Teoria dell'Informazione/RTI **Anno Accademico 2022–2023**

Lezioni 2 e 3

Ugo Vaccaro

Per i nostri scopi, una sorgente di informazione è una sequenza di variabili casuali $X_1, X_2, \ldots, X_i, \ldots$ le quali assumono valori in un di un insieme finito X (che chiameremo alfabeto sorgente).

Esempio 1 *Ad esempio,* $\mathsf{X} = \{a, b, c, \ldots, z\}$ *potrebbe essere l'alfabeto della lingua italiana e potremmo porre,* $\forall x \in \mathsf{X}$ *e* $\forall i \geq 1$ $P\{X_i = x\}$ *=probabilità di occorrenza della lettera x nella lingua italiana. Consultando la letteratura, scopriremmo che in tal caso avremmo* $P\{X_i = a\} \approx 11.74/100$*,* $P\{X_i = b\} \approx 0.92/100$*,* $P\{X_i = c\} \approx 4.50/100$*,* $P\{X_i = d\} \approx 3.73/100$*,* $P\{X_i = e\} \approx 11.79/100$*,* $P\{X_i = f\} \approx 0.95/100$*,* $P\{X_i = g\} \approx 1.64/100$*,* $P\{X_i = h\} \approx 1.54/100$*,* $P\{X_i = i\} \approx 11.28/100$*,* $+P\{X_i = \ell\} \approx 6.51/100$*,* $P\{X_i = m\} \approx 2.51/100$*,* $P\{X_i = n\} \approx 6.88/100$*,* $P\{X_i = o\} \approx 9.83/100$*,* $P\{X_i = p\} \approx 3.05/100$*,* $P\{X_i = q\} \approx 0.51/100$*,* $P\{X_i = r\} \approx 6.73/100$*,* $P\{X_i = s\} \approx 4.98/100$*,* $P\{X_i = t\} \approx 5.62/100$*,* $P\{X_i = u\} \approx 3.01/100$*,* $P\{X_i = v\} \approx 2.10/100$*,* $P\{X_i = z\} \approx 0.49/100$*, dove il segno* $\approx$ *è da intendersi che i valori prima scritti sono in realtà le frequenze di apparizione dei caratteri della lingua italiana, ma ai fini pratici tali frequenze possono essere considerate, sotto opportune ipotesi, buone approssimazioni delle relative probabilità.*

Effettueremo l'assunzione simplificatrice che le variabili casuali X_i siano indipendenti, ovvero che $\forall k \geq 1$ e $\forall \mathbf{x} = x_1 x_2 \ldots x_k \in \mathsf{X}^k$ valga che

$$P\{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\} = \prod_{i=1}^{k} P\{X_i = x_i\},$$

dove X^k =insieme di tutte le sequenze di lunghezza k che si possono costruire sull'alfabeto X. In più, assumeremo che tutte le X_i abbiano la stessa distribuzione P, ovvero

$$\forall x \in X_i \text{ e } \forall i \geq 1 \text{ vale che } P\{X_i = x\} = P(x).$$

In gergo, la sorgente $X_1, X_2, \ldots, X_i, \ldots$ in cui le X_i sono tra di loro indipendenti e hanno tutte la stessa distribuzione $P = \{P(x_1), \ldots, P(x_m)\}$ (dove abbiamo implicitamente assunto che $\mathsf{X} = \{x_1, \ldots x_m\}$) viene chiamata *sorgente discreta, stazionaria e senza memoria.* In queste ipotesi, useremo la seguente compatta notazione per descrivere la sorgente:

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ P(x_1) & P(x_2) & \cdots & P(x_m) \end{pmatrix}$$

e quindi $X_i = X$ per ogni $i \geq 1$. A meno di menzione esplicita, tutte le sorgenti che studieremo in questo corso sono da intendersi discrete, stazionarie e senza memoria, ed useremo l'acronimo DSSS per intendere ciò.

Il primo problema che vogliamo studiare può essere informalmente descritto nel modo seguente. Desideriamo memorizzare le sequenze emesse da una sorgente su di un supporto che permette la scrittura e lettura di simboli binari (ovvero, 0 e 1). Per ovvi motivi, vorremmo poter recuperare, dalla lettura del binario, qual è la sequenza originaria emessa dalla sorgente. In più, vorremmo usare, per ciascuna sequenza sorgente, il *minor* numero di 0 e 1 (bits) possibili. Un pò più formalmente, cerchiamo una funzione di codifica $\mathsf{c} : \mathsf{X}^k \to \{0,1\}^n$ ed una funzione di decodifica $\mathsf{d} : \{0,1\}^n \to \mathsf{X}^k$ (dove $\{0,1\}^n$ è l'insieme delle sequenze binarie di una data lunghezza n) che abbia le seguenti caratteristiche

i) n sia il più piccolo possibile (e quindi il rapporto n/k pari al numero di bits usati per ogni lettera sorgente sia il più piccolo possibile)

ii) $P\{\mathbf{x} : \mathbf{x} \in \mathsf{X}^k$ e per cui valga $\mathsf{d}(\mathsf{c}(\mathbf{x})) \neq \mathbf{x}\}$ sia "molto piccolo".

Date funzioni di codifica $\mathsf{c} : \mathsf{X}^k \to \{0,1\}^n$ e decodifica $\mathsf{d} : \{0,1\}^n \to \mathsf{X}^k$, e sorgente X, denotiamo con $e(\mathsf{d},\mathsf{c})$ la quantità

$$e(\mathsf{d},\mathsf{c}) = P\{\mathbf{x} : \mathbf{x} \in \mathsf{X}^k \text{ e per cui valga } \mathsf{d}(\mathsf{c}(\mathbf{x})) \neq \mathbf{x}\}.$$

Chiaramente, $e(\mathsf{d},\mathsf{c})$ denota la probabilità che la sorgente emetta una qualche sequenza su cui le nostre regole di codifica e decodifica sbagliano. Chiameremo quindi $e(\mathsf{d},\mathsf{c})$ la probabilità di errore dello schema di codifica/decodifica. Inoltre, dati $k =$ lunghezza delle sequenze sorgenti che vogliamo codificare, ed $\epsilon =$probabilità di errore che vogliamo tollerare, poniamo $n(k,\epsilon)$ pari al più *piccolo* valore di n per cui possiamo trovare una funzione di codifica $\mathsf{c} : \mathsf{X}^k \to \{0,1\}^n$ ed una funzione di decodifica $\mathsf{d} : \{0,1\}^n \to \mathsf{X}^k$ che abbiano $e(\mathsf{d},\mathsf{c}) \leq \epsilon$. Vogliamo determinare il valore $n(k,\epsilon)/k$ quando k è "molto grande". Tale valore $n(k,\epsilon)/k$ rappresenterà quindi il *minimo* numero di bits per simbolo sorgente che possiamo usare sotto la condizione checi garantisce l'esistenza di una codifica e decodifica con probabilità di errore inferiore ad ϵ. Sussiste il seguente risultato.

Teorema 1 *Per una generica sorgente $X_1, X_2, \ldots, X_i, \ldots$ discreta, stazionaria e senza memoria con distribuzione $P = \{P(x_1), \ldots, P(x_m)\}$, vale che*

$$\lim_{k\to\infty} \frac{n(k,\epsilon)}{k} = \sum_{x\in\mathsf{X}} P(x) \log \frac{1}{P(x)}.$$

La quantità $\sum_{x\in\mathsf{X}} P(x) \log \frac{1}{P(x)}$ verrà denotata con $H(P)$ e verrà chiamata l'entropia della sorgente (ovvero, della variabile casuale X con distribuzione $P = \{P(x_1), \ldots, P(x_m)\}$.

Per la prova del Teorema, osserviamo che l'esistenza di una funzione di codifica $\mathsf{c} : \mathsf{X}^k \to \{0,1\}^n$ e decodifica $\mathsf{d} : \{0,1\}^n \to \mathsf{X}^k$ che abbiano probabilità di errore $e(\mathsf{d},\mathsf{c}) \leq \epsilon$ è ovviamente equivalente a trovare un sottoinsieme $A \subset \mathsf{X}^k$ di sequenze di lunghezza k tale che

$$P(A) = \sum_{\mathbf{x}\in A} P(\mathbf{x}) \geq 1 - \epsilon \qquad \text{e} \qquad |A| \leq 2^n. \tag{1}$$

Quindi, informalmente, l'enunciato del Teorema equivale a mostrare che per un $n = n(k,\epsilon) \approx kH(P)$ esiste un $A \subset \mathsf{X}^k$ per cui

$$P(A) \geq 1 - \epsilon \quad \text{e} \quad |A| \leq 2^{kH(P)}$$

e, viceversa, per ogni sottoinsieme di sequenze di probabilità almeno $1 - \epsilon$, la sua cardinalità è, in prima approssimazione, pari almeno a $2^{kH(P)}$.

Andiamo alla ricerca di un siffatto insieme A. Fissati k e $\delta > 0$, definiamo

$$B(k,\delta) = \{\mathbf{x} \in \mathsf{X}^k : 2^{-k(H(P)+\delta)} \leq P(\mathbf{x}) \leq 2^{-k(H(P)-\delta)}\} \tag{2}$$

Per l'insieme $B(k,\delta)$ valgono le seguenti uguaglianze

$$B(k,\delta) = \{\mathbf{x} \in \mathsf{X}^k : k(H(P)+\delta) \geq \log \frac{1}{P(\mathbf{x})} \geq k(H(P)-\delta)\} \quad (\text{poichè } \log \frac{1}{x} = -\log x) \tag{3}$$

$$= \left\{\mathbf{x} \in \mathsf{X}^k : k(H(P)+\delta) \geq \log \frac{1}{\prod_{i=1}^k P(x_i)} \geq k(H(P)-\delta)\right\} \quad (\text{poichè la sorgente è DSSS}) \tag{4}$$

$$= \left\{\mathbf{x} \in \mathsf{X}^k : k(H(P)+\delta) \geq \sum_{i=1}^k \log \frac{1}{P(x_i)} \geq k(H(P)-\delta)\right\} \quad (\text{poichè } \log a \times b = \log a + \log b) \tag{5}$$

Consideriamo adesso, per ogni $i = 1, \ldots, k$ la variabile casuale

$$Y_i = \begin{pmatrix} \log \frac{1}{P(x_1)} & \log \frac{1}{P(x_2)} & \cdots & \log \frac{1}{P(x_m)} \\ \\ P(x_1) & P(x_2) & \cdots & P(x_m) \end{pmatrix}$$

dove, ricordiamo $\mathsf{X} = \{x_1, \ldots x_m\}$. Per tale variabile casuale vale ovviamente che il suo valor medio $E[Y_i]$ è pari a

$$E[Y_i] = \sum_{x \in \mathsf{X}} P(x) \log \frac{1}{P(x)} = H(P).$$

Continuando dalla (5) otteniamo

$$B(k, \delta) = \left\{ \mathbf{x} \in \mathsf{X}^k : \left| \frac{1}{k} \sum_{i=1}^{k} \log \frac{1}{P(x_i)} - H(P) \right| \leq \delta \right\} \quad \text{(dalla definizione di valore assoluto)} \tag{6}$$

$$= \left\{ \mathbf{x} \in \mathsf{X}^k : \left| \frac{1}{k} \sum_{i=1}^{k} Y_i - H(P) \right| \leq \delta \right\} \quad \text{(dalla definizione di } Y_i\text{)} \tag{7}$$

Per la legge debole dei grandi numeri sappiamo che

$$\lim_{k \to \infty} P \left\{ \mathbf{x} : \left| \frac{1}{k} \sum_{i=1}^{k} Y_i - H(P) \right| \leq \delta \right\} = 1 \qquad \forall \delta > 0 \tag{8}$$

ovvero che $\lim_{k \to \infty} P(B(k, \delta)) = 1$. Pertanto, sappiamo che $\forall \epsilon > 0 \; \exists k$ tale che $P(B(k, \delta)) \geq 1 - \epsilon$.

Il primo passo lo abbiamo completato, ovvero abbiamo trovato un insieme la cui probabilità è $\geq 1 - \epsilon$. Speriamo ora che la sua cardinalità non sia troppo grande. Osserviamo a tal proposito che

$$1 \geq P(B(k, \delta)) = \sum_{\mathbf{x} \in B(k, \delta)} P(\mathbf{x}) \geq \sum_{\mathbf{x} \in B(k, \delta)} 2^{-k(H(P) + \delta)} \qquad \text{(dalla (2))} \tag{9}$$

$$= |B(k, \delta)| \times 2^{-k(H(P) + \delta)} \tag{10}$$

da cui otteniamo che

$$|B(k, \delta)| \leq 2^{k(H(P) + \delta)}.$$

Riassumendo, abbiamo trovato un insieme $B(k, \delta)$ che soddisfa le seguenti due interessanti proprietà

1. $P(B(k, \delta)) \geq 1 - \epsilon$
2. $|B(k, \delta)| \leq 2^{k(H(P) + \delta)}$.

Se adesso fissiamo $n = \lceil k(H(P) + \delta) \rceil$, il numero di sequenze binarie di lunghezza n (pari a 2^n) sarà *almeno* pari a $2^{k(H(P) + \delta)} \geq |B(k, \delta)|$. Quindi ne potremmo assegnare una *distinta* a ciascuna sequenza $\mathbf{x} \in B(k, \delta)$, e quindi tale funziona di decodifica sarà chiaramente iniettiva e quindi invertibile (da cui la decodifica). E per le restanti sequenze $\mathbf{x} \in \mathsf{X}^k \setminus B(k, \delta)$? Per esse non ce ne importa nulla, in quanto esse avranno *in totale* probabilità al più ϵ.

Ricordiamo che avevamo definito $n(k, \epsilon)$ come il più piccolo valore di n per cui possiamo trovare una funzione di codifica $\mathsf{c} : \mathsf{X}^k \to \{0, 1\}^n$ ed una funzione di decodifica $\mathsf{d} : \{0, 1\}^n \to \mathsf{X}^k$ che abbiano $e(\mathsf{d}, \mathsf{c}) \leq \epsilon$. Avendo

visto che per $n = \lceil k(H(P)+\delta)\rceil$ è possibile trovare una tale coppia di funzioni c e d, ne possiamo concludere che $n(k,\epsilon) \leq \lceil k(H(P)+\delta)\rceil$, ovvero che $\frac{n(k,\epsilon)}{k} \leq \frac{\lceil k(H(P)+\delta)\rceil}{k}$. Pertanto

$$\limsup_{k\to\infty} \frac{n(k,\epsilon)}{k} \leq (H(P)+\delta),$$

e la prima parte del Teorema è dimostrato.

Per dimostrare la seconda parte, dobbiamo far vedere che

$$\liminf_{k\to\infty} \frac{n(k,\epsilon)}{k} \geq (H(P)-\delta),$$

ovvero, alla luce di quanto detto prima, *ogni* sottoinsieme $A \subset \mathsf{X}^k$ che abbia $P(A) \geq 1-\epsilon$, non può avere un numero di elementi significativamente inferiore a $2^{kH(P)}$.

Aprendo una parentesi, possiamo facilmente osservare che per insiemi arbitrari A e B, con $P(A) \geq 1-\epsilon$ e $P(B) \geq 1-\epsilon$, vale che $P(A\cap B) \geq 1-2\epsilon$. Notiamo innanzitutto che $P(A) \geq 1-\epsilon$ e $P(B) \geq 1-\epsilon$ implica che $P(\overline{A}) \leq \epsilon$ e $P(\overline{B}) \leq \epsilon$, dove con $\overline{A}$ e $\overline{B}$ abbiamo denotato, rispetivamente, il complemento dell'insieme A ed il complemento dell'insieme B. Otteniamo quindi che

$$P(\overline{A}\cup\overline{B}) \leq P(\overline{A}) + P(\overline{B}) \leq 2\epsilon,$$

da cui

$$P(\overline{\overline{A}\cup\overline{B}}) \geq 1-2\epsilon.$$

D'altra parte, vale ovviamente che

$$P(\overline{\overline{A}\cup\overline{B}}) = P(A\cap B),$$

ed quindi $P(A\cap B) \geq 1-2\epsilon$, come speravamo.

Sia ora un generico $A \subset \mathsf{X}^k$ che abbia $P(A) \geq 1-\epsilon$. Da quanto prima osservato abbiamo

$$\begin{aligned} 1-2\epsilon \leq P(A\cap B(k,\delta)) &= \sum_{\mathbf{x}\in A\cap B(k,\delta)} P(\mathbf{x}) \leq \sum_{\mathbf{x}\in A\cap B(k,\delta)} 2^{-k(H(P)-\delta)} \qquad \text{(dalla (2)} \\ &\leq |A\cap B(k,\delta)| \times 2^{-k(H(P)-\delta)} \leq |A| \times 2^{-k(H(P)-\delta)} \end{aligned}$$

da cui otteniamo che

$$|A| \geq (1-2\epsilon)\times 2^{k(H(P)-\delta)}.$$

Riassumendo, *per ogni* insieme A per cui probabilità $P(A) \geq 1-\epsilon$, (in particolare, per il più piccolo tale insieme A) che ci garantisce quindi probabilità di errore $\leq \epsilon$ qualora forniamo una codifica distinta ad ogni elemento in A, ci occorrono *almeno* $(1-2\epsilon)\times 2^{k(H(P)-\delta)}$ sequenze binarie. Ovvero, la lunghezza della codifica dovrà essere almeno pari a $\log\left((1-2\epsilon)\times 2^{k(H(P)-\delta)}\right)$. Quindi, varrà

$$\frac{n(k,\epsilon)}{k} \geq \frac{\log(1-2\epsilon)}{k} + (H(P)-\delta),$$

da cui

$$\liminf_{k\to\infty} \frac{n(k,\epsilon)}{k} \geq (H(P)-\delta),$$

il che è quanto occorreva provare per completare la dimostrazione del Teorema. □

Una delle conseguenze del Teorema appena dimostrato è che all'interno dell'insieme X^k di tutte le sequenze di lunghezza k costruibili sull'alfabeto X (che sono in numero pari a $|\mathsf{X}|^k$) ve ne è uno di cardinalità $\approx 2^{kH(P)}$ (che,

in generale è $\ll |\mathsf{X}|^k)$ e di probalità totale $\geq 1-\epsilon$. E non solo. In tale insieme, tutte le sequenze sono quasi equiprobabili, nel senso che la loro probabilità è $\approx 2^{kH(P)}$.

Vediamo un esempio. Supponiamo di avere la seguente sorgente

$$X = \begin{pmatrix} a & b \\ 0.2 & 0.8 \end{pmatrix}$$

Vogliamo codificare sequenze sorgenti di lunghezza $k = 25$. Se insistessimo a voler codifiche e decodifiche con probabilità di decodifica errata pari a 0, dovremmo necessariamente usare sequenze sorgenti di lunghezza n per cui $2^n \geq 2^{25}$, ovvero $n = 25$ e quindi un bit per ogni simbolo sorgente.

Decidiamo allora di tollerare una probabilità di errore al più pari a $\epsilon = 0.03$. Ricordiamo che una generica sequenza lunga 25 emessa da X con numero di a pari a t e numero di b pari a $25-t$ ha probabilità di essere emessa uguale a $(0.2)^t \times (0.8)^{25-t}$. Inoltre, l'insieme di *tutte* le sequenze sorgenti di lunghezza 25 con numero di a pari a t e numero di b pari a $25-t$ ha probabiliità pari a

$$\binom{25}{t}(0.2)^t \times (0.8)^{25-t}.$$

L'entropia di X è pari a

$$H(X) = -0.2\log(0.2) - (0.8)\log(0.8) \approx 0.721928....$$

In accordo al Teorema prima visto, calcoliamo

$$B = \{\text{tutte le sequenze } \mathbf{x} : H(X) - \delta \leq -\frac{1}{25}\log p(\mathbf{x}) \leq H(X) + \delta\}$$

dove scegliamo $\delta = 0.2$. Quindi

$$B = \{\mathbf{x} : 0.521928 \leq -\frac{1}{25}\log\left((0.2)^t \times (0.8)^{25-t}\right) \leq 0.921928\}.$$

Risolvendo per t, scopriamo che $3 \leq t \leq 7$. Quindi

$$\Pr\{B\} = \sum_{t=3}^{7}\binom{25}{t}(0.2)^t \times (0.8)^{25-t} = 0.9793... > 1-\epsilon = 1-0.003.$$

Inoltre, vale che

$$|B| = \sum_{t=3}^{7}\binom{25}{t} = 725880.$$

Ora, $\lceil \log_2 |B| \rceil = 20$, per cui ci basterà usare 20 bit per codificare, in maniera distinta, ogni possibile sequenza in B. Il numero di bits per simbolo sorgente che useremmo è quindi $20/25 = 0.8$ (meno di 1, ed abbastanza prossimo all'entropia della sorgente, se aumentassimo k potremmo avvicinarci ancora di più). Osserviamo anche che, in questo esempio, $|B|/2^{25} \approx 0.0216...$, quindi stiamo codificando solo una frazione pari a 0.0216 del totale di tutte le possibili 2^{25} sequenze sorgenti di lunghezza 25, e ciononostante riusciamo ad avere una probabilità di decodifica errata < 0.03.

Un'ultima osservazione. Supponiamo che l'alfabeto della sorgente sia esso stesso binario (cioè pari a $\{0,1\}$. In tale caso il problema che abbiamo affrontato (e risolto!) è essenzialmente il seguente. Vorremmo trovare una funzione di codifica $\mathsf{c} : \{0,1\}^k \to \{0,1\}^n$, con $k > n$, che sia "invertibile" (cioè decodificabile). Ciò è ovviamente impossibile in quanto $\{0,1\}^k = 2^k > 2^n = \{0,1\}^n$ e quindi *necessariamente* degli errori accadranno nella inversione di c, ovvero nella decodifica. Quello che abbiamo scoperto è che per $n \approx kH(P)$ (e tale numero è in generale inferiore a k in quanto $H(P) = 1$ solo nel caso in cui le probabilità di emissione di 0 e 1 sono entrambe pari a $1/2$) possiamo

costruire una inversa di c che funziona "con alta probabilità", ovvero che sbaglia solo di di un sottoinsieme di $\{0,1\}^k$ che, benchè possibilmente grande, ha una probabilità totale inferiore a ϵ, con ϵ piccolo a piacere.

Il risultato prima visto (ovvero il Teorema 1), non è molto utile in pratica, in quanto la costruzione dell'insieme $B(k,\delta)$ richiede l'esame di un numero esponenziale di sequenze. Ciò non è fattibile in pratica. Per dare una soluzione a tale problema, è prima necessario studiare più in dettaglio le proprietà matematiche della funzione entropia $H(X)$ di una variabile casuale X. Ricordiamo innanzitutto il concetto di concavità e convessità di funzioni reali di una variabile reale.

Definizione 1 *Una funzione $f : \mathbb{R} \to \mathbb{R}$ è detta convessa sull'intervallo $(a,b) \subset \mathbb{R}$ se*

$$\forall x_1, x_2 \in (a,b) \;\; e \;\; \forall\, 0 \le \lambda \le 1 \;\; vale\ che \;\; f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

f è detta strettamente convessa su (a,b) se l'uguaglianza vale solo per $\lambda = 0$ e $\lambda = 1$.

Intuitivamente, una funzione è convessa se il grafico della funzione è sempre *sotto* la corda $\lambda f(x_1) + (1-\lambda) f(x_2)$ che unisce i punti x_1 e x_2. Equivalentemente, f è convessa se in ogni punto x_0 di (a,b) la funzione f è *sopra* la tangente passante per x_0, ovverosia se la seguente diseguaglianza vale:

$$\forall x \in (a,b) \quad f(x) \ge \alpha(x - x_0) + f(x_0). \tag{11}$$

Una funzione $f : \mathbb{R} \to \mathbb{R}$ è detta concava sull'intervallo $(a,b) \subset \mathbb{R}$ se $-f$ è convessa. Vale il seguente risultato.

Teorema 2 *Se la funzione f ha derivata seconda che è ≥ 0 (ovvero > 0 sull'intervallo (a,b), allora f è convessa (ovvero strettamente convessa).*

Dall'ultimo Teorema discende, ad esempio, che le funzioni x^2, e^x, $x \log x$ sono strettamente convesse per $x \ge 0$, mentre $\log x$, $\sqrt{x}$ sono strettamente concave per $x \ge 0$.

Ricordiamo ora che se abbiamo una variabile casuale X che assume valori in un insieme $\mathcal{X}$ in accordo alla distribuzione di probabilità $P = \{P(x) : x \in \mathcal{X}\}$, il suo valor medio $E[X]$ è definito come $E[X] = \sum_{x \in \mathcal{X}} x P(x)$. E se abbiamo una funzione $f : \mathbb{R} \to \mathbb{R}$ allora f ed X naturalmente definiscono una nuova variabile casuale $f(X)$ di valor medio $E[f(X)] = \sum f(x) P(x)$. Sussiste il seguente importante risultato, noto sotto il nome di *Diseguaglianza di Jensen.*

Teorema 3 *Se f è convessa e X è una variabile casuale, allora vale che*

$$E[f(X)] \ge f(E[X]). \tag{12}$$

Inoltre, se f è strettamente convessa, allora la diseguaglianza vale con il segno di eguaglianza se e solo se la variabile casuale è concentrata in un punto x_0, ovvero assume sempre lo stesso valore x_0.

Dimostrazione. Sia $x_0 = E[X]$ ed applichiamo la formula (11). Otteniamo che

$$f(X) \ge \alpha(X - E[X]) + f(E[X]).$$

Ricordando le proprietà della media di una variabile casuale, otteniamo che

$$E[f(X)] \ge \alpha(E[X] - E[X]) + f(E[X]) = f(E[X]).$$

É inoltre chiaro che se la f fosse strettamente convessa la diseguaglianza è soddisfatta con uguaglianza solo nel punto di contatto tra la tangente $\alpha(x - x_0) + f(x_0)$ e la funzione, ovvero in $x_0 = E[X]$, da cui segue che se in (12) vale l'uguaglianza, allora $X = E[X]$. □

Proviamo ora il seguente risultato.

Corollario 1 *Sia* $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ P(x_1) & P(x_2) & \cdots & P(x_m) \end{pmatrix}$ *una variabile casuale. Vale che*

$$0 \leq H(X) = \sum_{i=1}^{m} P(x_i) \log \frac{1}{P(x_i)} \leq \log m.$$

Dimostrazione. La limitazione inferiore $0 \leq H(X)$ è ovvia, una volta che si osservi che l'argomento del logaritmo è sempre ≥ 1, per cui ogni termine nella sommatoria di $H(X)$ è ≥ 0. Per provare la limitazione superiore, useremo la diseguaglianza di Jensen, applicata alla funzione concava $f(x) = -\log x$ ed alla variabile casuale

$$Y = \begin{pmatrix} \frac{1}{P(x_1)} & \frac{1}{P(x_2)} & \cdots & \frac{1}{P(x_m)} \\ P(x_1) & P(x_2) & \cdots & P(x_m) \end{pmatrix}.$$

Si ha

$$H(X) = \sum_{i=1}^{m} P(x_i) \log \frac{1}{P(x_i)} = E[f(Y)] \leq f(E[Y]) = \log \sum_{i=1}^{m} P(x_i) \frac{1}{P(x_i)} = \log m.$$

Poichè la funzione $f(x) = \log x$ è strettamente concava, ne discende anche che l'uguaglianza $H(X) = \log m$ vale se e solo se $P(x_i) = 1/m$, per ogni $i = 1, \ldots, m$. □

Il seguente esempio dovrebbe ulteriormente cementare la nostra convinzione che l'entropia di una variabile casuale X è una corretta misura dell'incertezza che a priori abbiamo sui valori assunti da X.

Esempio 2 *Sia* X *una variabile casuale con distribuzione di probabilità* $\mathbf{p} = (p_1, \ldots, p_n)$. *Supponiamo che* $\exists i, j \in \{1, \ldots, n\}$ *tale che* $p_i > p_j$. *Definiamo la nuova variabile casuale* Y *avente distribuzione di probabilità* $\mathbf{q} = (q_1, \ldots, q_n)$ *così definita*

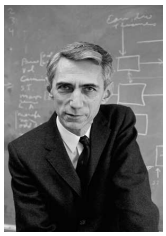
$$q_k = \begin{cases} p_k & \text{per } k \neq i, j \\ \frac{p_i + p_j}{2} & \text{per } k = i, j. \end{cases}$$

Per motivi evidenti, il nostro livello di incertezza su Y *è maggiore di quello che abbiamo su* X. *Infatti, per* Y *i valori* i*-esimo e* j*-esimo hanno uguale probabilità di accadere, e quindi nulla possiamo dire su di essi, mentre per* X *sappiamo che il valore* i*-esimo ha probabilità di occorrere maggiore della probabilità di occorrere dell'evento* j*-esimo, e questo dovrebbe descrescere il nostro livello di incertezza. Vediamo se ciò è riflesso anche dall'entropia,*

ovvero se è vero che l'entropia di Y è maggiore di quella di X. Calcoliamo

$$
\begin{aligned}
H(X) - H(Y) &= \sum_{k=1}^{n} p_k \log \frac{1}{p_i} - \sum_{i=k}^{n} q_i \log \frac{1}{q_k} \\
&= p_i \log \frac{1}{p_i} + p_j \log \frac{1}{p_j} - q_i \log \frac{1}{q_i} - q_j \log \frac{1}{q_j} \\
&= p_i \log \frac{1}{p_i} + p_j \log \frac{1}{p_j} - \frac{(p_i + p_j)}{2} \log \frac{2}{(p_i + p_j)} - \frac{(p_i + p_j)}{2} \log \frac{2}{(p_i + p_j)} \\
&= p_i \log \frac{1}{p_i} + p_j \log \frac{1}{p_j} - (p_i + p_j) \log \frac{2}{(p_i + p_j)} \\
&= p_i \log \frac{1}{p_i} + p_j \log \frac{1}{p_j} - p_i \log \frac{2}{(p_i + p_j)} - p_j \log \frac{2}{(p_i + p_j)} \\
&= p_i \log \frac{p_i + p_j}{p_i} + p_j \log \frac{p_i + p_j}{p_j} - (p_i + p_j) \log 2 \\
&= (p_i + p_j) \left(\frac{p_i}{p_i + p_j} \log \frac{p_i + p_j}{p_i} + \frac{p_j}{p_i + p_j} \log \frac{p_i + p_j}{p_j} \right) - (p_i + p_j) \\
&< (p_i + p_j) \left(\log \left(\frac{p_i}{p_i + p_j} \frac{p_i + p_j}{p_i} + \frac{p_j}{p_i + p_j} \frac{p_i + p_j}{p_j} \right) \right) - (p_i + p_j) \quad \text{(applicando la dis. di Jensen)} \\
&= (p_i + p_j) \log 2 - (p_i + p_j) = 0
\end{aligned}
$$

Concludiamo quindi che $H(Y) > H(X)$, come ci aspettavamo.



C.E. Shannon

Una nota storica: la maggior parte dei concetti di base che studieremo nel corso furono introdotti da C.E. Shannon nel 1948. Alla pagina iniziale del corso potrete trovare dei link a pagine che illustrano alcuni aspetti di uno dei giganti del pensiero del secolo scorso.