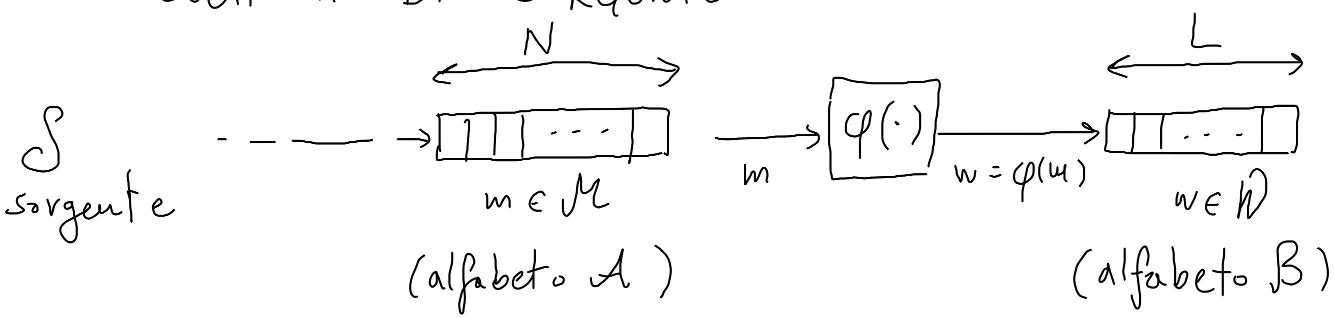
CODIFICA DI SORGENTE



Messaggi $m \in \mathcal{M}$: probabilità $q(m)$, lunghezza $n(m)$ $\longrightarrow$ valore atteso di N

$$\mathbb{E}[N] = \sum_{m \in \mathcal{M}} q(m)\, n(m)$$

Parole di codice $w \in \mathcal{W}$: probabilità $p(w)$, lunghezza $\ell(w)$ $\longrightarrow$ valore atteso di L

$$\mathbb{E}[L] = \sum_{w \in \mathcal{W}} p(w)\, \ell(w)$$

Tasso del codice : è il rapporto $\dfrac{\mathbb{E}[L]}{\mathbb{E}[N]} =: R$

Fattore di compressione : $1/R$

Nel caso di codici $\underline{B}$-B o $\underline{B}$-LV, $N = n$ costante $\longrightarrow$ $R = \dfrac{\mathbb{E}[L]}{n}$

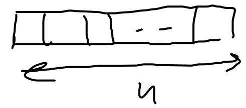
lunghezza dei blocchi della sorgente

Vogliamo minimizzare $\mathbb{E}[L]$

Problema di Codifica di sorgente: Data una sorgente S (stazionaria e senza memoria), individuare un codice

$C = (\mathcal{M}, \varphi, \mathcal{W})$ che sia u.d. e a tasso minimo (o approssimativamente minimo)

Codici B-LV. $\mathcal{M} = \mathcal{A}^n$



$(D = |\mathcal{B}|)$

Iniziamo da $\boxed{n = 1}$. In questo caso, $R = \dfrac{\mathbb{E}[L]}{n} = \mathbb{E}[L]$

Se $\mathcal{A} = \{a_1, \dots, a_K\} = \mathcal{M}$ $\qquad \mathcal{W} = \{w_1, w_2, \dots, w_K\}$

$$\underset{\text{D.d.p. della sorgente}}{} \longrightarrow p = (p_1, \dots, p_K) \qquad w_1, w_2, \dots, w_K \in \mathcal{B}^+$$

Lunghezza: $\ell_1, \ell_2, \dots, \ell_K$

$$R = \mathbb{E}[L] = \sum_{w \in \mathcal{W}} p(w)\, \ell(w) = \sum_{i=1}^{K} p_i \ell_i$$

Teo (3.13) (Teorema di Shannon). Se S è una sorgente stazionaria e senza memoria, con d.d.p. $p = (p_1, \dots, p_K)$, allora la lunghezza attesa delle parole di un <u>qualunque</u> codice D-ario u.d. è almeno pari all'entropia D-aria della sorgente:

$$\mathbb{E}[L] \geq H_D(p) = -\sum_{i=1}^{K} p_i \log_D p_i = \frac{H(p)}{\log_2 D}$$

Entropia D-aria

Dim. Consideriamo la d.d.p. $q = (q_1, \dots, q_K)$: $q_i \triangleq \dfrac{D^{-l_i}}{\underbrace{\sum_{j=1}^{K} D^{-l_j}}_{\alpha}} > 0$

$\boxed{\alpha = \sum_{j=1}^{K} D^{-l_j}}$; $q_i = \dfrac{D^{-l_i}}{\alpha}$

Scriviamo la divergenza informazionale tra p e q:

(Gibbs)

$$0 \le \frac{D(p\|q)}{\log_2 D} = \sum_{i=1}^{K} p_i \log_D\left(\frac{p_i}{q_i}\right) = \sum_i p_i \log_D p_i - \sum_i p_i \log_D\left(\frac{D^{-l_i}}{\alpha}\right) =$$

(H_0 "=" se e solo se $p = q$)

$$= -H_D(p) - \sum_i p_i \underbrace{\log_D(D^{-l_i})}_{-l_i} + \sum_i p_i \log_D \alpha = -H_D(p) + \underbrace{\sum_i p_i l_i}_{\mathbb{E}[L]} + \log_D \alpha$$

Per McMillan-Kraft, $\alpha \le 1 \Rightarrow \log_D \alpha \le 0$ (è = se $\alpha = 1$).

$$\sum_i p_i l_i \geq H_D(p) \underbrace{- \log_D \alpha}_{\geq 0} \Rightarrow \mathbb{E}[L] \geq H_D(p). \quad \text{QED.}$$

Abbiamo uguaglianza : $\mathbb{E}[L] = H_D(p)$

se e solo se : $p = q$ e $\alpha = 1$

ovvero : $p_i = D^{-l_i}$ per ogni $i = 1, \dots, K$

(distribuzione D-adica)

Esempio :

(a_1, a_2, a_3, a_4)

$p = (1/2, 1/4, 1/8, 1/8)$ $D = 2$

$2^{-1}, 2^{-2}, 2^{-3}, 2^{-3}$

$l_1 = 1, l_2 = 2, l_3 = 3, l_4 = 3$ $\rightarrow$ Ho $\mathbb{E}[L] = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1 + \frac{3}{4} = \frac{7}{4}$

e $H_D(p) = 7/4$.

Se p è uniforme, cioè $p_i = 1/K$, per $i = 1, 2, \dots, K$, allora $\mathbb{E}[L] \geq H_D(p) = \log_D K$

In particolare, se $D = K$, $\log_D K = 1$ $\rightarrow$ non posso sperare in un fattore di compressione ≥ 1

La differenza $r \triangleq \mathbb{E}[L] - H_D(p) \geq 0$

è detta ridondanza del codice.

Esistono codici con un tasso prossima all'entropia D-aria?

Se riuscissi ad avere $l_i \gtrsim -\log_D p_i$

allora $\mathbb{E}[L] \gtrsim -\sum_i p_i \log_D p_i = H_D(p)$ e $r \simeq 0$

Un approccio è di porre $l_i = \lceil -\log_D p_i \rceil$ ← parte intera superiore

(codifica di Shannon-Fano).

"Parte positiva"

Teo (3.14). (Teorema di codifica di sorgente per codici B-LV simbolici)

Se S è una sorgente con d.d.p. p, esiste un codice D-ario a prefisso

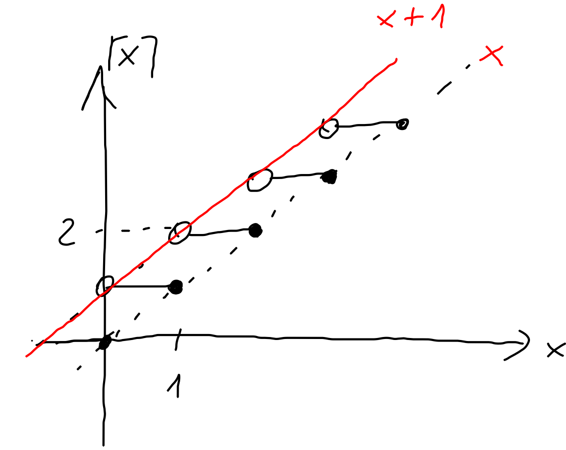
tale che $\mathbb{E}[L] < H_D(p) + 1$.

Dim. Definisco $l_i = \lceil -\log_D p_i \rceil$. Tali lunghezze soddisfano McMillan-Kraft, in quanto:

$$\sum_{i=1}^{k} D^{-\lceil -\log_D p_i \rceil} \le 1 \quad \checkmark$$

$\lceil x \rceil \geq x \longrightarrow$

$$\sum_{i=1}^{k} D^{-\lceil -\log_D p_i \rceil} \le \sum_{i=1}^{k} D^{\log_D p_i} = \sum_{i=1}^{k} p_i = 1$$



$x \le \lceil x \rceil < 1 + x$

$\Rightarrow$ Quindi esiste un codice a prefisso (quindi anche u.d.) che ha parole di codice di lunghezza $l_1, l_2, \dots, l_k$.

$$E[L] = \sum_{i=1}^{k} p_i l_i \overset{\lceil x \rceil < 1+x}{<} \sum_{i=1}^{k} p_i (1 - \log_D p_i) = \overbrace{\sum_{i=1}^{k} p_i}^{=1} - \sum_{i=1}^{k} p_i \log_D p_i$$

$$= 1 + H_D(p).$$

QED.

Esempio in cui $\mathbb{E}[L]$ è arbitrariamente vicino a $H_D(p)+1$:

$\mathcal{A} = \{0,1\}$, $p = (\overset{p(0)}{\varepsilon}, \overset{p(1)}{1-\varepsilon})$ per qualche $\varepsilon \in (0,1)$

La lunghezza media $\mathbb{E}[L]$ è necessariamente ≥ 1. (sia l_1 che l_2 sono ≥ 1)

$$H_D(p) = -\varepsilon \log_2 \varepsilon - (1-\varepsilon)\log_2(1-\varepsilon) = h_2(\varepsilon) \xrightarrow{\varepsilon \to 0} 0$$

$$\underbrace{\mathbb{E}[L]}_{\approx 1} \approx \underbrace{H_D(p)}_{\approx 0} + 1 \quad ; \quad H_D(p) + 1 - \mathbb{E}[L] \xrightarrow{\varepsilon \to 0} 0 .$$