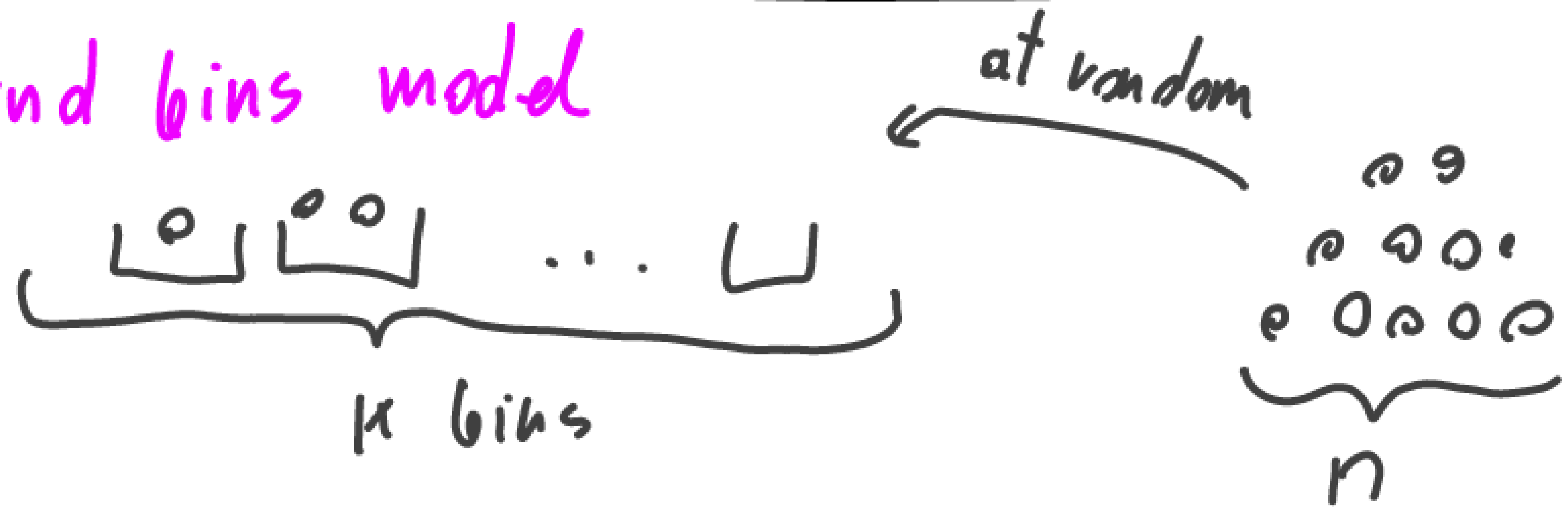
Balls and bins model, Hyperloglog

I) Balls and bins model



- when do we have a first collision: birthday paradox
  $n \approx \sqrt{k}$ to have prob of collision $> \frac{1}{2}$
- when all bins are non-empty: coupon collector problem $n \approx k \log(k)$ whp (exercise)
- what is the number of balls based only on number of empty bins: Linear Counting (1999)

II) Linear counting

1) Let V denote the number of empty bins after throwing n balls.

2) Note that $V = 1\!\!1_1 + 1\!\!1_2 + \ldots + 1\!\!1_k$; $1\!\!1_i = \begin{cases} 1, & i\text{-th bin empty} \\ 0, & \text{otherwise} \end{cases}$

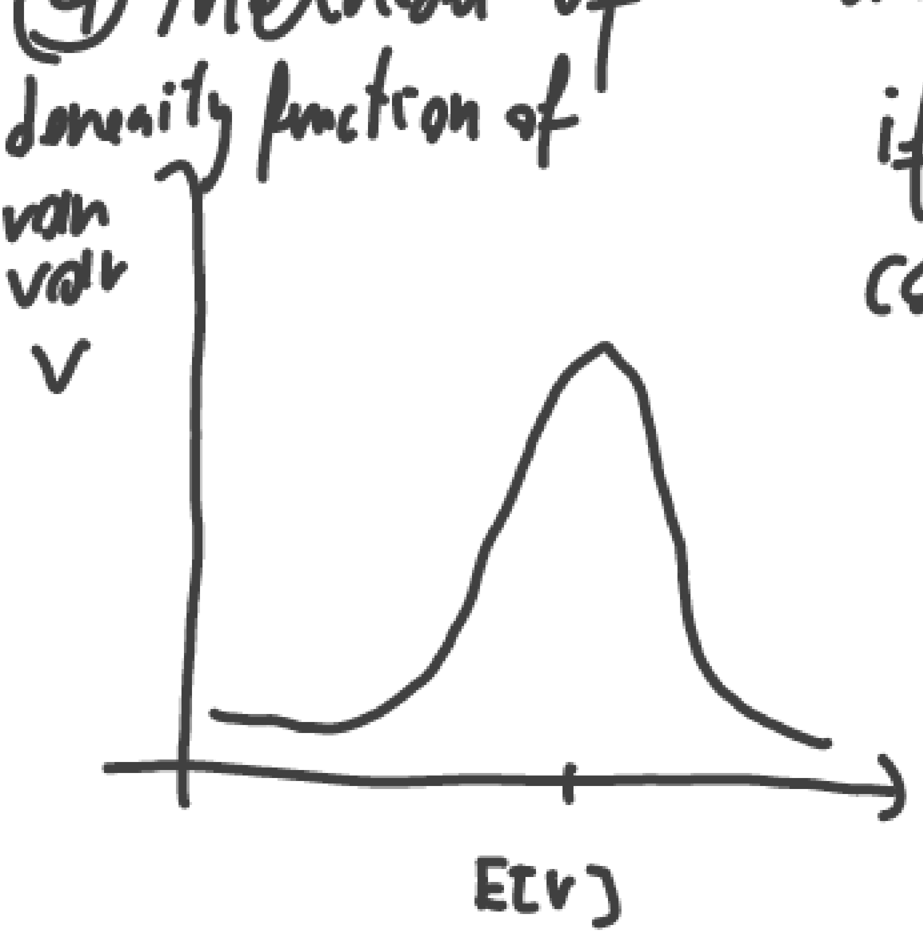
3) $\Pr[1\!\!1_i = 1] = \left(1 - \frac{1}{k}\right)^n$

$E[1\!\!1_i] = 1 \cdot \Pr[1\!\!1_i = 1] + 0 \cdot \Pr[1\!\!1_i = 0] = \left(1 - \frac{1}{k}\right)^n$

$E[V] = E[1\!\!1_1] + E[1\!\!1_2] + \ldots + E[1\!\!1_k] = k\left(1 - \frac{1}{k}\right)^n \rightarrow$

$\rightarrow n = \dfrac{\log\left(\frac{E[V]}{k}\right)}{\log\left(1 - \frac{1}{k}\right)}$

4) Method of moments: $\log\left(\frac{E[V]}{k}\right) \propto \log\left(\frac{V}{k}\right)$



if V is concentrated around its mean ($E[V]$) we can replace $E[V]$ with V.

$$\frac{1}{\ln(1-\frac{1}{k})} = -k + \frac{1}{2} + O\left(\frac{1}{k}\right)$$ // Wolfram-alpha : Series[f(k), {k, ∞, 2}]

yields $\underline{\hat{n} := -k \ln\left(\frac{V}{k}\right)}$ for $V > 0$

(III) Hyperloglog (memory : $O(\log\log(n))$)

History:

- 1977 : Morris Counter
- 1985 : probabilistic Counting
- 2003 : loglog
- 2007 : Hyperloglog
- 2016 : HLL ++

} Flajolet

} Google

Model:

- $\mathcal{M} = (S, m)$
- $h : S \to \{0,1\}^{32}$, $h(s) = h_1 h_2 \dots h_{32}$
- $g : \{0,1\}^{32} \to$ position of the first one from the left side " // $g(00101) = 3$
- note. $\Pr[h_1 = 0 \wedge h_2 = 0 \wedge \dots h_{i-1} = 0 \wedge h_i = 1] = \left(\frac{1}{2}\right)^i$, so $\Pr[g(h) = i] = \left(\frac{1}{2}\right)^i$
- note that pattern : $\underbrace{00..0}_{\log_2 n}1$ on average occurs once for n elements.

$1\dots \leftarrow \frac{1}{2}$

$01\dots \leftarrow \frac{1}{4}$

$001\dots \leftarrow \frac{1}{8}$

$\vdots$

$00..01 \leftarrow \left(\frac{1}{2}\right)^{\log_2 n} = \frac{1}{n}$

$\uparrow$

$\log_2 n$

- $S = \{s_1, s_2, \dots s_n\} \xrightarrow{h(\cdot)} H = \{h(s_1), h(s_2), \dots h(s_n)\} \xrightarrow{g(\cdot)} R = (r_1, r_2, \dots r_n) \xrightarrow{\max} M = \max(R) \approx \log_2(n)$
- $\hat{n} := 2^M \Rightarrow$ to store the value of M we need $\log_2 \log_2 n$ bits.
- note $\hat{n}$ has very big variance

• idea: lower the variance by taking the average of m experiments

$$Var[X] = a \rightarrow Var\left[\frac{X_1 + X_2 + \dots + X_m}{m}\right] = \frac{1}{m^2} m \cdot Var[X] = \frac{a}{m}$$

• in our case $\hat{n} := \frac{2^{M_1} + \dots + 2^{M_m}}{m}$, but first we do not have

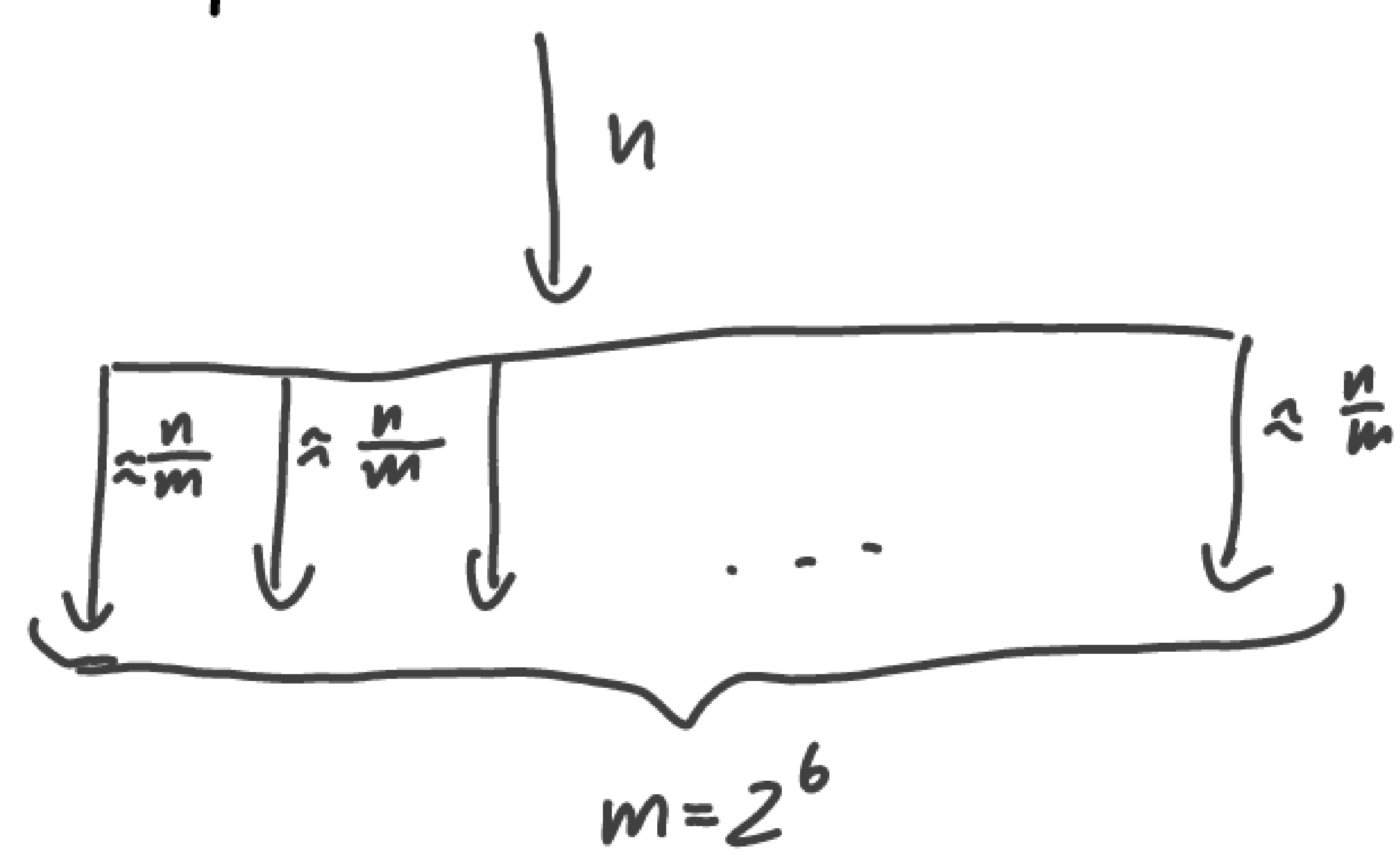
1) m independent hash functions

2) we have problem with outliers → the idea: use harmonic mean.

---

ad 1) Stochastic averaging – Simulate m hash functions based on a simple hash function

$$h(s) = \underbrace{h_1 h_2 \dots h_6}_{\text{number of experiment}} \underbrace{h_{6+1} \dots h_{32}}_{\text{here we use } \varrho \text{ to find first "one"}}$$



---

ad 2) harmonic mean

• $H_{AV}(x_1, x_2, \dots x_m) = \frac{m}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_m}}$, $x_i = 2^{M_i} \approx \frac{n}{m}$

• $\frac{m}{\frac{1}{2^{M_1}} + \dots + \frac{1}{2^{M_m}}} \approx \frac{n}{m} \rightarrow \hat{n}_{HLL} := \alpha_m \cdot m \cdot \left(\sum_{i=1}^{m} 2^{-M_i}\right)^{-1}$

constant $\alpha_m = \left(m \int_0^\infty \left(\log_2 \frac{2+u}{1+u}\right)^m du\right)^{-1}$

Hyperloglog (M, m, h)

we assume $m = 2^6$

① $M \leftarrow (0, 0 \dots 0)$ // length m

② for each $e \in M$

  $h \leftarrow h(e)$

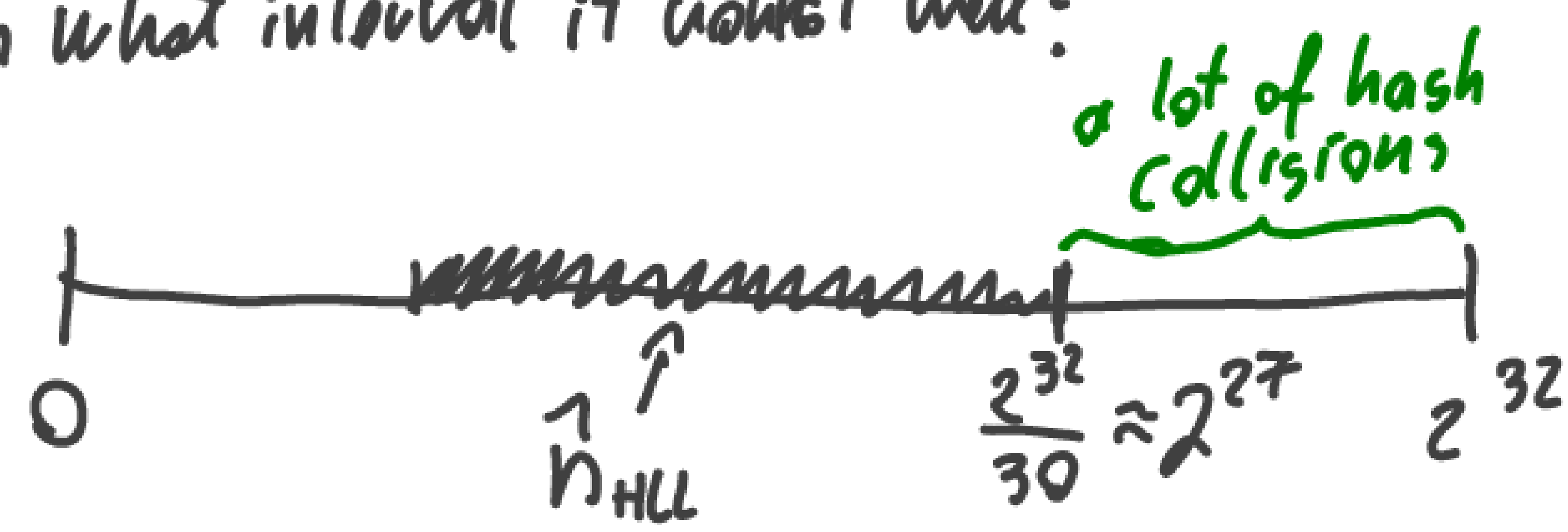
  $j \leftarrow \langle h_1 h_2 \dots h_6 \rangle_2 + 1$ // +1 because we do not want to start from 0.

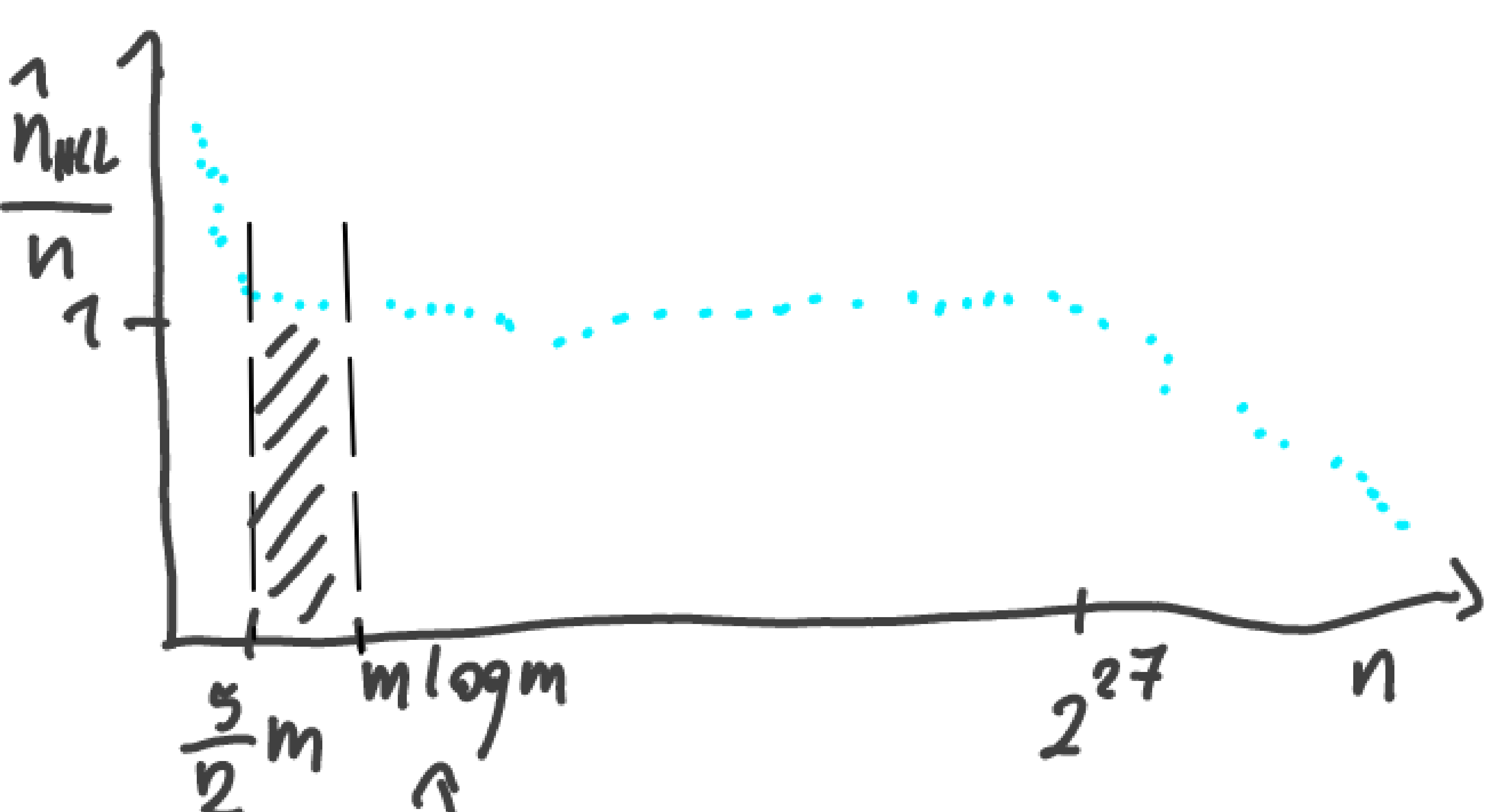
  $w \leftarrow h_{6+1} \dots h_{32}$

  $M_j \leftarrow \max(M_j, g(w))$

③ $\hat{n}_{HLL} := \alpha_m \cdot m^2 \cdot \left(\sum_{i=1}^{m} 2^{-M_i} \right)^{-1}$

On what interval it works well?





from coupon collector problem

but HLL works well in that region

---

④ Corrections

if $\frac{\hat{n}_{HLL}}{n} < \frac{5}{2} m$:

  $V \leftarrow |\{ i : M_i = 0 \}|$

  if $V \neq 0$:

    $\hat{n}_{HLL} \leftarrow -m \log\left(\frac{V}{m}\right)$

balls and bins model

(5) if $\frac{\hat{n}_{HLL}}{n} > \frac{2^{32}}{30}$

$H \leftarrow 2^{32}$

$\hat{n}_{HLL} \leftarrow -H \log\left(\frac{H - \hat{n}_{HLL}}{H}\right)$

ad (5)

- $H = 2^{32}$ hashvalues $\equiv$ bins
- elements $\equiv$ balls

2^{32} hashes

- $\hat{n}_{HLL}$ actually estimates the number of occupied hash values
- $H - \hat{n}_{HLL} \approx$ free hash values

! so in (5) we switch from estimating non-empty to empty bins. !