Licznik Morrisa:

C_n = wartość po n-aktual.

- $C_n \approx \log_2 n$
- $E[2^{C_n}] = n+1$

$\hat{n} = 2^{C_n} - 1$ ← estymator n

- ile bitów do zapisania C_n:

$\sim \log_2(\log_2 n)$

↑ nie można zbić „Teorie informacji"

$$0 \xrightarrow{1} 1 \xrightarrow{\frac{1}{2}} 2 \xrightarrow{\frac{1}{2^2}} 3 \xrightarrow{\frac{1}{2^3}} 4 \to \dots$$

$p = \frac{1}{2}$; tym p można żonglować

np. $p = \frac{3}{4}$; $E[C_n]$ rośnie, $\mathrm{var}(2^{C_n})$ maleje

Licznik geometryczny:

- stosunkowo mało pamięci
- liczenie różnych elementów

$\{a, a, b, a, a, b, c\} \equiv \{a, b, c\}$

"distinct counter"

IDEA: użyj dobrej funkcji hash: $h: \Sigma^* \longrightarrow \{0,1\}^m$

$\{a, a, b, a, a, b, c\} \rightarrow \{h(a), h(a), \dots, h(c)\}$

- $h(a) = \underbrace{\text{"}0000 1000 \dots 0\text{"}}_{m}$

$\boxed{m \approx 64}$

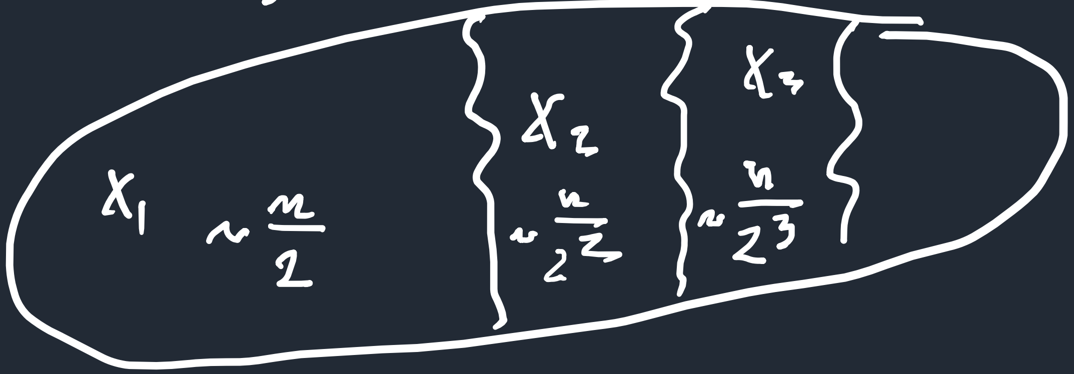
$$\varphi(a) = \begin{cases} \min\{k : h(a)_k = 1\} & : \exists k \; h(a)_k = 1 \\ m+1 & : \forall k \; h(a)_k = 0 \end{cases}$$

$$\varphi(a) = \min\{ i : h(a)_i = 1 \}$$

1. X – duży zbiór elementów

$$|X| = n$$

$$\varphi : \Sigma^* \to \{1, 2, \ldots, m\} \cup \{\mathrm{inf}\}$$



$$X_1 = \{ a \in X : \varphi(a) = 1 \}$$
$$= \{ a \in X : h(a) = \text{'}1*\text{'} \}$$
$$X_2 = \{ a \in X : h(a) = \text{'}01*\text{'} \}$$
$$X_i = \{ a \in X : h(a) = \text{'}\underbrace{00..0}_{i-1}1*\text{'} \}$$

$$|X_1| \simeq |X|/2$$

$$|X_i| \approx \frac{n}{2^i}$$

KIEDY: $E[|X_i|] \leqslant 1$

$$\frac{n}{2^i} \leqslant 1 \equiv 2^i \geqslant n$$
$$\equiv \boxed{i \geqslant \lg_2 n}$$

Licz. geom:

$L = 0$; onUpdate(a) {

$s = h(a)$;

$i = \min\{k : s_i = 1\}$ (lub mtl)

$L = \max\{L, i\}$

}

$L = 7$;

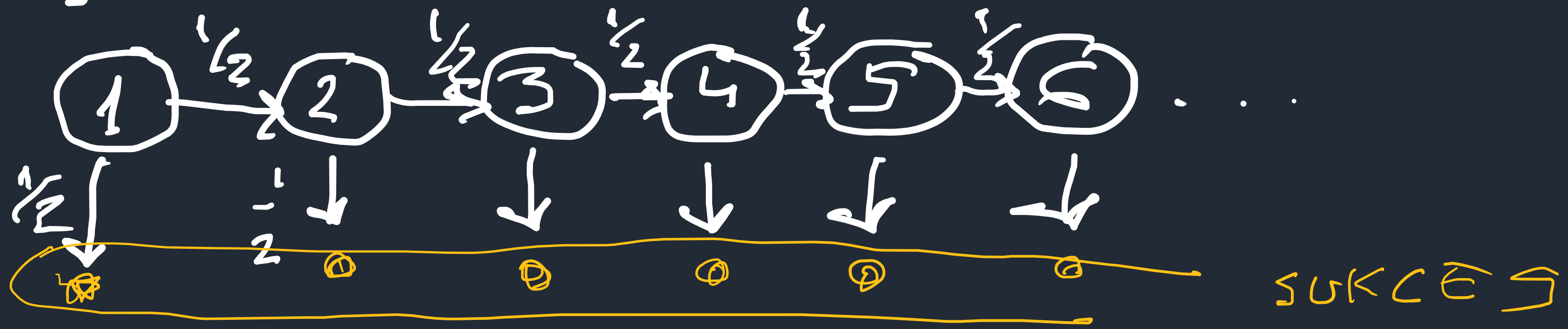
$a \overset{h}{\rightsquigarrow} (000100\ldots)$: $L = L$

$a \rightsquigarrow (0,0,0000000\underset{\substack{\uparrow \\ 9}}{1}000.)$ $L = 9$

Precyzja:

ciąg losowy 0-1 : $X_1 X_2 X_3 X_4 \dots X_m$

X :



$X \sim Geo(1/2)$ $\qquad$ $E[X] = 2$

mamy $X_1, X_2, \dots, X_n$ zm. los. niezal. o rozkł. $Geo(\frac{1}{2})$

$$Y = \max\{X_1, X_2, \dots, X_n\}$$

$$Y = \max\{X_1 \dots X_n\}$$

$$P[Y \leq k] = P[X_1 \leq k \wedge \dots \wedge X_n \leq k]$$

$$= P[X_1 \leq k]^n = (1 - P[X_1 > k])^n$$

$$= \left(1 - \left(\tfrac{1}{2}\right)^k\right)^n$$

$$= \left(1 - \frac{1}{2^k}\right)^n$$

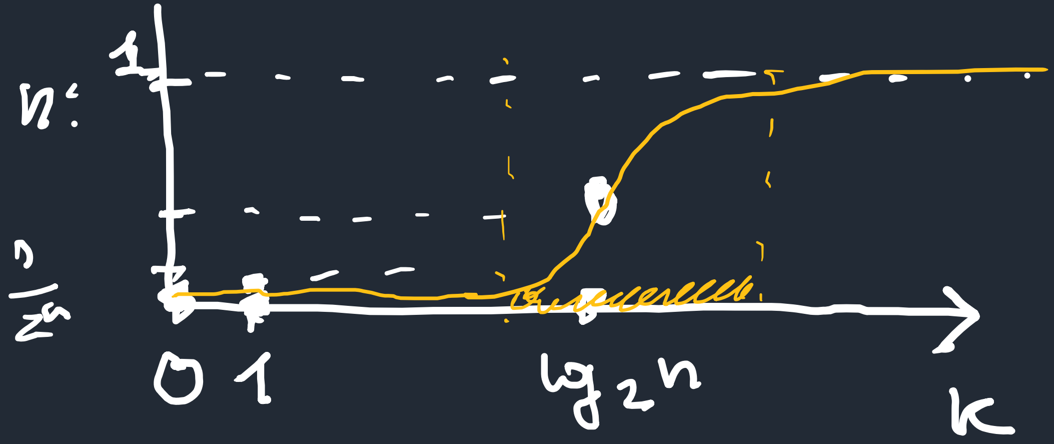
$Z \sim \mathrm{Geo}(p)$

$$1 \xrightarrow{p} 2 \xrightarrow{q} 3 \xrightarrow{q} 4 \xrightarrow{q} 5 \dots$$

$$P[Z > 3] = q^3$$

$q = 1 - p$

Fix n:



$$\left(1 - \frac{1}{2^n}\right)^n$$

$$k \leftarrow \lg_2 n$$

$$\left(1 - \frac{1}{2^{\lg_2 n}}\right)^n = \left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e}$$

$$\left(1 - \frac{1}{2}\right)^n = \frac{1}{2^n}$$

$Y = \max\{X_1 \dots X_n\}$ $\qquad X_i \sim \mathrm{Geo}(\frac{1}{2})$

$$E[Y] = \frac{1}{2} + \frac{H_n}{\ln 2} + r_n \qquad |r_n| < 0.01$$

$$H_n = 1 + \frac{1}{2} + \dots + \frac{1}{n} = \ln n + \gamma + O(\frac{1}{n}) \approx \ln n$$

nietrywialne $\approx \frac{\ln n}{\ln 2} = \lg_2 n$.

POMYSŁ: $\hat{n} = 2^L$

≈ 1990 (Martin, Flajolet)

Problem: dokładność

POMYSŁ: równoczesne liczniki $L_1, \dots, L_m$

użyj $\frac{L_1 + \dots + L_m}{m}$ $=$ średnia$(L_1, \dots, L_m)$

FAKT: $L_1, \dots, L_n$ – ten sam rozkład niezal.

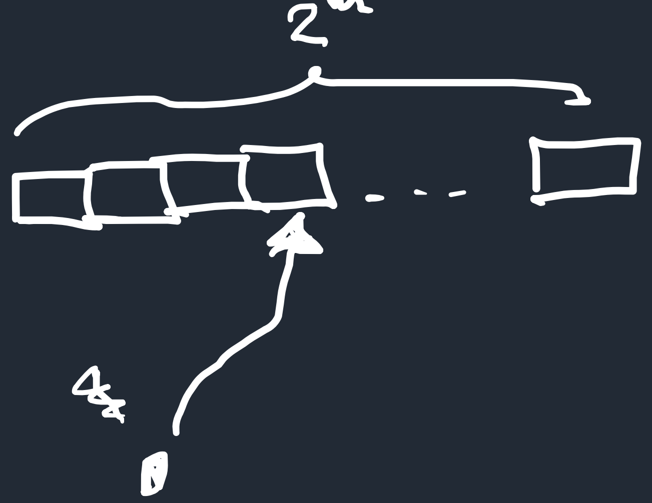
$Z_n = L_1 + \dots + L_n$

- $E[Z_n] = n \cdot E[L_1]$
- $var(Z_n) = n \cdot var(L_1)$: niezal.
- Czebyszew:

$$P(|Z_n - E[Z_n]| \geq \alpha E[Z_n]) \leq \frac{var(Z_n)}{\alpha^2 E[Z_n]^2} = \frac{n\, var(L_1)}{\alpha^2 \cdot n^2 E[L_1]} = \frac{1}{\alpha} \cdot \frac{1}{n} \cdot C$$

$$P(|Z_n - nE[L_1]| \geq C) = P\left(\left|\frac{Z_n}{n} - E[L_1]\right| \geq \frac{C}{n}\right) = \dots$$

POMYSŁ: $L[0, \ldots, 2^m - 1]$



$$h(a) = \text{'}x_1 \ldots x_M\text{'} =$$

$$= \text{'}\underbrace{x_1 \ldots x_m}_{m}\text{'} \,|\, \text{'}x_{m+1} \ldots x_M\text{'}$$

onUpdate(a) {

$h(a) = p \,\|\, s \quad ; \quad |p| = m$

$L[(p)_{(2)}] = \max\{L[(p)_{(2)}], \varphi(s)\}$

$|X| = n$

tu trafi średnio $\frac{n}{M}$ elem.

$|X| = n$

$$2^{L[i]} \simeq \frac{n}{M}$$

$$n \simeq M \cdot 2^{L[i]}$$

Ostateczny estymator

$$\text{średnia}\left(M \cdot 2^{L[0]}, \ldots, M \cdot 2^{L[M-1]}\right)$$

LOG-LOG

HYPER-LOG-LOG:

$$\text{średnia}(x_1 \ldots x_n) = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}$$

średnia harm.

ANALIZE $\overline{\underline{\text{II}}}$. $x_1, \dots, x_n > 0$

$$\frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}} \le \sqrt[n]{x_1 \cdot \dots \cdot x} \le \frac{x_1 + \dots + x_n}{n}$$

śr. harmon. śr. geom. śr. arytm.

{mało wrażliwa
{na outliers

$$x_i = m \cdot 2^{L[i]}$$

$$\frac{m}{\frac{1}{m}\left(\frac{1}{2}\right)^{L[0]} + \dots + \frac{1}{m}\left(\frac{1}{2}\right)^{L[m-1]}} = \frac{m^2}{\frac{1}{2^{L[0]}} + \dots + \frac{1}{2^{L[m-1]}}}.$$

on GetEstimator( ) {

$$Z = \frac{M^3}{\frac{1}{2^{L[0]}} + \dots + \frac{1}{2^{L[M-1]}}};$$

return : $\alpha_m \cdot Z$

}

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

$K = 2^m$

M	α_m
16	0.673
32	0.69..
64	0.709
≥ 128	$\frac{0.7213}{1 + \frac{1.079}{M}}$

OSTATECZNA KOREKTA

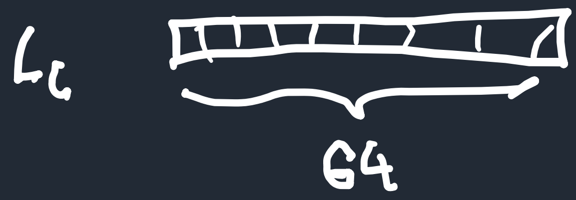
poniżej $M \log M$

- małe n

  Jeśli są zerowe liczniki, to zastosuj LINEAR COUNTING;
  czyli: $z \leftarrow$ liczba pustych rejestrów

  $$\boxed{M \cdot \ln \frac{M}{Z}}$$

- przekroczenie rozmiaru któregoś z liczników:



korekta $\longrightarrow$
$$\begin{cases} L > \dfrac{2^{64}}{30} \\ L = -2^{64} \ln\left(1 - \dfrac{L}{2^{64}}\right) \end{cases}$$

DOKŁADNOŚĆ : (stand. odchylenie)

$$\approx \frac{1.04}{\sqrt{M}},$$

CZYLI :

$$\begin{cases} M = 1024 \\ \sqrt{M} = 33 \end{cases} \qquad \sigma \approx \frac{1}{33} \qquad \boxed{3\%}$$

$$M \approx 128 \qquad \sqrt{M} = 11 \qquad \sigma \approx \frac{1}{10} \; ; \; 10\%$$

Z ZAIMPLEMENTUJ HLL