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 $h_1, \dots, h_d$

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

$h_1(x)$: binary string; $\circ$: concatenate

Gradually more independent

$h_1 : O(1)$-wise indep.

$h_2 : O(h_1)$-wise indep.

...

$h_d : O(\log n / \log\log n)$-wise indep.

(k-wise, k increase ↓)

Note: ① output length decreases at the same time

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

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.

$SD(X, Y) \Rightarrow$ statistical distance between two r.v. over finite domain Ω

$$= \frac{1}{2} \sum_{\omega \in \Omega} |Pr[X=\omega] - Pr[Y=\omega]|$$

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

unit cost RAM model

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

k-wise δ-dependent

For a family of $f : [n] \to [v]$, it is k-___ δ-___ iff

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

where X = distribution $(f(x_1), f(x_2), \ldots, f(x_k))$ for any distinct $x_1, \ldots x_k \in [n]$

Y = uniform distribution over $[v]^k$

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

Evaluation Time: $O(k)$

ε-biased Distribution [NN93]

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

$$|Pr(\oplus_{i \in S} X_i = 1) - Pr(\oplus_{i \in S} X_i = 0)| \leq \varepsilon$$

($\oplus$: XOR operation)

[AGHP92, Sec. 5] constructs an ε-biased distribution over $\{0,1\}^n$ where each point could be specified using $O(\log(n/\varepsilon))$ bits where each bit could be calculated using $O(\log(n/\varepsilon))$ times.

In RAM, for word size of $w = \Omega(\log(\frac{n}{\varepsilon}))$ bits, each $t \in [n]$ consecutive bits could be calculated in time $O(\log(n/\varepsilon)\, t)$

Proof in AGHP92

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

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

(T, k, ε)-block source

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

$$\Pr_{(x_1,\ldots,x_{i-1}) \leftarrow (X_1,\ldots,X_{i-1})} [H_\infty(X_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) \geq k] \geq 1 - \varepsilon$$

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

Theorem

Corollary 2.1

Fact

for any k an ε-biased distribution is also k-wise δ-dependent for $\delta = \varepsilon 2^{k/2}$

For any u, v, $v = 2^i$, there exists a family of k-wise δ-dependent function $f: [u] \to [v]$ 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 with/or word size of $w = \Omega(\log u + k \log v + \log(\frac{1}{\delta}))$.

check AGHP92 for proof.

Corollary 2.2

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

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

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

↑ How X diverge from μ.

proved on paper using Markov Ineq.

Lemma 2.2 from [BR94]

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

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

$\le \frac{E[|X-\mu|^{2k}]}{t^{2k}}$ ← markov inequality $Pr(X \ge a) \le \frac{E[X]}{a}$

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

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

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

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} E[(X_i - \mu_i)^2]$$

$= \sum_{i_1,i_2 \in [n]} E \prod_{j=1}^{2} (i_j - j)$

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

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

let $\hat{X}$ be estimator value of X

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

$\because 0 \le \delta \le 1 \quad n, k > 0$

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

$\le \sum_{i_1,\ldots,i_{2k} \in [n]} E\left[\prod_{j=1}^{2k} (\hat{X}_{ij} - \mu_{ij})\right] + \delta n^{2k}$ where $\mu_i = E[X_i]$

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

consider $Pr(|\hat{X}-\mu| \ge t) \le \frac{E[(\hat{X}-\mu)^{2k}]}{t^{2k}}$

From BR94: $Pr((\hat{X}-\mu) > A) \le \frac{E[(\hat{X}-\mu)^t]}{A^t}$

$\le C_t \cdot \left(\frac{nt}{A^2}\right)^{\frac{t}{2}}$ where $C_t \le 1.0004$

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

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

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

may not be used by our construction.

Lemma 2.4 [GW97] ← may need a proof (but see from 4.1/4.2 main rescale)

r.v. $x_1 \in [0,1]^{n_1}$, $x_2 \in [0,1]^{n_2}$, $H_\infty(x_1 x_2) \ge h_1 + h_2 - a$

① $H_\infty(x_1) \ge h_1 - a$

② for any $\varepsilon > 0$,

$$\Pr_{x_1 \in x_1}\left[H_\infty(x_2 | x_1 = x_1) < h_2 - a - \log(1/\varepsilon)\right] < \varepsilon$$

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 $f: [u] \to [u]$ 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 $S \subseteq [u]$

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

↑

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

Construction

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 $[u] \to \{0,1\}^{l_i}$, where:

① $n_0 = n$ $\quad n_i = \frac{n^{i-1}}{2^{l_i}}$ for every $i \in [d]$

$n_1 = \frac{n}{2^{l_1}}$ $\quad n_2 = \frac{n_1}{2^{l_2}} = \frac{n}{2^{l_1+l_2}}$ $\dots$ $n_d = \frac{n}{2^{\sum l_i}}$

② $l_i = \left\lfloor \frac{\log n_{i-1}}{4} \right\rfloor$ for every $i \in [d-1]$, $l_d = \log n - \sum_{i=1}^{d-1} l_i$

$l_i = \left\lfloor \frac{\log n_{i-1}}{4} \right\rfloor = \left\lfloor \frac{\log n - \sum l_i}{4} \right\rfloor$ for $i \in [d-1]$

$l_d = \log n - \sum_{i=1}^{d-1} l_i$

③ $k_i l_i = \Theta(\log n)$ for every $i \in [d-1]$, $k_d = \Theta(\log n / \log\log n)$

$k_i l_i = k_i \left\lfloor \frac{\log n - \sum l_i}{4} \right\rfloor$

④ $\delta = poly(\frac{1}{n})$

By corollary 2.1, we have a family H_i : need detail.

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

We defined our function family F as

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

for $h_i \in H_i$

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

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

layer 0: has 1 bin with all balls, expected load $n_0 = n$ $\quad \frac{n_0}{2^{l_1+l_2}} = \frac{n_0}{2^{l_1}\cdot 2^{l_2}} = \frac{n_1}{2^{l_2}}$

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

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 $n_2 = \frac{\#\text{ of balls}}{\#\text{ of bins}} = \frac{n_0}{2^{l_1+l_2}}$

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

$\vdots$

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

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

number of balls in any bin of layer $i+1$ $\le (1 + \Omega(\frac{1}{\log\log n}))(1+\alpha_i)n_{i+1}$ with high prob.

Applications

① storing elements using linear probing

② augmenting k-wise independence function using an construction without affect the space/time requirement

need some evidence or proof.

Proof For Lemma 3.2

- Fix $y \in \{0,1\}^{L_{i+1}}$, let $X = |h_{i+1}^{-1}(y) \cap S_i|$
- assume WLOG, $|S_i| \geq \lfloor (1+\alpha_i) n_i \rfloor$ or we could add dummy elements to enlarge a smaller S_i.

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

  we could treat S_i as elements in one bin from last layer

new {
$X_j = 1$ if element $j \in S_i$ is hashed into h_{i+1}

$X = \sum_j^{|S_i|} X_j$ $\quad E[X] = \sum_j^{|S_i|} E[X_j] = \sum_j^{|S_i|} \frac{1}{2^{L_{i+1}}} = \frac{|S_i|}{2^{L_{i+1}}}$

(if uniform, total $2^{L_{i+1}}$ bins for each set S_i)

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

$$k = \frac{k_{i+1}}{2}, \quad \mu = E[X] = |S_i|/2^{L_{i+1}}$$

$$\Rightarrow Pr[X > (1+\alpha)\mu] \leq 2\left(\frac{|S_i| k_{i+1}}{(\alpha\mu)^2}\right)^{k_{i+1}/2} + \delta\left(\frac{|S_i|}{\alpha\mu}\right)^{k_{i+1}/2}$$

replace μ by $|S_i|/2^{L_{i+1}}$ →

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

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

- Now, we will upper bound each section

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

and ② $|S_i| \geq (1+\alpha_i) n_i - 1 \geq n_i$ ③ $\alpha = \Omega(1/\log\log n)$

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

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

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

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

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

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

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

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

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

$k_{i+1} = \frac{k_i w}{l_i} = \frac{4\log n^c}{\log n_i} = \frac{4c\log n}{\log n_i}$

$\alpha = \Omega\left(n^{\frac{1}{\sqrt{k}}}\right) = \Omega\left(\frac{1}{\log n}\right)$

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

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

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

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

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

$$f\left(\frac{2^{l_{i+1}}}{\alpha}\right)^{k_{i+1}} = f\frac{2^{l_{i+1}k_{i+1}}}{\alpha^{k_{i+1}}}$$

$$= f\frac{2^{\log n^c}}{\alpha^{k_{i+1}}}$$

$$\le f n^c$$

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

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

Thus, we get

$$\Pr[X > (1+\alpha)(1+\alpha_i) n_{i+1}] = \Pr\left[X > (1+\alpha)(1+\alpha_i)\frac{n_i}{2^{l_{i+1}}}\right] \quad \text{since } n_{i+1} = \frac{n_i}{2^{l_{i+1}}}$$

$$\le \Pr\left[X > (1+\alpha)\frac{|S_i|}{2^{l_{i+1}}}\right] \quad \text{since } |S_i| \le (1+\alpha_i) n_i$$

$$= \Pr[X > (1+\alpha)\mu] \quad \text{since } \mu = \frac{|S_i|}{2^{l_{i+1}}}$$

$$\le \frac{1}{2n^{c+2}} + \frac{1}{2n^{c+2}} \quad \text{by upper bound.}$$

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

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

$\Rightarrow$ union bound over y and subtract from 1

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

□

Proof For Theorem 3.1

Description Length \ Evaluation time

i^{th} layer $\Rightarrow h_i$ is k_i-wise δ-dependent function: $[u] \to \{0,1\}^{l_i}$

$v = 2^{l_i}$, $u = poly(n)$, $\delta = poly(\frac{1}{n})$

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

$$\begin{aligned} space &= O(\log u + k \log v + \log(c/\delta)) \\ &= O(\log u + k_i \log 2^{l_i} \log(\frac{1}{poly(\frac{1}{n})})) \\ &= O(\log u + k_i l_i \log(\frac{1}{\delta})) \\ &= O(\log n) \end{aligned}$$

$time = O(\log u + k \log v + \log(c/\delta)) = O(\log n)$ also.

Maximum Load

- Fix a set $S \subseteq [u]$ 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.

  for $\alpha = \Omega(c/\log\log n)$

Base Case $i = 0$

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

$\text{Prob}(\text{number of balls in one bin of layer } i+1 > (1+\alpha)^{i+1} n_{i+1}) \le \frac{1}{n^{c+1}}$ (at most n bins.)

Union bound $\Rightarrow \text{Prob}(\text{there exists a bin of layer } i+1 > (1+\alpha)^{i+1} n_{i+1}) \le \frac{n}{n^{c+1}}$

$\Rightarrow \text{Prob}(\text{there doesn't exist a bin of layer } i+1 \le (1+\alpha)^{i+1} n_{i+1}) \ge 1 - \frac{n}{n^{c+1}}$

$\ge 1 - \frac{i+1}{n^{c+1}}$

$\Rightarrow \text{Prob}(\text{maximum load of layer } i+1 = (1+\alpha)^{i+1} n_{i+1}) \ge 1 - \frac{i+1}{n^{c+1}}$

Thus, 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 $(1+\alpha)^{d-1} n_{d-1} \le 2 n_{d-1}$ for $d = O(\log\log n)$.

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

$$l_i \ge c \log n_{i-1}/4 - 1$$

$$\Rightarrow n_i = n_{i-1}/2^{l_i} \le 2 n_{i-1}^{3/4}$$

since $\frac{1}{2^{l_i}} \le \frac{1}{2^{\log n_{i-1}/4}} = \frac{1}{n_{i-1}^{3/4}}$

$$\Rightarrow n_{i-1} \le 2 n_{i-2}^{3/4} \Rightarrow n_i \le 2(2 n_{i-2}^{3/4})^{3/4} = 2^{1+3/4} n_{i-2}^{(3/4)^2}$$

$$\Rightarrow n_{i-1} \le 2^{\sum_{j=0}^{i-1} (3/4)^j} n^{(3/4)^i} \le 2^4 n^{(3/4)^i} = 16 n^{(3/4)^i}$$

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

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

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

$$n_{d-1} \le 16 n^{(3/4)^d} = 16 n^{(3/4)^{\log_{3/4}\left(\frac{\log \frac{\log n}{16}}{\log n}\right)}}$$

$$= 16 n^{\frac{\log \frac{\log n}{16}}{\log n}}$$

$$= 16 n^{\log_n\left(\frac{\log n}{16}\right)}$$

$$= 16 \frac{\log n}{16}$$

$$= \log n.$$

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

Thus, we know that, with Prob $\geq 1-(d-1)/n^{c+1}$, the maximum load in level $d-1$ is $(1+\alpha)^{d-1} n_{d-1} \leq 2n_{d-1} \leq 2\log n$.

(since $l_d = \log n_{d-1} \Rightarrow 2^{l_d} = 2^{\log n_{d-1}} = n_{d-1}$ bins)

These elements (in each bins) are hashed into n_{d-1} bins using the function h_d which is k_d-wise ξ-dependent, where $k_d = \Omega(\log n / \log\log n)$. 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.

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

($\binom{2n_{d-1}}{t}$: all comb of t within $2n_{d-1}$; $\frac{1}{n_{d-1}}$: one bin)

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

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

(since $n_{d-1} < \log n$ $\longrightarrow$)

$$\leq \left(\frac{2e\log\log n}{\gamma\log n}\right)^{\frac{\gamma\log n}{\log\log n}} + \xi\left(\frac{2e\log\log n}{\gamma}\right)^{\frac{\gamma\log n}{\log\log n}}$$

(detail(c?) $\longrightarrow$)

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

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

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

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

$$\Pr(\text{a bin with more than } t \text{ elements}) \leq \frac{1}{n^{c+1}}$$ since $2^{l_{d-1}} < 2^{l_d} = n_{d-1} \leq n$

This implies that

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

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

$\geq 1 - \frac{d}{n^{c+1}}$

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

□

[BR94] Lemma 2.2

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

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

where $C_t = 2\sqrt{\pi t}\, e^{\frac{1}{6t} - \frac{t}{2}} \leq 1.0004$.

We prove Lemma 2.2 using this Lemma:

Lemma A.1

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

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

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., $E[Z] = \int_0^\infty \Pr[Z \geq x]\, dx$.

proof of A.2:

$$\int_0^\infty \Pr(Z \geq x)\, dx = \int_0^\infty \int_x^\infty f_Z(t)\, dt\, dx \qquad f_Z(t): \text{pdf}$$

$$= \int_0^\infty \int_0^t f_Z(t)\, dx\, dt \qquad \text{change order of integral}$$

$$= \int_0^\infty [x f_Z(t)]_0^t\, dt$$

$$= \int_0^\infty t f_Z(t)\, dt$$

$$= E[Z] \qquad \text{as the definition of expectation}$$

Lemma A.3

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

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

proof of Lemma A.3:

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

$\Rightarrow \Pr(|Y-\mu| \geq a) \leq 2e^{-\frac{a^2}{3\mu}}$ set $a = \varepsilon\mu$.

$\leq 2e^{-\frac{a^2}{3n}}$ since $\mu = E[Y] \leq n$

$\leq$ emmm 证不出来

直接用吧 :)

proof of Lemma A.1

By A.2,

$$E[(Y-\mu)^t] = \int_0^\infty \Pr[(Y-\mu)^t > x]\,dx$$

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

By A.3,

$$\Pr[|Y-\mu| > x^{1/t}] \leq 2e^{-\frac{x^{2/t}}{2n}}\,dx.$$

$$\Rightarrow E[(Y-\mu)^t] \leq 2\int_0^\infty e^{-\frac{x^{2/t}}{2n}}\,dx$$

$$= 2\cdot\frac{t}{2}\cdot(2n)^{t/2}\int_0^\infty y^{\frac{t}{2}-1}e^{-y}\,dy$$

$y = \frac{x^{2/t}}{2n}$

$\Rightarrow x = (2ny)^{\frac{t}{2}}$

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

$$= 2\cdot\frac{t}{2}\cdot(2n)^{t/2}\,\Gamma\left(\frac{t}{2}\right)$$

Gamma function: $\Gamma(k-1) = k! = \int_0^\infty y^k e^{-y}\,dy$

$$= 2(2n)^{t/2}\left(\frac{t}{2}-1\right)!\,\frac{t}{2}$$

$$= 2(2n)^{t/2}\left(\frac{t}{2}\right)!$$

$$\leq 2(2n)^{t/2}e^{\frac{1}{6t}}\sqrt{\pi t}\left(\frac{t}{2e}\right)^{t/2}$$

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

$$= 2e^{\frac{1}{6t}}\sqrt{\pi t}\left(\frac{nt}{e}\right)^{t/2}$$

Now, we will complete the proof for Lemma 2.2

proof

By A.1, since t is even

$$E[|Y-\mu|^t] \leq E[(Y-\mu)^t] \leq 2e^{\frac{1}{6t}}\sqrt{\pi t}\left(\frac{nt}{e}\right)^{t/2}$$

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:

$$\Pr[|X-\mu|^t \geq A^t] \leq \frac{E[|X-\mu|^t]}{A^t}$$

$$= \frac{E[(Y-\mu)^t]}{A^t}$$

$$\leq \frac{2e^{\frac{1}{6t}}\sqrt{\pi t}\left(\frac{nt}{e}\right)^{t/2}}{A^t}$$

$$= 2\sqrt{\pi t}\, e^{\frac{1}{6t}-\frac{t}{2}}\left(\frac{nt}{A^2}\right)^{t/2}$$

$$= C_t\left(\frac{nt}{A^2}\right)^{t/2}$$

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

□

AGHP92

- [AGHP92, Sec. 5] constructs an ε-biased distribution over $\{0,1\}^n$ where each point could be specified using $O(\log(n/\varepsilon))$ bits where each bit could be calculated using $O(\log(n/\varepsilon))$ times.
- In RAM, for word size of $w = \Omega(\log(\frac{n}{\varepsilon}))$ bits, each $t \in [n]$ consecutive bits could be calculated in time $O(\log(n/\varepsilon)\, t)$

Notes: seed contains $x, y \in GF[2^m]$ where $m = O(\log n/\varepsilon)$ and the ith output bit is the inner product modulo 2 of the binary representation of x^i and y.

- Combine the fact that for any k, ε-biased distribution is also k-wise δ-dependent if $\delta = \varepsilon 2^{k/2}$ [AGHP92, Cor. 1] ✓
- Set $\begin{cases} n = v \log v = u \text{ blocks and } \log v \text{ bits, each represents a single output value in } [v] \\ \varepsilon = \delta 2^{-k \log v/2} \end{cases}$
- Then we could derive **Corollary 2.1.**

Theorem

Definition 2 almost k-wise independence.

Let S_n be sample space and $X = x_1, \ldots, x_n$ be chosen uniformly randomly from S_n,

① S_n is (ε, k)-independent (in max norm) if for any k positions $i_1 < i_2 < i_3 \cdots < i_k$, and any k-bit string α, we have

$$\left| \Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k} \right| \le \varepsilon$$

In definition before, $SD(X, Y) \le \varepsilon$ for Y = uniform distribution

② S_n is ε-away (in L_1 norm) from k-wise independence if for any k positions $i_1 < i_2 \cdots < i_k$, we have

$$\sum_{\alpha \in \{0,1\}^k} \left| \Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k} \right| \le \varepsilon.$$ (2^k different α each $\le \varepsilon$)

S_n is (ε, k) independent $\Rightarrow$ at most $2^k \varepsilon$-away from k-wise indep.

S_n is ε-away from k-wise indep. $\Rightarrow$ S_n is (ε, k) independent

Linear Test

test which take the exclusive-or of the bits in some fixed location in the string.

Definition 3

- Let $(\alpha, \beta)_2$ denote the inner-product-mod-2 of the binary string α and β

$$(\alpha_1 \dots \alpha_n, \beta_1 \dots \beta_n)_2 = \sum_{i=1}^{n} \alpha_i \beta_i \bmod 2$$

- A 0-1 random variable X is ε-biased if

$$|Pr[X=0] - Pr[X=1]| \le \varepsilon$$

- Let S_n be a sample space and $X = x_1 \dots x_n$ be chosen uniformly randomly from S_n. The sample space S_n is said to be ε-biased with respect to Linear tests if for every $\alpha = \alpha_1 \dots \alpha_n \in \{0,1\}^n - \{0\}^n$, the r.v. $(\alpha, X)_2$ is ε-biased

$\{0\}^n$ is not valid

- Moreover, The sample space S_n is said to be ε-biased with respect to Linear tests of size at most k if for every $\alpha = \alpha_1 \dots \alpha_n \in \{0,1\}^n - \{0\}^n$ s.t. at most k of the α_i are 1, the r.v. $(\alpha, X)_2$ is ε-biased

Lemma 1

Let $S_n \subset \{0,1\}^n$ be a sample space that is ε-biased w.r.t. Linear tests of size at most k. Then, the sample space S_n is $((1-2^{-k})\varepsilon, k)$-independent (in max norm), and $(2^k-1)^{1/2}\varepsilon$-away (in L_1 norm) from k-wise indep.

or Corollary 1

Let $S_n \subset \{0,1\}^n$ be a sample space that is ε-biased w.r.t. Linear tests of size. Then, for every k, the sample space S_n is $((1-2^{-k})\varepsilon, k)$-independent (in max norm), and $(2^k-1)^{1/2}\varepsilon$-away (in L_1 norm) from k-wise indep.

Connection to Corollary 2.1 in ball & bins.

Fact

for any k an ε-biased distribution is also k-wise δ-dependent for $\delta = \varepsilon 2^{k/2}$

k-wise δ-dependent $\Leftrightarrow SD(X, U) \le \delta$, U = uniform dist. over $[v]^k$

$$\frac{1}{2}\sum_{w \in \Omega} \left| Pr[X=w] - Pr[U=w] \right| \le \delta$$

$$\Leftrightarrow \frac{1}{2}\sum_{w \in \Omega} \left| Pr[X=w] - \frac{1}{2^k} \right| \le \delta$$

$$\Leftrightarrow \sum_{w \in \Omega} \left| Pr[X=w] - \frac{1}{2^k} \right| \le 2\delta = \varepsilon 2^{\frac{k}{2}+1} \Leftrightarrow \boxed{\varepsilon 2^{\frac{k}{2}+1}\text{-away}}$$

for any **ε-biased distribution** $\Rightarrow (2^k - 1)^{\frac{1}{2}}\varepsilon$-away

$$\Rightarrow \sum_{w \in \Omega} \left| Pr[X=w] - \frac{1}{2^k} \right| \le (2^k-1)^{\frac{1}{2}}\varepsilon \le \varepsilon 2^{\frac{k}{2}+1}$$

$\Rightarrow \varepsilon 2^{\frac{k}{2}+1}$-away $\Rightarrow$ **k-wise δ-dependent**

$\Rightarrow$ Fact stands.

Remark

when apply this Lemma / Corollary, we will use the bounds ε and $2^{k/2}\varepsilon$ respectively.

ε: max norm.

$2^{k/2}\varepsilon$: L_1 norm

Construction | **AGHP92 Sec. 5** $S_n^{2m} \Rightarrow$ a sample space of 2^{2m} strings each of length n.

- Let $m = \log(\frac{n}{\varepsilon})$ and $bin : GF(2^m) \mapsto \{0,1\}^m$ be a one-to-one mapping function satisfies that

① $bin(0) = \{0\}^m$

② $bin(u+v) = bin(u) \oplus bin(v)$ where $\oplus$ is the bitwise XOR.

which should be fulfilled by the standard representation of $GF(2^m)$.

- A string $s \in S_n^{2m}$ is specified using two field elements x and y. The i^{th} bit in s is the inner-product-modulo-2 of x^i and $y \Rightarrow (bin(x^i), bin(y))_2$ where x^i is the i^{th} power of x when considered as an element in $GF(2^m)$.

Claim Sample space S_n^{2m} is $\frac{n-1}{2^m}$-biased with respect to linear tests.

Namely, For any non-zero α, the r.v. $(\alpha, r)_2$ is $(n-1)2^{-m}$-biased when r is uniformly selected from S_n^{2m}.

proof

Let $r(x,y) = r_0(x,y) r_1(x,y) \dots r_n(x,y)$. Then,

$$(\alpha, r(x,y))_2 = \sum_{i=0}^{n-1} \alpha_i r_i(x,y) \bmod 2$$

$$= \sum_{i=0}^{n-1} \alpha_i (bin(x^i), bin(y))_2 \bmod 2$$

$$= (bin(\sum_{i=0}^{n-1} \alpha_i x^i), bin(y))_2$$

$$= (bin(p_\alpha(x)), bin(y))_2 \quad \text{for } p_\alpha(t) = \sum_{i=0}^{n-1} \alpha_i t^i$$

as a poly over $GF(2)$ $\in GF(2^m)$

When $x, y \in GF(2^m)$ are chosen uniformly, we fix x:

① x is not a zero of $p_\alpha(x) \Rightarrow p_\alpha(x) \neq 0$

$\Rightarrow bin(p_\alpha(x)) \neq \{0\}^m$

set $bin(p_\alpha(x)) = p_0 p_1 \cdots p_{2^m-1}$

$\Rightarrow (bin(p_\alpha(x)), bin(y))_2 = \sum_{i=0}^{2^m-1} p_i y_i \bmod 2$

$\Rightarrow (\text{bin}(p_\alpha(x)), \text{bin}(y))_2$ is unbiased since

y is also uniformly selected from $GF(2^m)$

② x is a zero of $p_\alpha(x) \Rightarrow p_\alpha(x) = 0$

$\Rightarrow \text{bin}(p_\alpha(x)) = \{0\}^m$

$\Rightarrow (\text{bin}(p_\alpha(x)), \text{bin}(y))_2 = 0$

for all y.

However,

① $p_\alpha(t)$ has at most $n-1$ zeros.

② $x \in GF(2^m)$

Thus, there are $\leq n - 1/2^m$ of all values of x such that

$(\text{bin}(p_\alpha(x)), \text{bin}(y))_2 = 0$ for all values of y.

$\Rightarrow (\text{bin}(p_\alpha(x)), \text{bin}(y))_2$ is $\frac{n-1}{2^m}$-biased.

□

Connection to Corollary 2.1 in book & boss.

① - [AGHP92, Sec. 5] constructs an ε-biased distribution over $\{0,1\}^n$ where each point could be specified using $O(\log(n/\varepsilon))$ bits where each bit could be calculated using $O(\log(n/\varepsilon))$ times.

- In RAM, for word size of $w = \Omega(\log(\frac{n}{\varepsilon}))$ bits, each $t \in [n]$ consecutive bits could be calculated in time $O(\log(n/\varepsilon)\, t)$

Notice: seed contains $x, y \in GF[2^m]$ where $m = O(\log n/\varepsilon)$ and the ith output bit is the inner product modulo 2 of the binary representation of x^i and y.

$x, y \in GF[2^m]$ where $m = O(\log n/\varepsilon)$

Space: this construction output $\{0,1\}^n \Rightarrow O(\log \frac{n}{\varepsilon})$

time: i^{th} bit of the sample point is $C(bin(x^i), bin(y))_2$

$\Rightarrow O(m) = O(\log \frac{n}{\varepsilon})$ times for 1 bit.

$\Rightarrow O(\log \frac{n}{\varepsilon} \cdot t)$ for t bits.

② - Set $\begin{cases} n = u \log v = u \text{ blocks and } \log v \text{ bits, each represents a single output value in } [v] \\ \varepsilon = \delta 2^{-k \log v / 2} \end{cases}$

- For any u, v, $v = 2^i$, there exists a family of k-wise δ-dependent function $f: [u] \to [v]$ 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 with a word size of $w = \Omega(\log u + k \log v + \log(\frac{1}{\delta}))$.

space/time: $O(\log \frac{n}{\varepsilon}) = O(\log \frac{u \log v}{\delta 2^{-k \log v / 2}})$

$= O(\log u + \log \log v + \log(\frac{1}{\delta}) + \frac{k \log v}{2})$

$= O(\log u + k \log v + \log(\frac{1}{\delta}))$ since $\log \log v \in O(\log v)$

word size: $\Omega(\log \frac{n}{\varepsilon}) = \Omega(\log \frac{u \log v}{\delta 2^{-k \log v / 2}})$

$= \Omega(\log u + \log \log v + \log(\frac{1}{\delta}) + \frac{k \log v}{2})$

$= \Omega(\log u + \frac{k}{2} \log v + \log(\frac{1}{\delta}))$

$\approx \Omega(\log u + k \log v + \log(\frac{1}{\delta}))$