Random variables $X_1, \ldots, X_n$ over Ω.

Mutually independent if:

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

k-wise independent if:

$$\forall\, S \subset \{1, \ldots, n\} \text{ where } |S| \leq k,\ S = \{s_1, \ldots, s_k\},$$

$$\forall\, i_1, \ldots, i_k \in \Omega,$$

$$\Pr\left(\bigcap_{j=1}^{k} X_{s_j} = i_j\right) = \prod_{j=1}^{k} \Pr(X_{s_j} = i_j)$$

Pairwise independent is $k=2$, i.e.,

$$\forall\, j, \ell \in \{1, \ldots, n\},\ a, b \in \Omega,$$

$$\Pr(X_j = a, X_\ell = b) = \Pr(X_j = a) \times \Pr(X_\ell = b).$$

Lemma: Let $X_1, \ldots, X_n$ be pairwise independent,
& let $X = \sum_{i=1}^{n} X_i$

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

& if X_i's are binary/indicator random variables,
0-1

then $\text{Var}(X) \leq \sum_{i=1}^{n} E[X_i^2] = \sum_i E[X_i] = E[X]$.

Simple construction of pairwise independent random variables:

from ~~b~~ b mutually indp't. random bits,
generate $m = 2^b - 1$ pairwise indp't. random bits.

Let $X_1, \dots, X_b \in \{0,1\}$ be uniform, mutually indp't. random bits

Let $S_1, \dots, S_{2^b-1}$ be the nonempty subsets of $\{1, \dots, b\}$.

Set $Y_j = \bigoplus_{i \in S_j} X_i = \sum_{i \in S_j} X_i \bmod 2$

Note, $Y_j \in \{0,1\}$

__Lemma:__ The Y_j's are pairwise indp't.

__Proof:__ __Claim 1:__ $\Pr(Y_j = 1) = \Pr(Y_j = 0) = \frac{1}{2}$

Why? Let $S_j = \{z_1, \dots, z_\ell\} \subseteq \{1, \dots, b\}$

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

Reveal $X_{z_1}, \dots, X_{z_{\ell-1}}$. 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_ℓ and consider some $z \in S_j \setminus S_\ell$.

$$\Pr(Y_j = a, Y_\ell = b) = \Pr(Y_j = a \mid Y_\ell = b) \Pr(Y_\ell = b)$$

$$= \Pr(Y_j = a \mid Y_\ell = b) \cancel{\Pr(Y} \times \frac{1}{2}$$

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. $\frac{1}{2}$ $\cancel{z}\, X_z = 0$ & Y_j is the same

Therefore, $\Pr(Y_j = a \mid Y_\ell = b)$

$$= \Pr(Y_j = a \mid X_1, \ldots, X_b \setminus X_z)$$

$$= \frac{1}{2}$$

& thus, $\Pr(Y_j = a, Y_\ell = b) = \frac{1}{4}$.

More sophisticated construction:

For prime p, given a, b which are indep't & uniform over $\{0, 1, \dots, p-1\}$

then we construct $Y_1, \dots Y_{p-1}$ which are pairwise indept. & uniform over $\{0, 1, \dots, p-1\}$.

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

Lemma: The Y_i's are pairwise indep't.

Proof: First, Y_i is uniform over $\{0, 1, \dots, p-1\}$.

Why? By principle of deferred decisions again.

For any b & i & α in $\{0, \dots, p-1\}$,

$$\Pr(Y_i = \alpha) = \Pr(a + bi \equiv \alpha \bmod p)$$
$$= \Pr(a \equiv \alpha - bi \bmod p)$$
$$= \frac{1}{p}$$ Since there is a unique such a in $\{0, 1, \dots, p-1\}$.

Now consider $i, j \in \{0, \dots, p-1\}$, $i \neq j$ and $\alpha, \beta \in \{0, \dots, p-1\}$

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

$$Y_i = \alpha \iff a + ib \equiv \alpha \mod p$$
$$Y_j = \beta \iff a + jb \equiv \beta \mod p$$

Thus, $\alpha - \beta \equiv b(i-j) \mod p$

$$b \equiv \frac{\alpha - \beta}{i-j} \mod p$$

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, $\Pr(Y_i = \alpha, Y_j = \beta)$

$$= \Pr\left(b \equiv \frac{\alpha - \beta}{i-j} \mod p, a \equiv \alpha - \frac{i(\alpha-\beta)}{i-j} \mod p\right)$$
$$= \frac{1}{p^2}$$

$\boxtimes$

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 $S = \{s_1, \ldots, s_m\}$ for HUGE m.

where each $s_i \in \{0, 1, \ldots, n-1\}$

Let $f = (f_1, \ldots f_n)$ where

$f_i = |\{j : s_j = i\}| =$ ~~frequ.~~ # of occurrences of value i in S.

Let $D = F_0 = |\{i : f_i > 0\}| =$ # of distinct values in S.

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

aim for (ϵ, δ)-approx of D:

output $\hat{D}$ where $\Pr\left(\hat{D}(1-\epsilon) \leq D \leq (1+\epsilon)\hat{D}\right) \geq 1-\delta$

in ~~time poli~~ space $\text{Poly}\left(\log n, \frac{1}{\epsilon}, \log(1/\delta)\right)$

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 $\text{zeros}(k) \geq 1$ since it ends in 0.

AMS algorithm: (Diff. alg. than last class but same paper)

1. Choose a random hash function $h : [n] \to [n]$
   $\{0,1,\ldots,n-1\} \to \{0,1,\ldots,n-1\}$
   which is pairwise independent
   (if n is not prime, choose prime p where $n \leq p \leq 2n$ & think of n as p)
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 $(2^{z+\frac{1}{2}})$

Intuition for alg.:

$h(k)$ is uniformly random bitstring so prob. it has $\text{zeros}(h(k)) = \ell$ is prob. of last ℓ bits all $= 0$ which is prob. $2^{-\ell}$!

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 d to have $\text{zeros}(h(k)) = \log d$.

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

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

Analysis:

For $k \in \{0, \dots, n-1\}$, & integer $\ell \geq 0$,

let $X_{\ell,k} = \begin{cases} 1 & \text{if } \text{zeros}(h(k)) \geq \ell \\ 0 & \text{o/w} \end{cases}$

& let $Y_\ell = \sum_{k : f_k > 0} X_{\ell,k}$

Let t be the final value of z at end of algorithm.

$$t \geq \ell \iff Y_\ell > 0$$

$$t = \max_{k} \{\ell : Y_\ell > 0\} = \max \{\ell : \exists k, f_k > 0, \text{zeros}(h(k)) \geq \ell\}$$

this is same as:

$$t < \ell \iff Y_\ell = 0.$$

$$\begin{array}{c} \parallel \\ t \leq \ell - 1 \end{array}$$

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

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

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

$= 1/2^\ell$

$$\mathrm{Var}(Y_\ell) = \sum_{k: f_k > 0} \mathrm{Var}(X_{\ell,k}) \quad \text{since } X_{\ell,k} \,\&\, X_{\ell,k'} \text{ are pairwise indep't.}$$

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

$$= \sum_{k: f_k > 0} E[X_{\ell,k}]$$

$$= \frac{d}{2^\ell}$$

By Markov's ineq.,

$$\Pr(Y_\ell > 0) = \Pr(Y_\ell \geq 1)$$

$$\leq \frac{E[Y_\ell]}{1} = \frac{d}{2^\ell}$$

By Chebyshev's (can apply since pairwise indep't.)

$$\Pr(Y_\ell = 0) \leq \Pr\left(|Y_\ell - E[Y_\ell]| \geq \frac{d}{2^\ell}\right)$$

$$\leq \frac{\mathrm{Var}(Y_\ell)}{\left(\frac{d}{2^\ell}\right)^2} \leq \frac{2^\ell}{d}$$

Our goal is to output d.

Let's aim for $\hat{d}$ where $\frac{\hat{d}}{3} \le d \le 3\hat{d}$

So $\hat{d}$ is a 3-approx. of d.

Let a be the smallest integer s.t. $2^{a+\frac{1}{2}} \ge 3d$.

We want to show that the prob. that t = final value of Z is unlikely to be as large as a.

$$\Pr(\hat{d} \ge 3d) = \Pr(t \ge a)$$
$$= \Pr(Y_a > 0)$$
$$\le \frac{d}{2^a} = \frac{\sqrt{2}}{3} < .48 < \frac{1}{2}$$

Note, $2^{a+\frac{1}{2}} \ge 3d$

$$\frac{d}{2^a} \le \frac{\sqrt{2}}{3}$$

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. $2^{b+\frac{1}{2}} \leq \frac{d}{3}$

$$\Pr\left(\hat{d} \leq \frac{d}{3}\right) = \Pr(t \leq b)$$

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

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

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

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

How to boost this Prob. to $\geq 1-\delta$?

Do $k = O(\log(1/\delta))$ trials,

get outputs $D_1, \ldots, D_k$

Output $D = \text{Median}(D_1, \ldots, D_k)$

Let $Z_i = \begin{cases} 1 & \text{if } D_i < 3d \\ 0 & \text{o/w} \end{cases}$

If D is $\geq 3d$ then $\geq \frac{k}{2}$ of the trials exceed $\geq 3d$.

Let $Z = \sum_{i=1}^{k} Z_i$

$$E[Z] \geq .52k$$

$$\begin{aligned} \Pr(D \geq 3d) &\leq \Pr\left(Z < \frac{k}{2}\right) \\ &= \Pr(Z \leq .52k - .02k) \\ &\leq \Pr(Z \leq (1-.02)E[Z]) \\ &\leq e^{-(.02)^2 .52k/3} \leq \delta/2 \end{aligned}$$

for $k = c\log(\frac{1}{\delta})$ with c big enough constant.

This gives a $(3,\delta)$-approx.

using $O(\log n)$ bits per hash function

& $O(\log\log n)$ bits for z

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

Next class: $(1+\epsilon,\delta)$-approx. for all $\epsilon>0$.