# Codifica di Huffman / 23-04

$H_0 = \log_2 |V_x|$ dove $V_x$ = # di simboli

quantità di informazione grezza

è l'entropia max dei simboli $x_i$ (equiprobabilità)

Def. $C(x)$ $\forall x \in V_x$ una codifica

$L_C(x) < H_0$ $\forall x \in V_x$

dati M bit, posso identificare file
$2^n$ file grandi: M bit

0 00 $2^{2+1} - 2$
1 01
10
11

1 Gb: $2^{30}$ file grandi: 1 Gb: $2^{31}-2$

n° di file = $V_x$

$H_0 = \log_2 (2^{31}-2) \simeq 31$

0 00 000
1 01 001
10 010
11 100
101
110
111

tutti i file fino a 3 bit
con una compressione mi aspetto una riduzione
(a 2 bit)
$H_0$ 4 caselle con 2 bit, ma ho 7 caselle
da 3 bit, eppure devo codificarle in
tutte caselle diverse (univocità)

↓

dunque può succedere che comprimendo
molte volte, la dimensione (di un file
già compresso) aumenta

La lunghezza attesa $L(C, V_X)$ non può essere minore di $H(X)$

So $\log_2 e(t-1) \geq \log_2(t)$

la retta sta sopra il log

Le lunghezze delle codifica sono $L_1 \dots L_{|V_X|}$

Univocamente decifrabile (file diversi in codifiche diverse)

Considero $Z = \sum z^{-L_i}$ $\quad q_i = \frac{2^{-L_i}}{Z}$ ciascun termine della somma / sta normalizzando

$\sum q_i = 1$ (somma / somma)

$\log_2 q_i = \log_2 2^{-L_i} - \log_2 Z$

$\log_2 q_i = -L_i - \log Z$

per cui $L_i = -\log_2 q_i - \log_2 Z = \log \frac{1}{q_i} - \log Z = L_i$

lunghezza di ogni codifica

$\Phi$: $q_1 \dots q_{|V_X|}$ della $t = \frac{q_i}{p_i}$

$p_1 \dots p_{|V_X|}$

$\sum_i p_i \log \frac{1}{p_i} - \sum_i p_i \log \frac{1}{q_i} = \sum p_i \left(\log \frac{q_i}{p_i}\right)$

$H$

$= \sum p_i \log_2 t_i$

$\sum p_i \log_2 \frac{1}{q_i} \leq \sum p_i (t_i - 1)$ per la concavità del log di cui sopra

$= \sum p_i \log_2 \frac{1}{q_i} - \sum p_i$

$1 - 1 = 0$

$= 0$ se $q_i = p_i$

$L(C, V_X) = \sum p_i L_i = \sum p_i (\log_2 \frac{1}{q_i} - \log Z)$

$L_i = \log_2 \frac{1}{p_i} - \log Z$

$\geq H(X)$

VINCOLO INVALICABILE

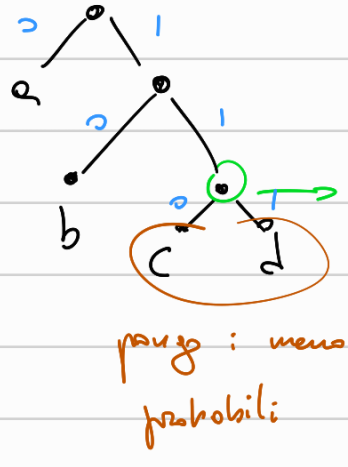
se vado sotto l'entropia qualcos'altro deve aumentare (la lunghezza della sequenza) x rispettare l'uguaglianza

## Codifica di Huffman

| | X | $p_i$ |
|---|---|---|
| * | | |
| 0 | a | $\frac{1}{2}$ |
| 10 | b | $\frac{1}{4}$ |
| 110 | c | $\frac{1}{8}$ |
| 111 | d | $\frac{1}{8}$ |

↓ possono essere lettere, numeri, file

Costruisco un albero




$p = \sum p$ figli

prendo i meno probabili

STEP 1: a $\frac{1}{2}$, b $\frac{1}{4}$, (c,d) $\frac{1}{4}$

STEP 2: a $\frac{1}{2}$, (b,(c,d)) $\frac{1}{2}$

STEP 3: (a,(b,(c,d))) 1

Ragionato attraverso una coda (estraggo man mano i nodi)

Se prendo una codifica di 0 e 1 ottengo * dopo aver costruito l'albero che è univoca e istantanea

Come input necessito di probabilità

$V_x$ e $p_i$ con $i = 1 \dots |V_x|$

- Huffman è, per simbolo, la codifica ottimale

prende un simbolo alla volta

| | | |
|---|---|---|
| 1/2 | a | 0 |
| 1/4 | b | 10 |
| 1/8 | c | 101 |
| 1/8 | d | 111 |

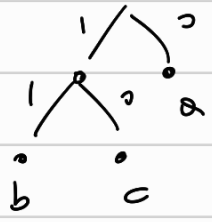
a b a c d = 0 10 0 101 111

codifico 1 simbolo x simbolo

$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{2}{8} \cdot 3 = 1{,}75$

(1 = lungh. cod)

| | | |
|---|---|---|
| 1/2 | a | |
| 1/4 | b | |
| 1/4 | c | |




a → 0

b → 11

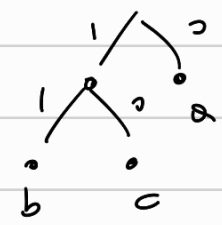
c → 10

$H = \log_2 V_x$ (n° di simboli)

$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 1{,}5$ lunghezza media

$H = 1{,}58$, posso comprimere ancora (anche se di poco) *

| | | |
|---|---|---|
| 1/3 | a | 0 |
| 1/3 | b | 11 |
| 1/3 | c | 10 |




$|-|(X) = \frac{1}{3} + \frac{2}{3} + \frac{2}{3} = \frac{5}{3} = 1{,}\overline{6}$

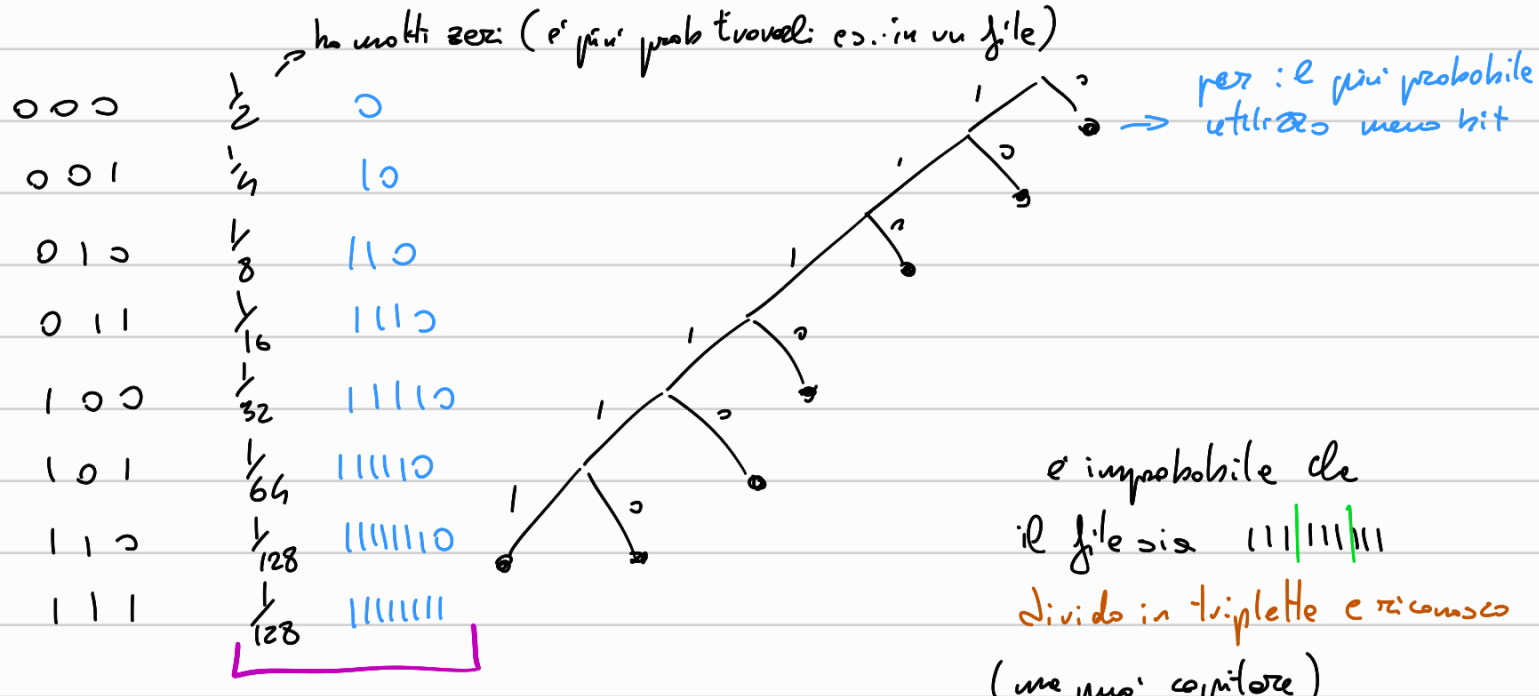
$H = 1{,}58 \longleftrightarrow \bar{L} \simeq 1{,}67$

ho un 10% di spreco, potrei comprimere ancora ma non ho potenze di due qui

Posso comprimere fino all'entropia (grezza) *

Quando ho molti simboli: (e quindi codificare un simbolo per volta)

Huffman statica




è improbabile che il file sia 111|111|111

divido in triplette e riconosco (uno più copilare)

8 simboli

3 bit

Quanti zeri ?

$$\underset{\text{zeri}}{3} \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + \frac{1}{16} + \frac{2}{32} + \frac{1}{64} + \frac{1}{128} = A$$

mi aspetto che questo numero sia > n° di uni

Perché è più probabile che io abbia più zeri ($p = \frac{1}{2}$ per 000)

$A \approx 2$ che sono più o meno il doppio degli uni

→ n° di zeri sulle codifiche : $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128}$

Debolezze di Huffman

- bisogna conoscere le probabilità a priori
- irrealistico il fatto che le probabilità siano

identicamente distribuite e che le var. casuali siano indipendenti

- Huffman si basa su scelte binarie; nel caso di sequenze di un solo simbolo, in cui l'entropia è vicina a zero (dove p di uno dei due risultati è circa 1)
  Huffman non compie bene, determinando una L media lontana dall'entropia ottenuta