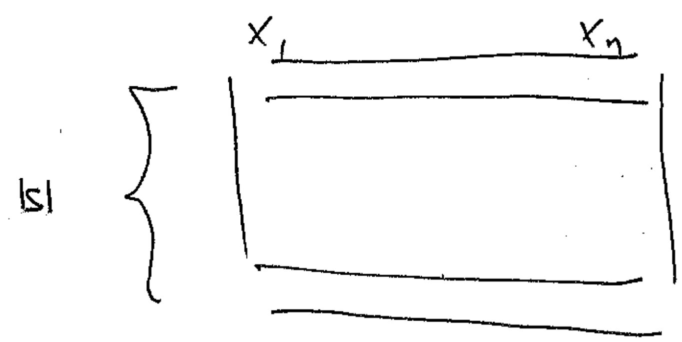
U.F.O.H.F

נמצא את התפלגות $X = X_1,\dots,X_n$ שהוא k-wise ind.

ונניח שהוא מתקבלת את התוך מרחב S.



אנחנו יכולים ליצור את ההתפלגות כדי שנבחר ע"י הבחירה

נבחר: (באופן אקראי את S) ואז

$$(X_1,\dots,X_n) = (h_s(1),\dots,h_s(n))$$

כי $h_s(i)$ הוא מה שקובע כאשר בוחרים את S ומציבים i.

הגדרה: משפחה $H=\{h: \Sigma \to \Gamma\}$

KUFOHF k-Universal family of hash functions

אם: לכל $\sigma_1,\dots,\sigma_k \in \Sigma$ שונים

ולכל $b_1,\dots,b_k \in \Gamma$

$$\Pr_{h\in H}\left[(h(\sigma_1)=b_1) \wedge \dots (h(\sigma_k)=b_k)\right] = \frac{1}{|\Gamma|^k}$$

הבעיה:

קלט: $X_1 \dots X_n \in \Sigma$

לכל $a \in \Sigma$ $\quad m_a = |\{i \in [n] \mid x_i = a\}|$

רוצים לקרב את $F_2 = \sum_{a \in \Sigma} m_a^2$.

המצב החישובי:

הקלט Read once access

רוצים רק זכרון משך $O(\lg(n \cdot |\Sigma|))$.

האלגוריתם: הנח $|\Sigma| \le n$.

נקח משפחה $H = \{h : \Sigma \to \{1,-1\}\}$

משפחה 4-U.F.O.N.F.

- האלגוריתם בוחר $h_1, \dots, h_T \in_R H$ (נקבע את T אח״כ)
- מחשב $S_t = \sum_{i=1}^{n} h_t(x_i)$ $(t=1,\dots,T)$
- הפלט הוא $\frac{1}{T} \sum_{t=1}^{T} (S_t)^2$.

סיבוכיות זכרון: $|H| = poly(n)$. $O(T \cdot \lg n)$. Read once

הוכחה :

נזכיר מה קורה כשבוחרים $h\in H$ מקרית.

ומחשבים $S^2=\left(\sum_{i=1}^{n} h(x_i)\right)^2$.

מתוך הנוסח : בוחרים $h\in H$ באופן אקראי.

$$E(S^2)=E\left(\sum_{i=1}^{n} h(x_i)\sum_{j=1}^{n} h(x_j)\right)=E\left(\sum_{a\in\Sigma} m_a\cdot h(a)\cdot\sum_{b\in\Sigma} m_b\cdot h(b)\right)$$

$$=\sum_{a,b\in\Sigma} m_a m_b \, E_{h\in H}\, h(a)\cdot h(b)$$

$$=\sum_{\substack{a:\\ b=a}} m_a^2 \underbrace{E_{h\in H}(h(a))^2}_{1}+\sum_{a\neq b} m_a m_b \underbrace{E_h\left(h(a)\,h(b)\right)}_{\substack{= \text{ אי-תלות בזוגות} \\ (E_h(h(a)))\cdot E_h(h(b)) \\ \overset{\shortparallel}{0}\ \cdot\ 0}}=\sum_a m_a^2=F_2$$

$$E(S^4)=\sum_{a,b,c,d\in\Sigma} m_a m_b m_c m_d\, E\left(h(a)\,h(b)\,h(c)\,h(d)\right)$$

$$=\sum_{\substack{a\\ a=b=c=d}} m_a^4+\sum_{a<b}\binom{4}{2} m_a^2 m_b^2 E_h\left(\underbrace{h^2(a)}_{1}\ \underbrace{h^2(b)}_{1}\right)+\sum_{a,b,c,d\in\Sigma} m_a m_b m_c m_d \underbrace{E\left(h(a)\ldots h(d)\right)}_{\overset{\shortparallel}{0}}$$

בוחרים 2 מקומות עבור a נוסף
ומתוכם 2 " " b נוסף.

לפי אי-תלות רביעיות $\xrightarrow{\text{ולכן}}$
פחות זוגות

$$=\sum_{a\in\Sigma} m_a^4+6\sum_{a<b} m_a^2 m_b^2$$

סימון: $. Z = S^2$

למה:

$$E_{h \in H}(Z) = F_2$$

$$Var_{h \in H}(Z) \le 2F_2$$

הוכחה:

$$Var(Z) = E\left((Z - E(Z))^2\right) = E\left(Z^2 - 2E(Z)Z + E^2(Z)\right)$$

$$= E(Z^2) - 2E^2(Z) + E^2(Z) = E(Z^2) - (E(Z))^2$$

$$Var(Z) = E(S^4) - (E(S^2))^2 = \sum_{a \in \Sigma} m_a^4 + 3\sum_{a \ne b} m_a^2 m_b^2 - \left(\sum_{a \in \Sigma} m_a^2\right)^2$$

$$= 3\underbrace{\sum_{a,b} m_a^2 m_b^2}_{(\sum_a m_a^2)^2} - 2\sum_a m_a^4 - \left(\sum_a m_a^2\right)^2$$

$$= 2(F_2^2 - F_4) \le 2F_2^2$$

סימון

$$Z = \cancel{\tfrac{1}{T}} \sum_{t=1}^{T} S_t^2$$

$$E(Z) = \cancel{\tfrac{1}{T}} \sum_t E(S_t^2) = T \cdot F_2$$

$$V(Z) = \cancel{\tfrac{1}{T^2}} \sum_t V(S_t^2) \le \frac{T}{\cancel{T^2}} 2F_2 = 2TF_2$$

אי-תלות הדדית.
מספיק: אי תלות בזוגות.

$$Pr\left(|Z - E(Z)| \ge \delta T F_2\right) \le \frac{Var(S)}{\delta^2 F_2^2 T^2} \le \frac{2F_2^2 \cancel{T}}{T^{\cancel{2}} \delta^2 F_2^2} = \frac{2}{T\delta^2} \le \frac{1}{3}$$

$\delta = \frac{1}{100}$ צ'בישב

נבחר T גדול מספיק, כלומר $T \ge \frac{6}{\delta^2}$

[illegible]

אם עכשיו נריץ $\Omega(\log\frac{1}{\delta})$ פעמים (כל פעם מריץ T פעמים)

ונקח את החציון, אז:

$$\Pr_{\text{חציון}}\left[\,|\text{חציון} - F_2| \ge \varepsilon\,\right] \le \delta$$

כי: חוק משהו-אזה

$$\delta = 2^{-c\cdot\lg\frac{1}{\delta}}$$

ורק $\frac{1}{2}$ מהניסיונות, נפלו לקצב ε F_2

ולכן גם החציון.

מה שאלו: $\frac{1}{100}$ קרוב בהסת' $\ge 1-\delta$

צריך זמן $O(\lg n \cdot \lg\frac{1}{\delta})$ כזמן.

- הנחות שלא מראה שאפשר גם לבחור p.i. $h_{\#}$, $h_{i,j}$

- למה זה אפשר לבחור H 2-wise ind

$h_{\lg\frac{1}{\delta}}, \dots, h_1$ מהלם כי ולהשיג רמת בטחון ישר?