Random variables $X_1, \ldots, X_n$ over $\mathcal{Z}$.

*Mutually independent* if:

$$\forall i_1, \ldots, i_n \in \mathcal{Z}, \quad \Pr\left( \bigcap_{j=1}^{n} X_j = i_j \right) = \prod_{j=1}^{n} \Pr(X_j = i_j)$$

*\(k\)-wise independent* if:

$$\forall S \subseteq \{1, \ldots, n\} \text{ where } |S| \leq k, \quad \forall i_1, \ldots, i_k \in \mathcal{Z},
\Pr\left( \bigcap_{j \in S} X_j = i_j \right) = \prod_{j=1}^{k} \Pr(X_j = i_j)$$

Pairwise independent is \(k=2\), i.e.,

$$\forall j, k \in \{1, \ldots, n\}, a, b \in \mathcal{Z},
\Pr(X_j = a, X_k = b) = \Pr(X_j = a) \times \Pr(X_k = b).$$
Lemma: Let \( X_1, \ldots, X_n \) be pairwise independent, and let \( X = \sum_{i=1}^{A} X_i \).

Then, \( \text{Var}(X) = \sum_{i=1}^{A} \text{Var}(X_i) \)

and if \( X_i's \) are binary/indicator random variables \( 0-1 \)

then \( \text{Var}(X) \leq \sum_{i=1}^{A} \mathbb{E}[X_i^2] = \sum_{i=1}^{A} \mathbb{E}[X_i] = \mathbb{E}[X] \).
Simple construction of pairwise independent random variables:

from a,b mutually indpt. random bits, generate \( m = 2^{b^2 - 1} \) pairwise indpt. random bits.

Let \( X_1, \ldots, X_b \in \{0,1\} \) be uniform, mutually indpt. random bits.

Let \( S_1, \ldots, S_{2^b - 1} \) be the nonempty subsets of \( \{1, \ldots, b\} \).

Set \( Y_j = \bigoplus_{i \in S_j} X_i \mod 2 \)

Note, \( Y_j \in \{0,1\} \)

**Lemma:** The \( Y_j \)'s are pairwise indpt.

**Proof:**

**Claim 1:** \( \Pr(Y_j = 1) = \Pr(Y_j = 0) = \frac{1}{2} \)

Why? Let \( S_j = \{ z_1, \ldots, z_{\ell} \} \subseteq \{1, \ldots, b\} \)

So, \( Y_j = (\sum_{i=1}^{\ell-1} X_{z_i} \mod 2) + X_{z_{\ell}} \mod 2 \)

Reveal \( X_{z_1}, \ldots, X_{z_{\ell-1}} \). Whatever this is, with prob. \( \frac{1}{2} \) \( X_{z_{\ell}} = 1 \) & \( Y_j \) is opposite & w.p. \( \frac{1}{2} \) \( X_{z_{\ell}} = 0 \) & \( Y_j \) is the same.

This is the principle of deferred decisions.
Now to see pairwise independence:

Fix $S_j$ & $S_k$ and consider some $z \in S_j \setminus S_k$.

$$\Pr(y_j = a, y_k = b) = \Pr(y_j = a | y_k = b) \Pr(y_k = b)$$

$$= \Pr(y_j = a | y_k = b) \frac{1}{2} \times \frac{1}{2}$$

we just showed this.

Reveal all $X_i$'s but $X_z$

Then with prob. $\frac{1}{2}$ $X_z = 1$ & $y_j$ flips

& w.p. $\frac{1}{2}$ $\exists X_z = 0$ & $y_j$ is the same

Therefore, $\Pr(y_j = a | y_k = b)$

$= \Pr(y_j = a | X_1, \ldots, X_{b \setminus X_z})$

$= \frac{1}{2}$

& thus $\Pr(y_j = a, y_k = b) = \frac{1}{4}$. 
More sophisticated construction:

For prime $p$, given $a, b$ which are independent and uniform over $\{0, 1, \ldots, p-1\}$, then we construct $Y_1, \ldots, Y_{p-1}$, which are pairwise independent and uniform over $\{0, 1, \ldots, p-1\}$.

Namely, let $Y_i = a + ib \mod p$ for $i = 0, \ldots, p-1$.

Lemma: The $Y_i$'s are pairwise independent.

Proof: First, $Y_i$ is uniform over $\{0, 1, \ldots, p-1\}$. Why? By principle of deferred decisions again.

For any $b$ & $i$, & $\alpha$ in $\{0, \ldots, p-1\}$,

$$\Pr(Y_i = \alpha) = \Pr(a + bi = \alpha \mod p)$$

$$= \Pr(a = \alpha - bi \mod p)$$

$$= \frac{1}{p} \quad \text{Since there is a unique such } a \text{ in } \{0, 1, \ldots, p-1\}.$$
Now consider \( i \neq j \rightarrow p \neq \| \text{ and } x, \beta \in \mathbb{Z}, \ldots, p-1 \).

We'll show: \( \Pr(Y_i = x, Y_j = \beta) = \frac{1}{p^2} \) & we're done.

\[
Y_i = x \iff a + ib = \alpha \mod p
\]
\[
Y_j = \beta \iff a + jb = \beta \mod p.
\]

Thus, \( x - \beta = b(i-j) \mod p \)

\[
b = \frac{x - \beta}{i-j} \mod p
\]

which is valid since \( i-j \neq 0 \)

& \( p \) is prime.

\& \( a = \alpha - bi \mod p \)

So there is a unique \((a, b)\) pair.

So that \( Y_i = \alpha, Y_j = \beta \)

Therefore, \( \Pr(Y_i = x, Y_j = \beta) \)

\[
= \Pr(b = \frac{x - \beta}{i-j} \mod p, a = \alpha - \frac{1}{i-j}(x - \beta) \mod p)
\]

\[
= \frac{1}{p^2}
\]

\[ \square \]

For \( n \) which is not prime, can choose \( p > n \)

where \( p \) is prime & \( p < 2n \).

Note, the random variables take \( O(\log n) \) bits to represent \( a, b \).
Streaming model:

Stream \( S = \{ s_1, ..., s_m \} \) for HUGE \( m \).

where each \( s_i \in \{0, 1, ..., n-1\} \)

Let \( f = (f_1, ..., f_n) \) where

\[
f_i = \left| \{ j : s_j = i \} \right| = \text{freq. of occurences of value } i \text{ in } S.
\]

Let \( Q = F_0 = \left| \{ i : f_i > Q \} \right| = \# \text{ of distinct values in } S \)

Goal: find \( Q \) with \( O(\log n) \) space.

aim for \((\varepsilon, \delta)\)-approx of \( Q \):

output \( \hat{Q} \) where \( \Pr[Q(1-\varepsilon) \leq \hat{Q} \leq (1+\varepsilon)Q] \geq 1-\delta \)

in space \( \text{poly}(\log n, \frac{1}{\varepsilon}, \log(1/\delta)) \)
For integer $k > 0$, let $\text{zeros}(k) = \#$ of 0's at end of binary representation of $k$.

$$= \max \{ l : 2^l \text{ divides } k \}$$

e.g., if $k$ is odd then $\text{zeros}(k) = 0$
even then $\text{zeros}(k) = 1$ since it ends in 0.

AMS algorithm: (2nd last class, but same paper)

1. Choose a random hash function $h : [n] \rightarrow [n]$

   $$\{0, 1, \ldots, n-1\} \rightarrow \{0, 1, \ldots, n-1\}$$

   which is pairwise independent

   (if $n$ is not prime, choose prime $p$ such that $n \leq p \leq 2n$)

2. Set $z = 0$

3. Go through stream & at element $k$:

   if $\text{zeros}(h(k)) > z$

   then $z = \text{zeros}(h(k))$

4. output $\left(2^{z + \frac{1}{2}}\right)$
Intuition for alg.

$h(k)$ is uniformly random bit string so prob. it has $\text{zeros}(h(k)) = \log d$ is prob. of last $\log d$ bits all = 0, which is prob. $2^{-\log d}$.

Thus, prob. that $\text{zeros}(h(k)) = \log d$ is $2^{-\log d} = \frac{1}{d}$.

So, if $d$ distinct items then expect 1 out of these to have $\text{zeros}(h(k)) = \log d$.

& for $l \gg \log d$, it's unlikely that $\text{zeros}(h(k)) \gg \log d$.

Thus, the max $\text{zeros}(h(k))$ is a good approx. for $\log d$.

Analysis:

For $k \in \{0, ..., n-1\}$, & integer $l \geq 0$,

let $X_{l,k} = \begin{cases} 1 & \text{if } \text{zeros}(h(k)) \geq l \\ 0 & \text{otherwise} \end{cases}$

& let $Y_l = \sum_{k: f_k \geq 0} X_{l,k}$
Let $+ l$ be the final value of $z$ at end of algorithm.

$$
+ l \iff Y_l > 0
$$

$$
+ = \max \{ l : Y_l > 0 \}
$$

$$
\text{this is same as:}

+ < l \iff Y_l = 0.
$$

\[ t \leq l - 1 \]

Note, $h(k)$ is uniform, i.e., $Pr(h(k) = j) = \frac{1}{n}$ for all $j \in \{0, \ldots, n-1\}$

Thus, $E[X_{l,k}] = Pr(\# \text{zeros}(h(k)) \geq l)$

$$
= Pr(\text{last } l \text{ bits of } h(k) \text{ are all } 0)
$$

$$
= \frac{1}{2^l}
$$
\[ \text{Var}(Y_e) = \sum_{k: f_k > 0} \text{Var}(X_{e,k}) \quad \text{since } X_{e,k} \text{ are pairwise indep.} \]

\[ \leq \sum_{k: f_k > 0} \mathbb{E}[X_{e,k}^2] \]

\[ = \sum_{k: f_k > 0} \mathbb{E}[X_{e,k}] \]

\[ = \frac{d}{2^l} \]

By Markov's ineq.,

\[ \mathbb{P}(Y_e > 0) = \mathbb{P}(Y_e \geq 1) \leq \frac{\mathbb{E}[Y_e]}{1} = \frac{d}{2^l} \]

By Chebyshev's (can apply since pairwise indep.),

\[ \mathbb{P}(Y_e = 0) \leq \mathbb{P}\left( |Y_e - \mathbb{E}[Y_e]| \geq \frac{d}{2^l} \right) \]

\[ \leq \frac{\text{Var}(Y_e)}{\left(\frac{d}{2^l}\right)^2} \leq \frac{2^l}{d} \]
Our goal is to output $\hat{d}$.
Let's aim for $\hat{d}$ where $\frac{2}{3} \leq \hat{d} \leq 3\hat{d}$
so $\hat{d}$ is a $3$-approx. of $d$.

Let $a$ be the smallest integer s.t. $2^{a + \frac{1}{2}} \geq 3d$.
We want to show that the prob. that $t = \text{final value of } z$ is unlikely to
be as large as $a$.

$$\Pr(\hat{d} \geq 3d) = \Pr(t \geq a) = \Pr(Y_a > 0) \leq \frac{d}{2^a} = \frac{\sqrt{2}}{3} \approx 0.41 $$

Note, $2^{a + \frac{1}{2}} \geq 3d$
$$\frac{3d}{2^a} \leq \frac{\sqrt{2}}{3}$$

Thus, $\Pr(\hat{d} < 3d) > .51$
On the other side, we want \( \hat{d} \geq \frac{d}{3} \)

let \( b \) be the largest integer s.t. \( 2^{b+\frac{1}{2}} \leq \frac{d}{3} \)

\[
Pr(\hat{d} \leq \frac{d}{3}) = Pr(t \leq b)
\]

\[
= Pr(Y_{b+1} = 0)
\]

\[
\leq \frac{2^{b+1}}{d} = \left(\frac{2^{b+\frac{1}{2}}}{d}\right)\sqrt{2} \leq \frac{\sqrt{2}}{3} < .48
\]

Therefore, \( Pr(\frac{d}{3} < d < 3\hat{d}) \geq .04 \)

Since \( Pr(\hat{d} \leq \frac{d}{3} \text{ or } \hat{d} \geq 3d) \leq 2 \times .48 = .96 \)

How to boost this prob. to \( \geq 1-\delta \)?
D. \( k = O(\log(1/\delta)) \) trials,
get outputs \( D_1, \ldots, D_k \)
Output \( D = \text{Median}(D_1, \ldots, D_k) \)

\[ Z_i = \begin{cases} 1 & \text{if } D_i < 3\delta \\ 0 & \text{otherwise} \end{cases} \]

If \( D \geq 3\delta \), then \( \geq \frac{k}{2} \) of the trials exceed \( \geq 3\delta \).

Let \( Z = \sum_{i=1}^{k} Z_i \)
\( \mathbb{E}[Z] \geq 0.52k \)

\( \Pr(D \geq 3\delta) \leq \Pr(Z < \frac{k}{2}) \)
\[ = \Pr(Z \leq \frac{k}{2} \cdot 0.52k - 0.02k) \]
\[ \leq \Pr(Z \leq (1-0.02)\mathbb{E}[Z]) \]
\[ \leq e^{-\frac{0.02 \cdot 0.52k^2}{3}} \leq \delta/2 \]
for \( k = c \log(1/\delta) \) with \( c \) big enough constant.
This gives a \((3, \delta)\)-approx.

Using \(O(\log n)\) bits per hash function

and \(O(\log \log n)\) bits for \(z\)

\[\Rightarrow O(\log(\frac{1}{\delta}) \log n)\] total bits.

Next class: \((\epsilon, \delta)\)-approx. for all \(\epsilon > 0\).