Recap: Bloom Filters:

- Limited ops: membership query
- Allows for mistakes:
  - Only false positives; no false negatives.

Array \( T \) of \( M \) bits
- Initialized to 0

- \( k \) hash functions: \( h_1, h_2, \ldots, h_k: U \rightarrow [M] \)
  - Assume fully random

Adding: \( x \in S \)
- Set bits \( T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)] \) to 1
Membership query:
Check those k locations.
If all k set to 1, return positive.
Else negative.

False positive probability:
Understand prob of all k bits being set to 1.

Let \( p = \text{prob that a bit in T is not set} \)

\[
p = \left(1 - \frac{1}{M}\right)^{kN} = \left(1 - \frac{1}{M}\right)^{M \cdot \frac{kN}{M}}
\]

- \( N \) elements/object
- \( M \) locations/bits
- \( k \) hash function
  (Assump. fully random)
\[ p = e^{-\frac{KN}{M}} \quad \text{(using the limit \( t \to \infty \))} \]

Prob. of a particular bit being set = 1 - \( p \)

Prob. of false positive = \( (1 - p)^k \)

\[ = \left(1 - e^{-\frac{KN}{M}}\right)^k \]

Differentiate \( e \) set to 0. (Ex.)

\[ k = \frac{N}{M} \cdot \ln 2 \]

Plug in \( p \) for prob. of false pos.
Let $e = p_{false\ positives}$

$e = \left(\frac{1}{2}\right)^{\frac{MN}{1.44N \cdot \log(1/e)}}$

$M = 1.44N \cdot \log\left(\frac{1}{e}\right)$

E.g. (1% false pos. prob)

$M = 10N$ bits

$k = 7$
Load Balancing

$N$ jobs, $M$ machines, equal sized jobs.

Consider $N = M$

When we use hashing to allocate,

question of interest: maximally loaded machine

captured in "balls and bins" problems

$N$ balls, $N$ bins

Randomly throw balls to bins
Thm: The max-loaded bin $O\left(\frac{\log N}{\log \log N}\right)$ balls with prob. at least $1 - \frac{1}{N}$.

Proof: High level steps:
1. Prob. of any bin receiving $\frac{\log N}{\log \log N}$ balls 
   $\Rightarrow$ want: $\frac{1}{N}$ (for union bounding over bins)
2. Pr's of there being at least one bin with at least there many balls.
   $\Rightarrow$ want: $\frac{1}{N}$

| Union bound: $p(A \cup B \cup C \ldots) \leq p(A) + p(B) \ldots$ |
\[ p \text{ (bin i has at least k balls)} \leq \binom{N}{k} \left( \frac{1}{N} \right)^k \]

\[ = \frac{n!}{(N-k)! \cdot k!} \cdot \frac{1}{N^k} \]

\[ = \frac{N^k}{k!} \cdot \frac{1}{N^k} \]

\[ = \frac{1}{k!} \]

Stirling's approximation:

\[ k! \approx \sqrt{2\pi k} \left( \frac{k}{e} \right)^k \]

Choose \( k = O \left( \frac{\log N}{\log \log N} \right) \) gives desired result.
\[ k = O \left( \frac{\log N}{\log \log N} \right) - \text{wise indep suffices} \]

More relaxed hash families?

Need concentration bounds.

\underline{Concentration bounds:}

1. Markov's Inequality: \( X \) non-negative R.V.
   \[ M = \text{mean} \]
   \[ P \left( X > \alpha \right) \leq \frac{M}{\alpha} \]
2. Chebyshev's Inequality: $X \sim \text{R.V.}$

$m = \text{mean}$

$\sigma^2 = \text{variance}$

\[
P\left(\left| X - m \right| \geq \varepsilon \right) \leq \frac{\sigma^2}{\varepsilon^2}
\]

E.g.: 2-wise indep.
N. balls N. bins

Lemma: max. load over all bins is $O(\sqrt{N})$ w.p. at least $\frac{1}{2}$

E.g.: Consider bin $i$
$$L_i = \text{load on bin } i$$

(\# balls)

$$= \sum_{j=1}^{N} X_{ij} \quad \text{subject to} \quad X_{ij} \in \{0, 1\}$$

$$= \begin{cases} 1 & \text{ball } j \text{ lands in bin } i \\ 0 & \text{o.w.} \end{cases}$$

$$N_i = E(L_i) = \sum_{j=1}^{N} X_{ij} = 1 \quad \text{(Ex.)}$$

(2-wise indy)

$$\sigma_i^2 = \text{Var}(L_i) = \text{Var}\left( \sum_{j=1}^{N} X_{ij} \right)$$

$$= \sum_{j=1}^{N} \text{Var}(X_{ij}) \quad \text{(\# 2-wise indy)}$$

$$= N \cdot \left( \frac{1}{N} - \frac{1}{N^2} \right) = 1 - \frac{1}{N}$$
By Chebyshev,

\[ P \left( \left| x_i - \mu_i \right| > \sqrt{2N \cdot \sigma_i^2} \right) \leq \frac{\sigma_i^2}{2N \cdot \sigma_i^2} \]

union bound gives the result.

---

\( p \)-wise indep.

higher-moment Chebyshev:

\[ P \left( \left| x - \mu \right| > \varepsilon \right) \leq \frac{\varepsilon \left[ \left( x - \mu \right)^p \right]}{\varepsilon^p} \]
**Chernoff Bounds**

Any r.v. $X$, for any $t > 0$

$$P(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}$$

$$\implies P(X \geq a) \leq \min_{t > 0} \frac{E[e^{tX}]}{e^{ta}}$$

**Hoeffding Bound:** $X_i$'s independent r.v. taking values in $[0, 1]$

$$X = X_1 + X_2 + \ldots + X_n, \quad M = E[X]$$

$$P(X > M + \lambda) \leq e^{-\frac{\lambda^2}{2M + \lambda}}, \quad P(X < M - \lambda) \leq e^{-\frac{\lambda^2}{3M}}$$
Ex:- \( \lambda = c M \).
Compare to Markov & Chebyshev

Using Hoeffding in balls & bins:

\[
L_i = \sum_{j=1}^{n} X_{ij}
\]

(Ex)

\[
\lambda = c \log N
\]

Then: w.h.p., max load is \( \Omega \left( \frac{\log N}{\log \log N} \right) \)
power-2 - 2-choice :

$O(\log \log N)$