Y_i Y_j] - (\sum_{j \neq a} f_j E[g(j)Y_j])^2 && (2)\\
&= \sum_{j \neq a} f_j^2/k + 0 + 0 = \|f\|_2^2 - f_a^2/k && (3)
\end{aligned}$$

where f is frequency distribution determine by σ.
From Chebyshev inequality

$$\begin{aligned}
\mathbf{Pr}[|\hat{f}_a - f_a| \geq \epsilon\sqrt{\|f\|_2^2 - f_a^2}] &= \mathbf{Pr}[|X - E[X]| \geq \epsilon\sqrt{\|f\|_2^2 - f_a^2}] && (4)\\
&\leq \frac{var[x]}{\epsilon^2(\|f\|^2 - f_a^2} && (5)\\
&= 1/k\epsilon^2 = 1/3 && (6)
\end{aligned}$$

For $j \in [n]$ let f_{-j} denote the (n-1) dimensional vector obtained by dropping j'th entru of f then

$$\mathbf{Pr}[|\hat{f}_a - f_a| \geq \epsilon \|f_a\|_2^2] \leq 1/3$$

3.2 Count sketch alg

```
Init
  C[1..t][1..k] <- 0, k = 3/epsilon, t = O(log(1/delta))
  choose t indep 2-universal hash funsc h_1,..,h_t:[n]->[k]
  choose t indep 2-universal hash funcs g_1,..,g_t:[k]->{-1,1}

Process(j,c)
  for i=1 to t C[i][h_i(j)] += c g_i(j)

Query(a)
  f^_a = median_{1\leq i \leq t} C[i][h_i(a)]g_i(a)
```

Chernoff bound argument proves $\hat{f}_a$ satiesfies

$$\mathbf{Pr}[|\hat{a}_a - f_a| \geq \epsilon \|f_{-a}\|] \leq \delta$$

Hash fucntion stored in $O(t \log n)$ space. each of the tk countes uses $\log m$ space. This gives space bound

$$O(t \log n + tk \log m) = O(1/\epsilon^2 \log 1/\delta(\log n + \log m))$$

4 Count-Min Sketch