

maximal load: $O(\log n / \log \log n)$ — aim

Space required: $O(\log^2 n / \log \log n)$

Construction

each function described in $O(\log n / \log \log n)$

evaluated in $O(\log n / \log \log n)$

IDEA: concatenate output of $O(\log \log n)$ functions which are gradually more independent. Each function f is described using d functions h_1, \dots, h_d

$$f(x) = \underbrace{h_1(x)}_{\substack{\uparrow \\ \text{binary string}}} \circ h_2(x) \circ \dots \circ \underbrace{h_d(x)}_{\substack{\uparrow \\ \text{concatenate}}}$$

Gradually more independent

$$\begin{array}{c} h_1: O(1)\text{-wise indep.} \\ h_2: O(h_1)\text{-wise indep.} \\ \dots \\ h_d: O(\log n / \log \log n)\text{-wise indep.} \end{array}$$

↓
k wise
k increase

Note: ① output length decreases at the same time

$$h_1: \sum \log n \Rightarrow h_d: O(\log \log n)$$

② each of h_1, \dots, h_d could be described/computed in $O(\log n)$ bits/time

Definitions

$$[n] \Rightarrow \{1, \dots, n\}$$

$U_n \Rightarrow$ uniform distribution over set $\{0, 1\}^n$

$x \in X \Rightarrow$ sample x from X for a r.v. X .

$x \in S \Rightarrow$ sample x uniformly from finite set S .

$$SD(X, Y) \Rightarrow \text{statistical distance between two r.v. over finite domain } \Sigma \\ = \frac{1}{2} \sum_{\omega \in \Sigma} |\Pr[X=\omega] - \Pr[Y=\omega]|$$

$x \circ y \Rightarrow$ concatenate x, y bit string

unit cost RAM model

\Rightarrow elements are taken from a universe of size n and each element can be stored using $c = O(\log n)$ bits.

k -wise f -dependent

For a family of $f: [n] \rightarrow [v]$, we're k - f -dependent iff

$$SD(X, Y) \leq \delta$$

where $X = \text{distribution}(f(x_1), f(x_2), \dots, f(x_k))$ for any distinct $x_1, \dots, x_k \in [n]$
 $Y = \text{uniform distribution over } [v]^k$

Describe Space: $O(k \max\{\log n, \log v\})$ bits

Evaluation Time: $O(k)$

ϵ -biased Distribution [M/N 93]

r.v.s X_1, X_2, \dots, X_n over $\{0, 1\}$ is ϵ -biased if for any $S \neq \emptyset \subseteq [n]$,

$$|\Pr[\bigoplus_{i \in S} X_i = 1] - \Pr[\bigoplus_{i \in S} X_i = 0]| \leq \epsilon$$

\uparrow
XOR operation

[Ahuja 92, Sec. 5] constructs an ϵ -biased distribution over $\{0, 1\}^n$ where each point could be specified using $O(\log(n/\epsilon))$ bits where each bit could be calculated using $O(\log(n/\epsilon))$ times.

In RAM, t bits could be calculated in time $O(\log(n/\epsilon)t)$

may need proof

min-entropy

for a r.v. X , its min-entropy is

$$H_{\infty}(X) = -\log(\max_x \Pr[X=x])$$

negative log of max probability of $X=x$.

k-source

k-source is a r.v. X with its min-entropy $H_{\infty}(X) \geq k$.

(T, k)-block source

r.v. $X = (X_1, \dots, X_T)$, for any $i \in [T]$

$$H_{\infty}(X_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \geq k$$

(T, k, ε)-block source

r.v. $X = (X_1, \dots, X_T)$, for any $i \in [T]$

$$\Pr[(H_{\infty}(X_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \geq k)] \geq 1 - \epsilon$$

$$(x_1, \dots, x_{i-1}) \leftarrow (X_1, \dots, X_{i-1})$$

Pr of X is (T, k) -block source $\geq 1 - \epsilon$

7 lemmas

Corollary 2.1

Fact

for any k an ϵ -biased distribution is also k -wise δ -dependent

$$\text{for } \delta = \epsilon 2^{k/2}$$

For any $u, v, v=2^i$, there exists a family of k -wise δ -dependent function $f:[u] \rightarrow [v]$ described in $O(\log u + k \log v + \log(1/\delta))$ bits and calculated in $O(\log u + k \log v + \log(1/\delta))$ in RAM.

check AGH92 for proof.

Corollary 2.2

Let $X_1, \dots, X_n \in \{0, 1\}$ be $2k$ -wise δ -dependent r.v. for $k \in \mathbb{N}, 0 \leq \delta < 1$,

let $X = \sum X_i$ and $\mu = E[X]$. Then, for any $t > 0$,

$$\Pr[|X - \mu| > t] \leq 2 \left(\frac{2^{nk}}{t^2} \right)^k + \delta \left(\frac{n}{t} \right)^{2k}.$$

How X change from μ . proceed on paper using Markov Ineq.

(Lemma 2.2 from [BR94])

$$\text{proof: } \Pr(|X - \mu| > t)$$

$$= \Pr(|X - \mu|^{2k} > t^{2k})$$

$$\leq \frac{E[|X - \mu|^{2k}]}{t^{2k}} \quad \xleftarrow{\text{markov inequality}} \quad \Pr(X \geq a) \leq \frac{E[X]}{a}.$$

$$\text{consider } E[X - \mu] = \sum_{i \in [n]} E[X_i - \mu_i]$$



$$E[X - E[X]] = \sum_{i=1}^n E[X_i - E[X_i]] = \sum_{i=1}^n E[X_i - \mu_i]$$

$$\text{similarly } E[(X - \mu)^2] = E[(X - E[X])^2]$$

$$= \sum_{i=1}^n \sum_{j=1}^n E[(X_{ij} - \mu_{ij})^2]$$

$$= \sum_{i_1, i_2, \dots, i_{2k} \in [n]} E \left[\prod_{j=1}^{2k} (X_{ij} - M_{ij}) \right]$$

\therefore we know that $E[(x-\mu)^{2k}]$

$$= \sum_{i_1, i_2, \dots, i_{2k} \in [n]} E \left[\prod_{j=1}^{2k} (X_{ij} - M_{ij}) \right]$$

let \hat{X} be estimator value of X

$$= \sum_{i_1, \dots, i_{2k} \in [n]} E \left[\prod_{j=1}^{2k} (\hat{X}_{ij} - M_{ij}) \right]$$

$\because 0 \leq \delta < 1, n, k > 0$

$$\therefore \leq \sum_{i_1, \dots, i_{2k} \in [n]} E \left[\prod_{j=1}^{2k} (\hat{X}_{ij} - M_{ij}) \right] + \delta n^{2k}.$$

$$\leq \sum_{i_1, \dots, i_{2k} \in [n]} E \left[\prod_{j=1}^{2k} (\hat{X}_{ij} - M_i) \right] + \delta n^{2k} \quad \text{where } M_i = E(X_i)$$

$$= \frac{E[(\hat{X} - \mu)^{2k}]}{t^{2k}} + \delta \left(\frac{n}{t}\right)^{2k}.$$

$$\text{consider } \Pr((\hat{X} - \mu) \geq t) \leq \frac{E[(\hat{X} - \mu)^t]}{t^t}$$

$$\text{From BR94: } \Pr((\hat{X} - \mu) \geq t) \leq \frac{E[(\hat{X} - \mu)^t]}{t^t}$$

$$\leq C \cdot \left(\frac{nt}{A^t} \right)^{\frac{t}{2}} \quad \text{where } C \leq 1.0004$$

$$\therefore \Pr((\hat{X} - \mu) \geq t) \leq \frac{E[(\hat{X} - \mu)^{2k}]}{t^{2k}} \leq C_{2k} \cdot \left(\frac{n \cdot 2k}{t^2} \right)^{\frac{2k}{2}}$$

$$\leq 2 \cdot \left(\frac{2nk}{t^2} \right)^k$$

$$\therefore \leq 2 \left(\frac{nk}{t^2} \right)^k + \delta \left(\frac{n}{t} \right)^{2k} \text{ as required.}$$

Lemma 2.4 [GW97] \leftarrow may not be used by our construction.
 \leftarrow may need a proof cut some from 4.1/4.2 main results

r.v. $x_1 \in [0,1]^n$, $x_2 \in [0,1]^{n_2}$, $H_\infty(x_1, x_2) \geq h_1 + h_2 - \Delta$

① $H_\infty(x_1) \geq h_1 - \Delta$

② for any $\epsilon > 0$,

$$\Pr_{x_1 \in X_1} [H_\infty(x_2 | x_1 = x_1) < h_2 - \Delta - \log(1/\epsilon)] < \epsilon$$

Corollary 2.5

Any $X = (X_1, \dots, X_T)$ over $(\{0,1\}^n)^T$, $H_\infty X \geq Tn - \Delta$
 is a $(T, n-\Delta - \log(1/\varepsilon), \varepsilon)$ -block source for any $\varepsilon > 0$.

↑ def of (T, k, ε) -block source
 and Lemma 2.4.

Theorem 3.1

any constant $c > 0$, integer n , $u = \text{poly}(n)$, there exists a family F of $f: [u] \rightarrow [u]$ that:

- ① describe using $O(\log n \log \log n)$ bits
- ② $f(x)$ can be computed in $O(\log n \log \log n)$ time
- ③ $\exists \gamma > 0$ such that for any $S \subseteq [u]$

$$\Pr_{f \in F} \left[\max_{i \in [u]} |f^{-1}(i) \cap S| \leq \frac{\gamma \log n}{\log \log n} \right] > 1 - \frac{1}{n^c}$$

\uparrow
 maximum load = $O(\frac{\gamma \log n}{\log \log n})$ with high probability.

Construction

$$\text{assume } n = 2^k$$

let $d = O(\log \log n)$, and for every $i \in [d]$, let H_i be a family of k_i -wise f -dependent functions $[u] \rightarrow \{0, 1\}^{L_i}$, where:

$$① n_0 = n \quad n_v := \frac{n}{2^{L_v}} \text{ for every } v \in [d]$$

$$n_1 = \frac{n}{2^{L_1}} \quad n_2 = \frac{n_1}{2^{L_2}} = \frac{n}{2^{L_1+L_2}} \dots n_d = \frac{n}{2^{\sum_{v=1}^{d-1} L_v}}$$

$$② L_i = \left\lfloor \frac{\log n_{i-1}}{4} \right\rfloor \text{ for every } i \in [d-1], \quad \text{load} = \log n - \sum_{v=1}^{d-1} L_v$$

$$L_i = \left\lfloor \frac{\log n_{i-1}}{4} \right\rfloor = \left\lfloor \frac{\log n - \sum_{v=1}^{i-1} L_v}{4} \right\rfloor \text{ for } i \in [d-1]$$

$$\text{load} = \log n - \sum_{i=1}^{d-1} L_i$$

$$③ k_i L_i = O(\log n) \text{ for every } i \in [d-1], \quad k_d = O(\log n / \log \log n)$$

$$k_i L_i = k_i \left\lfloor \frac{\log n - \sum_{v=1}^{i-1} L_v}{4} \right\rfloor$$

$$\textcircled{4} \quad f = \text{poly}(\frac{1}{n})$$

By corollary 2.1, we have a family H_i :

read detail.

h_i is represented using $O(\log n + k_i l_i + \log \frac{1}{\delta}) = \log n$ space/time

we defined our function family F as

$$f(x) = h_1(x) \circ h_2(x) \circ \dots \circ h_d(x)$$

for $h_i \in H_i$

Given the sets of balls $S \subseteq [n]$ of size n .

We visualize the construction as a tree of $d+1$ layers.

layer 0: has 1 bin with all balls, expected load $n_0 = n$ $\frac{n_0}{2^{0+0}} = \frac{n_0}{2^0 \cdot 1}$

layer 1: has 2^{l_1} bins, each bin has $n_1 = \frac{n_0}{2^{l_1}}$ $= \frac{n_0}{2^{1+0}}$

layer 2: has $2^{l_1+l_2}$ meaning each bin in layer 1 is split into 2^{l_2} bins

total of $2^{l_1} \cdot 2^{l_2} = 2^{l_1+l_2}$ bins, expected load $n_2 = \frac{\# \text{ of balls}}{\# \text{ of bins}} = \frac{n_0}{2^{l_1+l_2}}$

$$= \frac{\text{expected load in layer 1}}{\# \text{ of bins split into}} = \frac{n_1}{2^{l_2}}$$

layer i : $2^{\sum_{j=1}^i l_j}$ bins, expected load $n_i = \frac{n_{i-1}}{2^{l_i}}$

Lemma 3.2

For any $i = \{0, \dots, d-2\}$, $\alpha = \Delta(1/\log \log n)$, $0 < \alpha_i < 1$ and set $S_i \subseteq [n]$ of size at most $(1+\alpha_i)n_i$,

$$\Pr_{h_{i+1} \in H_{i+1}} \left[\max_{y \in [0, 1]^{l_{i+1}}} |h_{i+1}^{-1}(y) \cap S_i| \leq (1+\alpha)(1+\alpha_i)n_{i+1} \right] \geq 1 - \frac{1}{n^{c+1}}$$

number of balls in any bin of layer $i+1 \leq (1 + \Delta(1/\log \log n))(1 + \alpha_i)n_{i+1}$

Application

with high prob.

① storing elements using linear probing

② augmenting k -wise independence function using our construction

without affect the space/time requirement

read some evidence or proof.

Proof For Lemma 3.2

- Fix $y \in \{0, 1\}^{L_{i+1}}$, let $\chi = |\tilde{h}_{i+1}(y) \cap S_i|$

- assume WLOG, $|S_i| \geq \lfloor c(1+\alpha_i)n_i \rfloor$ or we could add dummy elements to enlarge S_i .

- Then, χ = sum of $|S_i|$ indicator random variables that are k_{i+1} -wise δ -dependent.

$$\text{new } \left\{ \begin{array}{l} X_j = 1 \text{ if element } j \in S_i \text{ is hashed into } h_{i+1} \\ X = \sum_j X_j \end{array} \right. \quad E[X] = \sum_j E[X_j] = \sum_j \frac{|S_i|}{2^{L_{i+1}}} = \frac{|S_i|}{2^{L_{i+1}}} \quad \begin{array}{l} \text{we could treat } S_i \text{ as elements in one bin from last layer} \\ \downarrow \\ \text{if uniform, total } 2^{L_{i+1}} \text{ bins for each set } S_i \end{array}$$

- Since χ = sum of $|S_i|$ k_{i+1} -wise δ -dependent r.v., we could apply Lemma 2.2:

$$k = \sum_{j=1}^{k_{i+1}} \mu = E[\chi] = |S_i|/2^{L_{i+1}} \Rightarrow \Pr[\chi > (1+\alpha_i)\mu] \leq 2 \left(\frac{|S_i| k_{i+1}}{\alpha_i \mu^2} \right)^{k_{i+1}/2} + \delta \left(\frac{|S_i|}{\alpha_i \mu} \right)^{k_{i+1}/2}$$

*replace μ
by $|S_i|/2^{L_{i+1}}$*

$$= 2 \left(\frac{|S_i| k_{i+1}}{|S_i|^2 / 2^{2L_{i+1}} \alpha^2} \right)^{k_{i+1}/2} + \delta \left(\frac{|S_i|}{|S_i|/2^{L_{i+1}} \alpha} \right)^{k_{i+1}/2}$$

$$= 2 \left(\frac{2^{2L_{i+1}} k_{i+1}}{|S_i| \alpha^2} \right)^{k_{i+1}/2} + \delta \left(\frac{2^{L_{i+1}}}{\alpha} \right)^{k_{i+1}/2}$$

- Now, we will upper bound each section.

$$1) \text{ Notice that, in our construction, } \textcircled{1} \quad l_{i+1} = \left\lfloor \frac{\log n_i}{4} \right\rfloor \leq \frac{\log(n_i)}{4}$$

$$\text{and } \textcircled{2} \quad |S_i| \geq (1+\alpha_i)n_i - 1 \geq n_i \quad \textcircled{3} \quad \alpha = \sqrt{2} c/\log \log n$$

$$2 \left(\frac{2^{2L_{i+1}} k_{i+1}}{\alpha^2 |S_i|} \right)^{k_{i+1}/2} \leq 2 \left(\frac{\frac{\log(n_i)}{2} k_{i+1}}{\alpha^2 |S_i|} \right)^{k_{i+1}/2} \quad \text{since } \textcircled{1}$$

$$= 2 \left(\frac{\frac{\log(n_i)}{2} k_{i+1}}{\alpha^2 |S_i|} \right)^{k_{i+1}/2}$$

$$\leq 2 \left(\frac{\sqrt{n_i} k_{i+1}}{\alpha^2 n_i} \right)^{k_{i+1}/2} \quad \text{since } \textcircled{2} \Rightarrow \frac{1}{|S_i|} < \frac{1}{n_i}$$

$$= 2 \left(\frac{k_{i+1}}{\alpha^2 \sqrt{n_i}} \right)^{k_{i+1}/2}$$

$$\leq 2 \left(\frac{k_{i+1}}{\alpha^2 2^{2L_{i+1}}} \right)^{k_{i+1}/2} \quad \text{since } 2^{2L_{i+1}} \leq 2^{\frac{\log(n_i)}{2}} = \sqrt{n_i}$$

$$= 2 \left(\frac{k_{i+1}}{\alpha^{k_{i+1}} 2^{L_{i+1} k_{i+1}}} \right)^{k_{i+1}/2} \quad \Rightarrow \frac{1}{\sqrt{n_i}} \leq \frac{1}{2^{2L_{i+1}}}$$

$$= 2 \left(\frac{2^{k_{i+1}}}{\alpha^{k_{i+1}} 2^{\log n_i}} \right) \text{ by construction.}$$

$$= 2 \left(\frac{2^{k_{i+1}}}{\alpha^{k_{i+1}} 2^{\log n_i}} \right) \text{ set } L_{k_{i+1}} = \log n_i$$

$$= 2 \left(\frac{2^{k_{i+1}}}{\alpha^{k_{i+1}}} \right)^{k_{i+1}/2} \cdot \frac{1}{n^c}$$

$$= 2 \left(\frac{2^{k_{i+1}}}{\sqrt{n} \cdot 2^{k_{i+1}}} \right)^{k_{i+1}/2} \cdot \frac{1}{n^c}$$

$$= 2 \left(\frac{1}{n^2} \right)^{\frac{2c \log n_i}{\log n_i}} \cdot \frac{1}{n^c}$$

$$\leq 2 \frac{1}{n^2} \cdot \frac{1}{n^c} \quad \text{since } \frac{\log n_i}{\log n_i} > 1$$

$$2 \left(\frac{2^{2^{k_{i+1}} k_{i+1}}}{\alpha^{2^{k_{i+1}}}} \right)^{k_{i+1}/2} \leq 2 \frac{1}{n^{c+2}} \quad \Leftarrow \text{result.}$$

2) Notice that $f = \text{poly}(\frac{1}{n})$, then

$$\begin{aligned} f\left(\frac{2^{k_{i+1}}}{\alpha}\right) &= f \frac{2^{k_{i+1}}}{\alpha^{k_{i+1}}} \\ &= f \frac{2^{\log n_i^c}}{\alpha^{k_{i+1}}} \end{aligned}$$

$$\leq f n^c$$

$$\leq \frac{1}{2^{n^{c+2}}} \cdot n^c$$

$$f = \frac{1}{2^{n^{c+2}}}$$

$$\leq \frac{1}{2^{n^{c+2}}}$$

Thus, we get

$$\begin{aligned} \Pr[X > c(1+\alpha)(1+\alpha_i)n_{i+1}] &= \Pr[X > c(1+\alpha)(1+\alpha_i)\frac{n_i}{2^{k_{i+1}}}] \text{ since } n_{i+1} = \frac{n_i}{2^{k_{i+1}}} \\ &\leq \Pr[X > c(1+\alpha)\frac{|S_i|}{2^{k_{i+1}}}] \text{ since } |S_i| \leq (1+\alpha_i)n_i \\ &= \Pr[X > c(1+\alpha)\mu] \text{ since } \mu = \frac{|S_i|}{2^{k_{i+1}}} \\ &\leq \frac{1}{2^{n^{c+2}}} + \frac{1}{2^{n^{c+2}}} \quad \text{by upper bound.} \\ &= \frac{1}{n^{c+2}} \end{aligned}$$

We have at most n different y since $y = \{0, 1\}^{L_{i+1}} \Rightarrow 2^{L_{i+1}} \leq 2^{\log n_i - 1} \leq 2^{\log n_i} = n$.

\Rightarrow union bound over y and subtract from 1

$$\frac{1}{n^{c+1}} \rightarrow \frac{1}{1 - \frac{1}{n^{c+1}}}$$

Th

Proof For Theorem 3.1

Description length / Evaluation time

i th layer $\Rightarrow h_i$ is k_i -wise f -dependent function: $[n] \rightarrow \{0,1\}^{L_i}$

$$v = 2^{L_i}, w = \text{poly}(n) f = \text{poly}\left(\frac{1}{n}\right)$$

By Corollary 2.1, we know there exists a family of k_i -wise f -dependent functions for:

$$\begin{aligned} \text{space} &= O(\log n + k \log v + \log(1/\delta)) \\ &= O(\log n + k_i \log_2^{L_i} \log\left(\frac{1}{\text{poly}(n)}\right)) \\ &= O(\log n + k_i L_i \log\left(\frac{1}{\delta}\right)) \\ &= O(\log n) \end{aligned}$$

$$\text{time} = O(\log n + k \log v + \log(1/\delta)) = O(\log n) \text{ also.}$$

Maximum load

- Fix a set $S \subseteq [n]$ of size n . We inductively argue that:

for every level $i \in \{0, \dots, d-1\}$, with probability $\geq 1 - \frac{i}{n^{c+1}}$,

the maximum load in level i is at most $(1+\alpha)^i n_i$ per bin.

$$\text{for } \alpha = \sqrt{C \log(n)}$$

Base Case $i=0$

$$\text{single bin with } n_0 = n = (1+\alpha)^0 n = (1+\alpha)^0 n_0$$



Inductive Hypothesis

claim holds for level i .

Inductive Case

Lemma 3.2 with $(1+\alpha_i) = (1+\alpha)^i$: number of balls in any bin of layer $i+1$

$$\leq (1+\alpha)^{i+1} n_{i+1} \text{ with prob } \geq 1 - \frac{i}{n^{c+1}}. \text{ Union bound over all } n_i \leq n \text{ bins,}$$

we could know that with prob $\geq 1 - \frac{i}{n^{c+1}} \geq 1 - \frac{i+1}{n^{c+1}}$, the maximum

Load in level $i+1$ is $(1+\alpha)^{i+1} n_{i+1}$.

Another way:

$$\begin{aligned} \text{Prob(number of balls in one bin of layer } i+1 > (1+\alpha)^{i+1} n_{i+1}) &\leq \frac{1}{n^{c+1}} \\ \text{Union bound} \Rightarrow \text{Prob(there exists a bin of layer } i+1 > (1+\alpha)^{i+1} n_{i+1}) &\leq \frac{n}{n^{c+1}} \\ \Rightarrow \text{Prob(there doesn't exist a bin of layer } i+1 \leq (1+\alpha)^{i+1} n_{i+1}) &\geq 1 - \frac{n}{n^{c+1}} \\ &\geq 1 - \frac{i+1}{n^{c+1}} \\ \Rightarrow \text{Prob(maximum load of layer } i+1 = (1+\alpha)^{i+1} n_{i+1}) &\geq 1 - \frac{i+1}{n^{c+1}} \end{aligned}$$

Then, claim holds in inductive case.

- Now, we want to upper bound n_{d-1} , the expected load in layer $d-1$.

By the induction claim, we know, with probability at least $1 - (d-1)/n^{c+1}$,

the maximum load in level $d-1$ is $(1+\alpha)^{d-1} n_{d-1} \leq 2 n_{d-1}$ for $\alpha = O(\log \log n)$.

For every $i \in [d-1]$,

$$\begin{aligned} l_i &\geq (\log n_{i-1})/4 - 1 \\ \Rightarrow n_i &= n_{i-1}/l_i \leq 2 n_{i-1}^{3/4} \quad \text{since } 2^{-l_i} \leq \frac{1}{\log n_{i-1}/4} = \frac{1}{n_{i-1}^{3/4}} \\ \Rightarrow n_{i-1} &\leq 2^{3/4} n_{i-2} \Rightarrow n_i \leq 2^{3/4} (2 n_{i-2})^{3/4} = 2^{1+3/4} n_{i-2}^{(3/4)^2} \\ \Rightarrow n_{i-1} &\leq 2^{\sum_{j=0}^{i-1} (3/4)^j} n^{(3/4)^i} \leq 2^4 n^{(3/4)^i} = 16 n^{(3/4)^i} \end{aligned}$$

Thus, for an appropriate choice of $\alpha = O(\log \log n)$ it holds

$$n_{d-1} \leq \log n$$

For example, $\alpha = \log_{3/4}(\frac{\log \frac{\log n}{16}}{\log n}) \in O(\log \log n)$

$$\begin{aligned} n_{d-1} &\leq 16 n^{(3/4)^d} = 16 n^{(3/4)^d \log_{3/4}(\frac{\log \frac{\log n}{16}}{\log n})} \\ &= 16 n^{\frac{\log \frac{\log n}{16}}{\log n}} \\ &= 16 n^{\log_n(\frac{\log n}{16})} \\ &= 16 \frac{\log n}{16} \\ &= \log n. \end{aligned}$$

Also, by definition of n_i , we know $n_i = \frac{n}{2^{l_i}}$, $\alpha = \log n - \sum_{i=1}^{d-1} l_i = \log n_{d-1}$

Thus, we know that, with $\text{Prob} \geq 1 - (cd-1)/n^{c+1}$,
 the maximum load on level $d-1$ vs $(1+d)^{d-1} n_{d-1} \leq 2 n_{d-1} \leq 2 \log n$.
 These elements are hashed into n_{d-1} bins using the function h_d which is
 k_d -wise f -dependent, where $k_d = \lceil 2c \log n / \log \log n \rceil$. Therefore, the
 probability that any $t = \gamma \log n / \log \log n < k_d$ elements from level $d-1$ are
 hashed into any specific bin in level d is at most.

$$\binom{2n_{d-1}}{t} \left(\left(\frac{1}{n_{d-1}} \right)^t + \delta \right) \leq \left(\frac{2n_{d-1}}{t} \right)^t \left(\left(\frac{1}{n_{d-1}} \right)^t + \delta \right)$$

\uparrow call comb of t
 \uparrow in bin
 within $2n_{d-1}$

$$= \left(\frac{2e}{t} \right)^t + \delta \left(\frac{2n_{d-1}}{t} \right)^t$$

$$= \left(\frac{2e \log n}{\gamma \log \log n} \right)^{\frac{2n_{d-1}}{\log \log n}} + \delta \left(\frac{2n_{d-1} \log n}{\gamma \log n} \right)^{\frac{2n_{d-1}}{\log \log n}}$$

Since $n_{d-1} < \log n \rightarrow$

$$\leq \left(\frac{2e \log n}{\gamma \log n} \right)^{\frac{n_{d-1}}{\log \log n}} + \delta \left(\frac{2e \log n}{\gamma} \right)^{\frac{n_{d-1}}{\log \log n}}$$

detail (?) \rightarrow

$$\leq \frac{1}{2n^{c+3}} + \frac{1}{2n^{c+3}}$$

$$= \frac{1}{n^{c+3}}$$

for $t = \frac{\gamma \log n}{\log \log n}$ and $\delta = \text{poly}(\frac{1}{n})$

This holds for any pair of bins in level d and $d-1$, which over them
 implies that:

$$\Pr(\text{a bin with more than } t \text{ elements}) \leq \frac{1}{n^{c+1}} \text{ since } 2^{L_d} \leq 2^{L_{d-1}} \leq n$$

This implies that

Union bound $\Rightarrow \Pr(\text{there exists a bin with more than } t \text{ elements}) \leq \frac{n}{n^{c+1}}$

$$\Pr(\text{there doesn't exist a bin with more than } t \text{ elements}) \geq 1 - \frac{n}{n^{c+1}} \geq 1 - \frac{c}{n^{c+1}}$$

$$\Rightarrow \Pr(\text{maximum load is } t) \geq 1 - \frac{c}{n^{c+1}} > 1 - \frac{1}{n^c}$$

B)

[BR94] Lemma 2.2

Let $t > 4$ be an even integer. Suppose X_1, \dots, X_n are t -wise independent r.v. taking values in $[0, 1]$. Let $X = X_1 + X_2 + \dots + X_n$ and $\mu = E[X]$.

Let $A > 0$. Then,

$$\Pr[|X - \mu| \geq A] \leq C_t \left(\frac{nt}{A^2} \right)^{t/2}$$

$$\text{where } C_t = 2\sqrt{\pi t} e^{\frac{-t}{6t}} \leq 1.0004.$$

We prove Lemma 2.2 using this lemma:

Lemma A.1

Let $t \geq 2$ be an even integer. Suppose Y_1, \dots, Y_n are independent r.v. taking values in $[0, 1]$. Let $Y = Y_1 + \dots + Y_n$ and $\mu = E[Y]$. Then,

$$E[(Y - \mu)^t] \leq 2e^{\frac{t}{6t}} \sqrt{t} \left(\frac{nt}{e} \right)^{t/2}$$

Now, we first need to provide a proof for Lemma A.1. It will use following Lemmas:

Lemma A.2

Let Z be the non-negative real valued r.v., $E[Z] = \int_0^\infty \Pr[Z > x] dx$.

proof of A.2:

$$\begin{aligned} \int_0^\infty \Pr(Z > x) dx &= \int_0^\infty \int_x^\infty f_Z(t) dt dx. & f_Z(t) : \text{pdf} \\ &= \int_0^\infty \int_0^t f_Z(t) dx dt & \text{change order of integral} \\ &= \int_0^\infty [x f_Z(t)]_0^t dt \\ &= \int_0^\infty t f_Z(t) dt \\ &= E[Z] & \text{as the definition of expectation} \end{aligned}$$

Lemma A.3

Suppose Y_1, \dots, Y_n are independent r.v. taking values in $[0, 1]$, $Y = Y_1 + \dots + Y_n$, $\mu = E[Y]$, $a > 0$. Then,

$$\Pr[|Y - \mu| > a] < 2e^{-a^2/2n}$$

proof of Lemma A.3:

Chernoff bound $\Pr(|Y-\mu| \geq \epsilon\mu) \leq 2e^{-\frac{\mu\epsilon^2}{3}}$

$$\Rightarrow \Pr(|Y-\mu| \geq \alpha) \leq 2e^{-\frac{\alpha^2}{3\mu}} \quad \text{set } \alpha = \epsilon\mu.$$
$$\leq 2e^{-\frac{\alpha^2}{3n}} \quad \text{since } \mu = E[Y] \leq n$$

$\leq \text{emmm 证不出来}$

直接用吧 :)

proof of Lemma A.1

By A.2,

$$E[(Y-\mu)^t] = \int_0^\infty \Pr[Y-\mu > x]^t dx$$
$$= \int_0^\infty \Pr[|Y-\mu| > x^{1/t}] dx \quad t \text{ is even}$$

By A.3,

$$\Pr[|Y-\mu| > x^{1/t}] \leq 2e^{-\frac{x^{2/t}}{2n}} dx.$$
$$\Rightarrow E[(Y-\mu)^t] \leq 2 \int_0^\infty e^{-\frac{x^{2/t}}{2n}} dx$$
$$= 2 \cdot \frac{t}{2} \cdot (2n)^{\frac{t}{2}} \int_0^\infty y^{\frac{t}{2}-1} e^{-y} dy$$

$y = \frac{x^{2/t}}{2n}$
 $\Rightarrow x = (2ny)^{\frac{t}{2}}$
 $dx = (2n)^{\frac{t}{2}} \cdot \frac{t}{2} \cdot y^{\frac{t}{2}-1} dy$

$$= 2 \cdot \frac{t}{2} \cdot (2n)^{\frac{t}{2}} \Gamma\left(\frac{t}{2}\right) \quad \begin{matrix} \text{Gamma function} \\ \Gamma(k-1) = k! = \int_0^\infty y^k e^{-y} dy \end{matrix}$$
$$= 2(2n)^{\frac{t}{2}} \left(\frac{t}{2}-1\right)! \frac{t}{2}$$
$$= 2(2n)^{\frac{t}{2}} \left(\frac{t}{2}\right)!$$
$$\leq 2(2n)^{\frac{t}{2}} e^{\frac{1}{8t}} \sqrt{\pi t} \left(\frac{t}{2e}\right)^{\frac{t}{2}}$$

$\text{sterling's formula}$
 $k! < e^{\frac{1}{12k}} \sqrt{2\pi k} \left(\frac{k}{e}\right)^k$

$$= 2e^{\frac{1}{8t}} \sqrt{\pi t} \left(\frac{nt}{e}\right)^{\frac{t}{2}}$$

Now, we will complete the proof for Lemma 2.2

proof

By A.1,

since t is even

$$E[|Y-\mu|^t] \stackrel{\downarrow}{=} E[(Y-\mu)^t] \leq 2e^{\frac{1}{8t}} \sqrt{\pi t} \left(\frac{nt}{e}\right)^{\frac{t}{2}}$$

may
 be
 useless... } Notice that $(X - \mu)^t$ is a polynomial in X_1, \dots, X_n of total degree t . By
 Linearity of expectation, $E[(Y - \mu)^t]$ can be computed under the assumption
 - that X_1, \dots, X_n are independent.

Then, apply Markov Inequality:

$$\begin{aligned}
 \Pr[|X - \mu|^t \geq A^t] &\leq \frac{E[|X - \mu|^t]}{A^t} \\
 &= \frac{E[(Y - \mu)^t]}{A^t} \\
 &\leq \frac{2e^{\frac{1}{6t}} \sqrt{nt} \left(\frac{nt}{e}\right)^{t/2}}{A^t} \\
 &= 2\sqrt{nt} e^{\frac{1-t}{6t}} \left(\frac{nt}{A^2}\right)^{t/2} \\
 &= C_t \left(\frac{nt}{A^2}\right)^{t/2}
 \end{aligned}$$

for $C_t = 2\sqrt{nt} e^{\frac{1-t}{6t}}$

□