Chapter 1

Teoria dell'informazione

Come evidenziato fin dal cap. **??**, il motivo per voler trasmettere un segnale deriva *dall'informazione* che lo stesso convoglia. Descrivere e misurare questo concetto è indissolubilmente legato alla caratterizzazione della entità che produce il segnale, indicata come *sorgente*, che può essere continua o discreta, con o senza memoria, e che si assume di tipo stazionario, ovvero invariante nel tempo. Rispetto al *testo originario* da cui queso capitolo è tratto, ci si limita a misurare l'informazione media (chiamata *entropia*) contenuta nei segnali prodotti da una sorgente, tralasciando l'argomento di come rappresentare gli stessi in una forma (chiamata *codifica di sorgente*) in grado di *ridurre* la quantità di dati da trasmettere (ovvero la banda da occupare) senza pregiudicare la qualità del messaggio, ossia senza *perdere* informazione. Al § 1.5 si determina poi il massimo tasso informativo che un canale può trasportare, ossia la sua *capacità*. Per motivi di tempo (e data l'assenza a lezione di studenti che non parlano italiano) questi contenuti non sono ancora stati tradotti. In alternativa, si veda l'ottimo articolo *The application of information theory to biochemical signaling systems* presso https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820280/.

Tipi di sorgente e quantità di informazione Una sorgente informativa può essere di natura discreta, come nel caso di un documento scritto, o continua, come nel caso di un segnale analogico, ad esempio audio e video. In entrambi i casi, considerazioni di tipo statistico conducono a *misurare* (in bit/simbolo) la quantità media di informazione presente nei messaggi prodotti, mediante la definizione di una grandezza, l'*entropia*.

Per informazione si intende quella ricevuta dalla notizia di qualcosa di cui (a priori) si conosce la probabilità, e dunque prima che l'evento si verifichi sussiste una condizione *di incertezza*. Da questo punto di vista l'informazione *ricevuta* a seguito della conoscenza dell'evento che si è verificato misura *la riduzione* dell'incertezza a suo riguardo.

1.1 Informazione di una sorgente discreta senza memoria

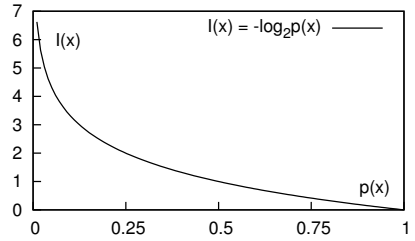
Iniziamo l'analisi considerando una sorgente discreta di informazione, che produce sequenze $x(n)$ composte da simboli x_k appartenenti ad un alfabeto di cardinalità L (ossia con $k = \{1, 2, \cdots, L\}$), e che si presentano con probabilità $p_k = Pr(x_k)$ non

dipendente da n, ovvero la sorgente è stazionaria.

Sorgente senza memoria Con questo termine si intende che i simboli vengono emessi in modo *statisticamente indipendente* (§ **??**), ovvero indicando con x_h, x_k una coppia di simboli consecutivi (ossia $x(n) = x_h$, $x(n+1) = x_k$), la probabilità del secondo non dipende dall'identità del primo, ossia $p(x_k/x_h) = p(x_k) = p_k$.

Misura dell'informazione La conoscenza di ognuno dei simboli emessi x_k apporta una quantità di informazione (espressa in *bit*) definita come[1]

$$I_k = I(x_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k \text{ bit} \quad (1.1)$$



che rappresenta il *livello di dubbio* a riguardo del verificarsi dell'evento x_k prima che questo si verifichi, ovvero di quanto possiamo ritenerci sorpresi nel venire a conoscenza dell'evento x_k, di cui riteniamo di conoscere la probabilità p_k. Osserviamo infatti che la (1.1) attribuisce un valore di informazione tanto più elevato quanto minore è la probabilità di emissione del simbolo.

La scelta di esprimere la relazione tra probabilità e informazione mediante il logaritmo in base 2 consente di verificare le seguenti osservazioni:

Prob. p_k	Inf. $-\log_2 p_k$	Commento
1	0	L'evento certo non fornisce informazione
0	∞	L'evento impossibile dà informazione infinita
$1/2$	1	Conoscere quale tra due eventi equiprobabili si sia verificato apporta un'informazione pari ad una cifra binaria (0/1) o **bit** = *binary digit*
$1/2^n$	n	Es. probabilità $1/4 \rightarrow$ due bit, $1/8 \rightarrow$ tre bit ...

Notiamo inoltre che, essendo la sorgente senza memoria, due simboli emessi consecutivamente sono statisticamente indipendenti ovvero $p(x_h x_k) = p(x_h)\, p(x_k)$ e dunque $I(x_h, x_k) = -\log_2 p_h p_k = -\log_2 p_h - \log_2 p_k = I(x_h) + I(x_k)$.

1.1.1 Entropia

Come in termodinamica al concetto di entropia si associa il grado di *disordine* in un sistema, così per una sorgente informativa l'entropia misura il livello *medio* di *casualità* dei simboli emessi. Definiamo infatti *entropia* (indicata con H) di una sorgente discreta S *il valore atteso* (§ **??**) della quantità di informazione apportata dalla conoscenza dei simboli (scelti tra L possibili) da essa generati

$$H_s = E\{I_k\} = \sum_{k=1}^{L} p_k I_k = \sum_{k=1}^{L} p_k \log_2 \frac{1}{p_k} \text{ bit/simbolo} \quad (1.2)$$

[1] Per calcolare il logaritmo in base 2, sussiste la relazione $\log_2 \alpha = \frac{\log_{10} \alpha}{\log_{10} 2} \simeq 3.322 \cdot \log_{10} \alpha$. O più in generale, $\log_2 \alpha = \frac{\log_\beta \alpha}{\log_\beta 2}$

che, pesando in probabilità il valore di informazione associato ai diversi simboli, rappresenta il *tasso medio* di informazione per simbolo delle sequenze osservabili. Come dimostriamo sotto, da tale definizione ne consegue che

- se i simboli sono *equiprobabili* ($p_k = \frac{1}{L}$ con $\forall k$) la sorgente è *massimamente informativa*, e la sua entropia è la massima possibile per un alfabeto ad L simboli, pari a $H_{s_{Max}} = \frac{1}{L}\sum_{k=1}^{L} \log_2 L = \log_2 L$ bit/simbolo;
- se i simboli non sono equiprobabili, allora $H_s < \log_2 L$;
- se la sorgente emette sempre e solo lo stesso simbolo, allora $H_s = 0$.

Questi predicati possono essere riassunti nell'espressione

$$0 \leq H_s \leq \log_2 L \tag{1.3}$$

Dimostrazione Osserviamo innanzitutto che $H_s \geq 0$ in quanto la (1.2) comprende tutti termini positivi o nulli, essendo $\log_2 \alpha \geq 0$ per $\alpha = 1/p_k \geq 1$. Mostriamo ora che $H_s - \log_2 L \leq 0$: riscriviamo innanzitutto il primo membro della diseguaglianza come

$$\begin{aligned} H_s - \log_2 L = \textstyle\sum_k p_k \log_2 \frac{1}{p_k} - \log_2 L \cdot \sum_k p_k = \\ \textstyle\sum_k p_k \left(\log_2 \frac{1}{p_k} - \log_2 L \right) = \sum_k p_k \log_2 \frac{1}{L \cdot p_k} \end{aligned} \tag{1.4}$$

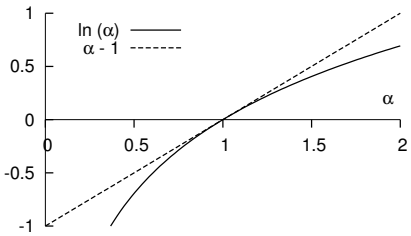
dato che $\sum_k p_k = 1$, ove le sommatorie su k si intendono da 1 ad L . Esprimiamo poi questo risultato parziale nei termini di logaritmi *naturali*, tenendo conto che $\log_2 \alpha = \frac{\ln \alpha}{\ln 2}$, ovvero

$$\sum_k p_k \log_2 \frac{1}{L \cdot p_k} = \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k} \tag{1.5}$$

A questo punto utilizziamo la relazione

$$\ln \alpha \leq \alpha - 1$$

mostrata in figura, con l'uguaglianza valida solo se $\alpha = 1$; ponendo quindi $\alpha = \frac{1}{L \cdot p_k}$ e sostituendo la (1.5) nella (1.4) si ottiene



$$\begin{aligned} H_s - \log_2 L &= \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k} \leq \frac{1}{\ln 2} \sum_k p_k \left(\frac{1}{L \cdot p_k} - 1 \right) = \\ &= \frac{1}{\ln 2} \left(\sum_k \frac{1}{L} - \sum_k p_k \right) = \frac{1}{\ln 2} (1 - 1) = 0 \end{aligned}$$

con il segno di uguale solo se $\frac{1}{L \cdot p_k} = 1$ ovvero $p_k = \frac{1}{L}$.

1.1.1.1 Entropia di sorgente binaria

Nel caso particolare di una sorgente *binaria*, ovvero che emette uno tra due simboli $\{x_0, x_1\}$ con probabilità rispettivamente $p_0 = p$, $p_1 = q = 1 - p$, la formula dell'entropia (1.2) fornisce l'espressione

$$H_b(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \text{ bit/simbolo} \tag{1.6}$$

il cui grafico è mostrato a sinistra in figura 1.1, in funzione di p.

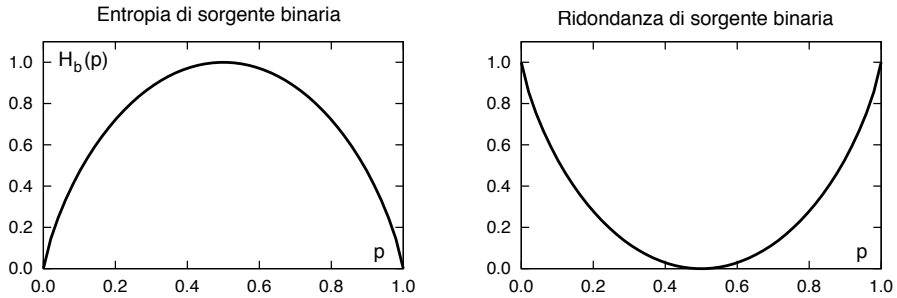


Figure 1.1: Entropia di sorgente binaria, e ridondanza associata

I due simboli $\{x_0, x_1\}$ possono essere rappresentati dalle 2 cifre binarie $\{0, 1\}$, che in questo caso chiamiamo *binit*, per non confonderli con la misura dell'informazione (il bit). Osserviamo quindi che se $p \neq 0.5$ si ottiene $H_b(p) < 1$, ossia la sorgente emette informazione con un tasso inferiore a un bit/simbolo, mentre a prima vista non potremmo usare meno di un binit per rappresentare ogni simbolo binario.

1.1.1.2 Entropia di sorgente L-aria

L'applicazione della (1.3) al caso di una sorgente che emette simboli *non* equiprobabili ed appartenenti ad un alfabeto di cardinalità L, determina per la stessa un valore di entropia $H_L < \log_2 L$ bit/simbolo.

Esempio Nel caso di una sorgente quaternaria con $p_0 = 0.5$, $p_1 = 0.25$, $p_2 = 0.125$, $p_3 = 0.125$, l'applicazione della (1.2) fornisce $H_4 = 1.75$ bit/simbolo, inferiore ai 2 bit/simbolo di una sorgente con quattro simboli equiprobabili. La relativa ridondanza è ora pari a $1 - {}^{1.75}/_{2} = 0.125$ ovvero il 12.5 %.

1.2 Sorgente discreta con memoria

Passiamo ora ad affrontare il caso in cui i simboli emessi dalla sorgente non possano essere ritenuti statisticamente indipendenti. Indicando con $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ una sequenza di n di simboli, la sua probabilità *congiunta* si calcola ora come

$$p(\mathbf{x}) = p(x_1)\, p(x_2/x_1)\, p(x_3/x_1, x_2) \ldots p(x_n/x_1, x_2, \ldots x_{n-1}) \neq \prod_{k=1}^{n} p(x_k) \qquad (1.7)$$

dato che appunto la dipendenza statistica comporta l'uso delle probabilità condizionali. L'espressione dell'entropia si modifica dunque in

$$H_n = E_{\mathbf{x}}\{I(\mathbf{x})\} = -\frac{1}{n} \sum_{\textit{tutte le possibili sequenze } \mathbf{x}} p(\mathbf{x}) \log_2 p(\mathbf{x}) \text{ bit/simbolo}$$

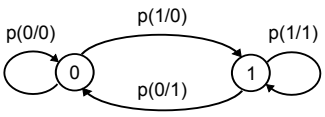
in cui $p(\mathbf{x})$ è la probabilità congiunta (1.7) di una possibile sequenza di simboli $\mathbf{x}$, e la media di insieme è effettuata su tutte le possibili sequenze $\mathbf{x}$ di lunghezza n. La grandezza H_n prende il nome di *entropia a blocco*, e si dimostra che al crescere di n il suo valore è *non crescente*, ossia $H_{n+1} \leq H_n \leq H_{n-1}$, mentre per $n \to \infty$, H_n tende ad un valore $H_\infty \leq H_s$, in cui l'uguaglianza è valida solo per sorgenti senza memoria.

1.2.1 Sorgente Markoviana

Se oltre ad un certo valore $n_{Max} = M$ la sequenza H_n non decresce più la sorgente è detta a *memoria finita* o *di Markov* di ordine M, ed è caratterizzata dal fatto che le probabilità condizionate dipendono solo dagli ultimi M simboli emessi.

Esempio Analizziamo il caso di una sorgente binaria di Markov del primo ordine, per la quale sono definite le probabilità condizionate mostrate a lato, a cui corrisponde il *diagramma di transizione* raffigurato. Essendo $M = 1$, lo *stato* della sorgente è determinato dal simbolo emesso per ultimo, che condiziona le probabilità di emissione del simbolo successivo: con i valori dell'esempio, si osserva come la sorgente *preferisca* continuare ad emettere l'ultimo simbolo prodotto, piuttosto che l'altro.

$$p(0/0) = 0.9 \quad p(1/0) = 0.1$$
$$p(0/1) = 0.4 \quad p(1/1) = 0.6$$



In pratica è come se ora vi fossero L^M diverse sorgenti S_i (nel caso dell'esempio, $2^1 = 2$), ognuna associata ad una diversa *storia passata* rappresentata dagli ultimi M simboli emessi (nell'esempio $M = 1$), identificativi dello *stato,* o *memoria,* della sorgente. In questo caso l'entropia di sorgente può essere calcolata applicando la (1.6) ad ognuno dei possibili stati, ottenendo in tal modo dei valori di *entropia condizionata* $H(x/S_i)$, mentre l'entropia di sorgente *complessiva* si ottiene come *valore atteso* dell'entropia condizionata rispetto alle probabilità di trovarsi in ognuno degli stati del modello Markoviano.

Tornando all'esempio, i valori di entropia condizionata risultano pari a

$$\begin{aligned} H(x/S_0) &= -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.47 \\ H(x/S_1) &= -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.97 \end{aligned}$$

bit/simbolo, mentre il valore della probabilità di trovarsi in uno dei due stati si ottiene risolvendo il sistema

$$\begin{cases} p(S_0) &= p(0/0)\, p(S_0) + p(0/1)\, p(S_1) \\ 1 &= p(S_0) + p(S_1) \end{cases} \tag{1.8}$$

in cui la prima equazione asserisce che la probabilità di trovarsi in S_0 è pari alla somma di quella di esserci già, per quella di emettere ancora zero, più la probabilità di aver emesso uno, ed ora emettere zero. Procedendo per sostituzione si ottiene $p(S_0) = 0.8$ e $p(S_1) = 0.2$, ossia gli stessi valori dell'esempio binario senza memoria di pag. **??**. Ma mentre in quel caso il valore dell'entropia risultava pari a 0.72 bit/simbolo, ora si ottiene

$$H = p(S_0)\, H(x/S_0) + p(S_1)\, H(x/S_1) = 0.58 \text{ bit/simbolo}$$

mostrando come la presenza di memoria aumenti la predicibilità delle sequenze emesse dalla sorgente.

Esercizio Si ripeta il calcolo dell'entropia per un modello di Markov del primo ordine, caratterizzato dalle probabilità $p(0) = p(1) = 0.5$ e $p(1/0) = p(0/1) = 0.01$, mostrando che in questo caso si ottiene una entropia di 0.08 bit/simbolo.

1.3 Contenuto informativo di sorgente continua

Sebbene l'estensione del concetto di entropia già definito per sorgenti discrete (§ 1.1.1) sia abbastanza diretto, la sua applicazione al caso di sorgenti tempo-continue presenta risvolti particolari, che andiamo a discutere.

1.3.1 Entropia differenziale di sorgente continua

L'espressione (1.2) valida per le sorgenti discrete può essere formalmente estesa al caso di una sorgente continua che produce un processo $x(t)$ stazionario ed incorrelato, descritto da una d.d.p. del primo ordine $p_x(x)$, portando all'espressione

$$h(X) = E\left\{-\log_2 p_x(x)\right\} = -\int p_x(x)\log_2 p_x(x)\,dx \tag{1.9}$$

indicata con la h minuscola per distinguerla dal caso discreto, e chiamata *entropia differenziale* a seguito delle proprietà che andiamo ad illustrare.

Dipendenza dalla dinamica Il valore ottenuto dalla (1.9) può risultare positivo, negativo o nullo, in funzione della dinamica della variabile aleatoria X.

Esempio Se calcoliamo il valore di entropia differenziale per un processo i cui valori sono descritti da una variabile aleatoria a distribuzione uniforme $p_x(x) = \frac{1}{A} rect_A(x)$, otteniamo il risultato $h(X) = -\frac{1}{A}\int_{-A/2}^{A/2} \log_2\left(\frac{1}{A}\right) dx = \log_2 A$ il cui valore effettivo, appunto, dipende dal valore di A. In particolare, se $A = 1$ si ottiene $h(X) = 0$.

L'esempio è un modo per osservare che, in presenza di un fattore di una v.a. $Y = \alpha X$ *scalata* di un fattore α, si ottenga $h(Y) = h(X) + \log_2|\alpha|$.

Invarianza rispetto alla media L'entropia differenziale non dipende dal valor medio della variabile aleatoria, ovvero è invariante rispetto alle traslazioni. Per verificare la veridicità di tale affermazione, calcolare per esercizio il valore di $h(X)$ per una d.d.p. $p_x(x) = \frac{1}{A} rect_A(x - m)$.

Confronto tra entropia di processi Essendo il valore $h(X)$ dipendente dalla dinamica della v.a., l'entropia differenziale sembra inadatta ad esprimere il contenuto informativo *assoluto* di una sorgente continua; ciononostante può comunque essere utile per confrontare due sorgenti con *uguale varianza* σ_x^2, come mostrato alla nota[2]. A

[2] In effetti esiste una misura di entropia *assoluta* per sorgenti continue, che però ha la *sgradevole caratteristica* di risultare sempre infinita. Infatti, approssimando la (1.9) come limite a cui tende una sommatoria, e suddividendo l'escursione dei valori di x in intervalli uguali Δx, possiamo scrivere

$$\begin{aligned} h_{abs}(x) &= \lim_{\Delta x\to 0} \textstyle\sum_i p(x_i)\,\Delta x \log_2 \dfrac{1}{p(x_i)\,\Delta x} \\ &= \lim_{\Delta x\to 0} \textstyle\sum_i \left[p(x_i)\,\Delta x \log_2 \dfrac{1}{p(x_i)} + p(x_i)\,\Delta x \log_2 \dfrac{1}{\Delta x}\right] = h(x) + h_0 \end{aligned}$$

in cui $h(x)$ è proprio la (1.9) mentre $h_0 = -\lim_{\Delta x\to 0}\log_2 \Delta x \int_{-\infty}^{\infty} p(x)\,dx = -\lim_{\Delta x\to 0}\log_2 \Delta x = \infty$. D'altra parte, la differenza tra le entropie assolute di due sorgenti z e x risulta pari a $h_{abs}(z) - h_{abs}(x) = h(z) - h(x) + h_0(z) - h_0(x)$, in cui la seconda differenza tende a $-\log_2 \frac{\Delta z}{\Delta x}$ che, se z ed x hanno la medesima dinamica, risulta pari a zero.

tale proposito, valutiamo il valore di entropia differenziale per un caso particolarmente rilevante.

1.3.2 Entropia differenziale di sorgente gaussiana

Applicando la (1.9) al caso $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-\frac{x^2}{2\sigma_x^2}}$, dopo aver osservato (vedi nota 1) che

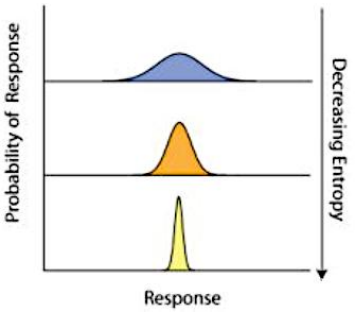
$$-\log_2 p\,(x) = -\frac{\ln p\,(x)}{\ln 2} = \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \frac{x^2}{2\sigma_x^2}\right)$$

possiamo scrivere

$$\begin{aligned} h_G\,(X) &= -\int p\,(x)\log_2 \frac{1}{\sqrt{2\pi}\sigma_x}\mathrm{e}^{-\frac{x^2}{2\sigma_x^2}}dx = \int p\,(x)\,\frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \frac{x^2}{2\sigma_x^2}\right)dx = \\ &= \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2}\int_{-\infty}^{\infty} p\,(x)\,dx + \frac{1}{2\sigma_x^2}\int_{-\infty}^{\infty} p\,(x)\,x^2 dx\right) = \\ &= \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \frac{1}{2}\right) = \frac{1}{\ln 2}\ln\sqrt{2\pi\mathrm{e}\sigma_x^2} = \log_2\sqrt{2\pi\mathrm{e}\sigma_x^2} = \\ &= \frac{1}{2}\log_2\left(2\pi\mathrm{e}\sigma_x^2\right) \end{aligned} \tag{1.10}$$

essendo $\frac{1}{2} = \ln e^{1/2}$, ed avendo di nuovo applicato la nota 1.

Dispersione come aumento dell'incertezza L'espressione (1.10) ci permette di osservare come la misura di entropia differenziale sia proporzionale (con legge logaritmica) alla varianza, che come sappiamo è una misura della dispersione della v.a. attorno al suo valor medio. A ciò infatti corrisponde un aumento dell'incertezza a riguardo dei possibili valori della v.a. La figura a lato esprime questo concetto con riferimento[3] ad una v.a. che rappresenta la risposta di un sistema biochimico ad uno stimolo,



1.3.2.1 Massima informazione per processo gaussiano

Al § **??** si mostra che il processo gaussiano è quello che consegue il massimo valore di entropia differenziale per σ_x^2 assegnata, ovvero è valida la diseguaglianza

$$h_G\,(X) = \frac{1}{2}\log_2\left(2\pi\mathrm{e}\sigma_x^2\right) > h\,(X) \quad \text{data } \sigma_x^2 \tag{1.11}$$

Principio di massima entropia In presenza di informazioni incomplete a riguardo di un sistema stocastico, come ad es. la conoscenza della sola varianza di una v.a., assumere l'ipotesi di gaussianità a riguardo della d.d.p. che lo governa equivale[4] ad adottare *le ipotesi meno restrittive* (ovvero più informative) a riguardo..

[3] Vedi Rhee, A. et al, *The application of information theory to biochemical signaling systems.* Physical biology (2012), https://doi.org/10.1088/1478-3975/9/4/045011

[4] Approfondimenti presso https://en.wikipedia.org/wiki/Principle_of_maximum_entropy

1.4 Misure di informazione per una coppia di v.a.

Descrivono da un punto di vista informativo i messaggi prodotti da una coppia di sorgenti, oppure osservate in ingresso ed in uscita da un canale trasmissivo, e che trattiamo come una v.a. bidimensionale. Per il momento, le definizioni vengono espresse nei termini di v.a. discrete.

1.4.1 Entropia congiunta

Si riferisce a due v.a. X e Y le cui realizzazioni sono descritte dalle d.d.p. marginali $p(x)$ e $p(y)$ e dalla d.d.p. congiunta $p(x,y)$, ed è definita come

$$H(X,Y) = H(Y,X) = -\sum_x \sum_y p(x,y) \log_2 p(x,y)$$

L'entropia congiunta risulta sempre non negativa, e delimitata tra

$$0 \leq \max\{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y)$$

in cui vale $H(X,Y) = \max\{H(X), H(Y)\}$ quando tra le due grandezze sussiste un legame deterministico, mentre si ha $H(X,Y) = H(X) + H(Y)$ qualora le v.a. siano statisticamente indipendenti, ovvero $p(x,y) = p(x)\,p(y)$. Nel caso di v.a. continua sussiste l'equivalente definizione per l'entropia differenziale congiunta

$$h(X,Y) = -\int_x \int_y p(x,y) \log_2 p(x,y)\, dx dy \tag{1.12}$$

1.4.2 Entropia condizionata

Ci si riferisce ancora a due v.a. X e Y ma si intende esprimere l'incertezza *residua* a riguardo di una di esse *qualora l'altra sia nota*. L'incertezza residua di (ad es.) Y per X nota è misurata, per una specifica realizzazione di (x,y), da $I(y/x) = -\log_2 p(y/x)$, mentre la relativa media di insieme è calcolata come valore atteso $E_{X,Y}\{I(y/x)\}$ rispetto ad entrambe le v.a., giungendo alla definizione di *entropia condizionata*

$$H(Y/X) = -\sum_x \sum_y p(x,y) \log_2 p(y/x) \tag{1.13}$$

Per la (1.13) risulta

$$0 \leq H(Y/X) \leq H(Y)$$

in cui la prima relazione è una uguaglianza se (e solo se) $p(y/x)$ è una funzione deterministica e non una d.d.p., mentre $H(Y/X) = H(Y)$ se (e solo se) Xed Y sono statisticamente indipendenti ovvero $p(y,x) = p(x)\,p(y)$, e quindi $p(y/x) = p(y)$.

Dalla (1.13) è facile[5] derivare la relazione

$$H(Y/X) = H(X,Y) - H(X) \tag{1.14}$$

che può essere *interpretata* considerando che la conoscenza di X apporta una informazione media $H(X)$, dunque per risolvere l'incertezza relativa ad Y sia necessario specificare (in media) solamente $H(X,Y) - H(X)$ ulteriori bit; in altre parole,

[5] Considerando infatti che $p(y/x) = \frac{p(x,y)}{p(x)}$, la (1.13) si riscrive come

$H(Y/X) = -\sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)} = -\sum_x \sum_y p(x,y) \log_2 p(x,y) + \sum_x \sum_y p(x,y) \log_2 p(x)$

in cui il primo termine è pari a $H(X,Y)$, ed il secondo (previa saturazione $p(x) = \sum_y p(x,y)$) ad $H(X)$.

l'incertezza residua $H(Y/X)$ è pari a quella congiunta $H(X,Y)$ meno quella ottenuta dalla conoscenza di X.

Dato inoltre che si possono sviluppare passaggi analoghi a quelli di nota 5 per quanto riguarda $H(X/Y)$, si ottiene anche che

$$H(X/Y) = H(X,Y) - H(Y) \tag{1.15}$$

e dunque sussiste *l'equivalente* del teorema di Bayes (§ **??**), ovvero

$$H(Y/X) = H(X/Y) + H(Y) - H(X) \tag{1.16}$$

Infine, nel caso di v.a. continue la definizione di entropia differenziale condizionale è

$$h(Y/X) = -\int_x \int_y p(x,y) \log_2 p(y/x)\, dxdy = -\int_x p(x) \int_y p(y/x) \log_2 p(y/x)\, dydx$$

i cui valori possono però risultare anche negativi o indeterminati (pag. 6).

1.4.3 Informazione mutua media

Anche questa grandezza tiene conto di due v.a. X e Y[6] descritte dalle d.d.p. marginali $p(x)$ e $p(y)$, nonché dalla d.d.p. congiunta $p(x,y)$; la sua definizione è

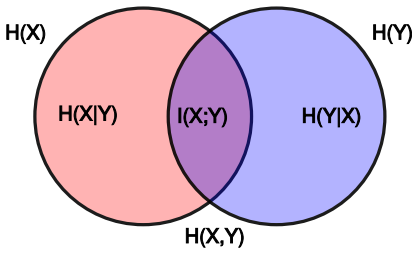
$$I(X;Y) = I(Y;X) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)} \tag{1.17}$$

ed ha un valore positivo o nullo, quest'ultimo se (e solo se) la v.a. sono indipendenti, cioè $p(x,y) = p(x)\,p(y)$.

A differenza delle due espressioni (1.14) e (1.15) per l'entropia condizionata, l'informazione mutua media $I(X;Y)$ *è simmetrica* rispetto alle due v.a. Il suo valore misura l'informazione che X e Y *condividono*, ovvero quanta informazione la conoscenza di una *apporti* nei confronti dell'altra. Per essa sussistono le eguaglianze[7]

$$\begin{aligned} I(X;Y) &= H(X) - H(X/Y) = H(Y) - H(Y/X) \\ &= H(X) + H(Y) - H(X,Y) = \\ &= H(X,Y) - H(X/Y) - H(Y/X) \end{aligned}$$

che possono essere meglio apprezzate nei termini di unione, differenza ed intersezione di insiemi, come raffigurato nel diagramma mostrato a lato. In particolare, in base alle prime due eguaglianze possiamo dire che $I(X;Y)$ è pari all'entropia di una delle due v.a., meno il numero di bit a simbolo necessari a descriverla qualora l'altra v.a. sia nota, ovvero meno l'*incertezza residua* qualora una delle due sia nota.



Anche questo concetto si applica al caso di v.a. continue, ottenendo l'espressione

[6] Vedi anche la trattazione al § 1.5.1 e seguenti nel caso in cui X ed Y siano le grandezze in ingresso ed in uscita da un canale di comunicazione.

[7] Vedi ad es. https://en.wikipedia.org/wiki/Mutual_information, ma anche la nota 10 a pag. 11

dell'informazione mutua media *differenziale*

$$I(X;Y) = \int_x \int_y p(x,y) \log_2 \frac{p(x,y)}{p(x)\, p(y)} dx dy$$

che *non dipende* dalla dinamica[8] delle v.a. X e Y come invece accadeva per l'entropia differenziale di una (1.9) o due (1.12) v.a.

1.5 Capacità di canale

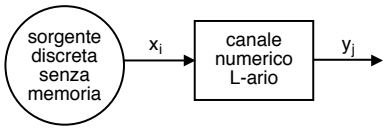
Lo scopo di ciò che segue è stabilire *i limiti* entro cui è possibile operare, ovvero quale sia *il massimo* teorico della quantità di informazione R trasmissibile su un determinato canale *rumoroso*, a cui cioè è associata una probabilità di errore P_e di osservare in uscita qualcosa di diverso da ciò che è entrato. Tale massimo è noto come *capacità* C del canale, espressa in bit/simbolo, e in canale di telecomunicazione dipende da grandezze di natura continua come *potenza*, *banda*, e *livello di rumore* in ricezione.

Sviluppi teorici che non affrontiamo in questa sede mostrano che finché l'entropia della sorgente H_s in ingresso al canale si mantiene inferiore al valore della *capacità di canale* C (§§ 1.6 e **??**), l'informazione può essere trasportata (teoricamente) *senza errori!* Mentre se al contrario $R > C$, non è possibile trovare nessun procedimento in grado di ridurre gli errori - che anzi, divengono praticamente *certi*.

1.5.1 Informazione mutua media per canale numerico $L-$ario

Approfondiamo questa nozione introdotta al § 1.4.3 mostrando come *l'informazione condivisa* tra ingresso ed uscita di un canale consenta di determinare anche la quantità di informazione che viene *persa* a causa degli errori che si sono verificati.

Consideriamo una sorgente discreta che emette simboli x appartenenti ad un alfabeto finito di cardinalità L, ossia $x \in \{x_i\}$ con $i = 1, 2, \cdots, L$, ed indichiamo con $y \in \{y_j\}$ (sempre per $j = 1, 2, \cdots, L$) il corrispondente simbolo ricevuto mediante un canale discreto, in generale diverso da x, a causa di errori introdotti dal canale. Conoscendo le densità di probabilità $p(x_i)$, $p(y_j)$, e le probabilità congiunte $p(x_i, y_j)$, possiamo definire la quantità di informazione *in comune* tra x_i e y_j, denominata *informazione mutua*, come[9]



$$I(x_i, y_j) = \log_2 \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)} = \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \log_2 \frac{p(y_j/x_i)}{p(y_j)} \quad \text{bit} \tag{1.18}$$

da cui deriva che

1. se ingresso ed uscita del canale sono *statisticamente indipendenti* si ha $p(x_i, y_j) = p(x_i)\, p(y_j)$, e di conseguenza l'informazione mutua *è nulla* per qualunque coppia x_i, y_j;

[8] Ciò deriva dall'essere le d.d.p. presenti sia a numeratore che a denominatore dell'argomento di $\log_2$.

[9] Per ottenere le diverse forme della (1.18) si ricordi che $p(x_i, y_j) = p(x_i/y_j)\, p(y_j) = p(y_j/x_i)\, p(x_i)$

2. se $p\left(y_j/x_i\right) > p\left(y_j\right)$ significa che l'essere a conoscenza della trasmissione di x_i rende la ricezione di y_j *più probabile* di quanto non lo fosse a priori, e corrisponde ad una informazione mutua *positiva*;
3. la definizione di informazione mutua è *simmetrica*, ovvero $I\left(x_i, y_j\right) = I\left(y_j, x_i\right)$;
4. rifrasando la 2. in virtù della 3., se $p\left(x_i/y_j\right) > p\left(x_j\right)$ allora ricevere y_j rende la trasmissione di x_i *più probabile* di quanto non lo fosse a priori, manifestando lo stesso valore di informazione mutua *positiva* del punto 2.

Per giungere ad una grandezza $I\left(X, Y\right)$ che tenga conto del comportamento *medio* del canale, ovvero per coppie ingresso-uscita qualsiasi, occorre pesare i valori di $I\left(x_i, y_j\right)$ con le relative probabilità congiunte, ossia calcolarne il valore atteso rispetto a tutte le possibili coppie $\left(x_i, y_j\right)$:

$$I\left(X;Y\right) = E_{X,Y}\left\{I\left(x_i, y_j\right)\right\} = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p\left(x_i/y_j\right)}{p\left(x_i\right)} \tag{1.19}$$

$$= \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p\left(y_j/x_i\right)}{p\left(y_j\right)} \tag{1.20}$$

ri-ottenendo così l'*informazione mutua media* (§ 1.4.3), misurata in bit/simbolo, e che rappresenta (in media) quanta informazione ogni simbolo ricevuto trasporta a riguardo di quello trasmesso. In virtù della simmetria di questa definizione, ci accorgiamo che il valore di $I\left(X, Y\right)$ può essere espresso[10] nelle due forme alternative

$$I\left(X;Y\right) = H\left(X\right) - H\left(X/Y\right) \tag{1.21}$$

$$= H\left(Y\right) - H\left(Y/X\right) \tag{1.22}$$

in cui l'entropia *condizionale* (§ 1.4.2)

$$H\left(X/Y\right) = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(x_i/y_j\right)} \tag{1.23}$$

prende il nome di *equivocazione* e rappresenta la quantità media di informazione *persa*, rispetto all'entropia di sorgente $H\left(X\right)$, a causa della rumorosità del canale. Nel caso in cui il canale non introduca errori, e quindi $p\left(x_i/y_j\right)$ sia pari a 1 se $j = i$ e zero altrimenti, è facile vedere[11] che $H\left(X/Y\right)$ è pari a zero, e $I\left(X;Y\right) = H\left(X\right)$, ossia tutta l'informazione della sorgente si trasferisce a destinazione. D'altra parte

$$H\left(Y/X\right) = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(y_j/x_i\right)} \tag{1.24}$$

prende il nome di *noise entropy* dato che considera il processo di rumore come se fosse

[10] Infatti

$$\sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p\left(x_i/y_j\right)}{p\left(x_i\right)} = \sum_i \sum_j p\left(x_i, y_j\right) \left[\log_2 \frac{1}{p\left(x_i\right)} - \log_2 \frac{1}{p\left(x_i/y_j\right)}\right] =$$

$$= \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(x_i\right)} - \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(x_i/y_j\right)}$$

L'ultimo termine è indicato come entropia condizionale $H\left(X/Y\right)$ (eq. (1.23)), mentre il penultimo è pari all'entropia di sorgente $H\left(X\right)$ dato che *saturando* la prob. congiunta $p\left(x_i, y_j\right)$ rispetto ad j, ovvero $\sum_j p\left(x_i, y_j\right) = p\left(x_i\right)$, si perviene alla (1.21) in base al risultato $\sum_i \log_2 \frac{1}{p\left(x_i\right)} \sum_j p\left(x_i, y_j\right) = \sum_i p\left(x_i\right) \log_2 \frac{1}{p\left(x_i\right)}$. Per la (1.22) il passaggio è del tutto simile.

[11] Infatti in tal caso la (1.23) diviene $\sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(x_i/y_j\right)} = \sum_i p\left(x_i, y_i\right) \log_2 1 = 0$

un segnale informativo: infatti, sebbene si possa essere tentati di dire che l'informazione media ricevuta è misurata dalla entropia $H(Y)$ della sequenza di osservazione, una parte di essa $H(Y/X)$ *è falsa*, perché in realtà è introdotta dagli errori.

1.6 Capacità di canale discreto

Le relazioni fin qui discusse permettono di valutare la perdita di informazione causata dai disturbi, ma dipendono sia dalle probabilità *in avanti* $p\left(y_j/x_i\right)$ che descrivono il comportamento del canale, sia da quelle *a priori* $p(x_i)$, che attengono alle caratteristiche della sorgente. Vogliamo invece definire una grandezza che esprima esclusivamente l'attitudine (o *capacità*) del canale a trasportare informazione, indipendentemente dalle caratteristiche della sorgente. Questo risultato può essere ottenuto variando le probabilità a priori in tutti i modi possibili, fino a trovare il valore

$$C_s = \max_{p(x)} I(X;Y) \quad \text{bit/simbolo} \tag{1.25}$$

che definisce la *capacità di canale per simbolo* come il massimo valore dell'informazione mutua media, ottenuto in corrispondenza della migliore sorgente possibile. Il pedice $_s$ sta per *simbolo*, e serve a distinguere il valore ora definito da quello che esprime la massima *intensità* di trasferimento dell'informazione espressa in bit/secondo, ottenibile una volta nota la frequenza f_s con cui sono trasmessi i simboli, fornendo per la capacità di canale il nuovo valore[12]

$$C = f_s \cdot C_s \quad \text{bit/secondo} \tag{1.26}$$

L'importanza di questa quantità risiede nel *teorema fondamentale per canali rumorosi*[13] già anticipato più volte, che asserisce che per ogni canale discreto senza memoria di capacità C

- esiste una tecnica di codifica che consente la trasmissione di informazione a velocità R e con probabilità di errore per simbolo p_e *piccola a piacere*, purché risulti $R < C$;
- se è accettabile una probabilità di errore p_e, si può raggiungere (con la miglior codifica possibile) una velocità $R(p_e) = \frac{C}{1-H_b(p_e)} > C$ in cui $H_b(p_e)$ è l'entropia di una sorgente binaria (1.6);
- per ogni valore di p_e non è possibile trasmettere informazione a velocità maggiore di $R(p_e)$.

Il teorema non suggerisce come individuare la tecnica di codifica, né fa distinzioni tra codifica di sorgente e di canale, ma indica le prestazioni limite ottenibili mediante la migliore tecnica possibile, in grado di ridurre a piacere la p_e purché $R < C$, mettendoci al tempo stesso in guardia a non tentare operazioni impossibili. Da questo punto di vista, le prestazioni conseguibili adottando le tecniche di codifica note possono essere valutate

[12] Notiamo l'invarianza di (1.26) rispetto al numero di livelli con cui è effettuata la trasmissione: se M bit sono raggruppati per generare simboli ad $L = 2^M$ livelli, come noto f_s si riduce di M volte, mentre C_s aumenta della stessa quantità, dato che ogni simbolo trasporta ora M bit anziché uno.

[13] http://it.wikipedia.org/wiki/Secondo_teorema_di_Shannon

confrontandole con quelle *ideali* predette dal teorema. Inoltre, dato che la capacità di canale è definita come massimo valore di $I(X,Y)$ per la migliore $p(x)$, qualora la statistica dei messaggi prodotti dal codificatore di sorgente differisca da quella ottima per il canale, l'effettiva informazione mutua media risulterà ridotta rispetto al valore della capacità, così come la massima velocità R.

Illustriamo l'applicazione di questi risultati al campo dei segnali biologici, attingendo dal riferimento citato alla nota 3.

1.7 Application of information theory to biochemical signaling systems

Ci troviamo nella circostanza in cui X ed Y rappresentano rispettivamente ad es. la concentrazione di un ligando e quella di una proteina, di una espressione genica, o di un fattore di trascrizione, oppure ancora Y può assumere valori discreti come quando rappresenta il fato di una cellula. A sua volta la prob. condizionata $p(y/x)$ rappresenta i fattori di variabilità intrinseca (legati alle singole reazioni chimiche coinvolte) che estrinseca (legati a cofattori ambientali).

Conseguenze della simmetria dell'informazione mutua An important consequence of the symmetry

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)$$

is that although we usually wish to quantify the reduction in uncertainty about X provided by the response Y (represented the former equality), it is usually *far easier* to experimentally measure the distribution $p(y)$ of responses Y (represented by the latter equality) from which evaluate $H(Y)$, as well as to measure $p(y/x)$ when knows stimuli are given at the input of the channel, from which to evaluate $H(Y/X)$.

Vantaggio rispetto alla correlazione Mentre nel caso della correlazione tra v.a., la loro indipendenza statistica determina una correlazione nulla ma non il viceversa (a meno che per v.a. gaussiane), nel caso dell'informazione mutua un suo valore pari a zero è condizione necessaria e sufficiente alla loro indipendenza statistica. In altri termini, mentre la correlazione cattura solamente relazioni *lineari* tra le variabili, la misura dell'informazione *trascende* dalla natura del legame.

Utilità del canale in funzione di quanti possibili valori in ingresso ed uscita Dato che sia $H(X/Y)$ che $H(Y/X)$ possono annullarsi in caso di una dipendenza *deterministica* tra X ed Y, dalle eq. (1.21) e (1.22)

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)$$

ne discende che $I(X;Y) \leq \min\{H(X), H(Y)\}$.

This upper bound also implies that the range of values that the input and output can take can limit the effectiveness of the communication channel. For instance, if we have a signal X that can take on 1.000.000 values (an entropy as high as $\log_2 10^6 \simeq 20$ bits) but the output Y that can only take on one of two values (entropy at most 1 bit)

then the mutual information between X and Y is necessarily 1 bit or less. As a result, a communication channel relying on a rich signal but poor output, or vice versa, can be limited in its ability to transmit information.

Misura della capacità For many biological signaling channels, $p\,(y/x)$ can be readily experimentally measured, whereas $p\,(x)$ cannot be easily estimated, particularly if X corresponds to commonly very low ligand concentrations and infrequent signaling events. However, channel capacity can be easily inferred (through optimization algorithms) by determining which $p\,(x))$ yields the maximum amount of information. Nonetheless, for biological channels the capacity may yield insights as to the magnitude of the actual amount of information transmitted, because under the efficient coding hypothesis, biological systems whose primary function is communication can be expected to have evolved to be optimally matched to the information sources that feed them.

Data processing inequality Essentially states that at every step of information processing, information cannot be gained, only lost. More precisely, for a Markov chain $X \rightarrow Y \rightarrow Z$, the data processing inequality states that $I(X;Z) \leq I(X;Y)$. That is, Z contains no more information about X as Y does.

The relevance of the data processing inequality is twofold. First, it places bounds on the performance of a biological system that contains multiple communication channels in series. For instance, consider $X \rightarrow Y$ to represent cytokine signaling to a transcription factor and $Y \rightarrow Z$ to represent transcription factor signaling to the concentration of an expressed protein. Assuming no other sources of information, then the amount of information that the expressed protein (Z) provides about the cytokine signal (X) cannot be more than the information that the transcription factor (Y) provides about the cytokine (X). If the information flow between X and Y is particularly limiting, this can place strict bounds on the fidelity of the response Z.

Second, the data processing inequality has implications for experimental measurements. For instance, consider the chain X (*signal*) $\rightarrow$ Y (*actual response*) $\rightarrow$ Y' (*measured response*). Although an experimentalist might wish to quantify the mutual information between the signal and actual response, $I(X;Y)$, she is confined to measuring $I(X;Y')$. For $I(X;Y')$ to be close in value to $I(X;Y)$ the noise between Y and Y' resulting from experimental error must be minimal. Thus, it is critical to pay close attention to the degree of experimental noise when attempting to measure mutual information.

Bias To understand the origin of the bias, recall that entropy depends on the range of values that a random variable can realize. A finite data sample, by its nature, will not reflect the full range of the underlying distribution and will give the perception that the distribution is thinner than it really is. Consequently, the entropy computed on a finite data sample will be smaller than the true entropy, although this negative bias will diminish as the sample size increases. Similarly, the conditional entropy $H(Y/X)$ is also negatively biased but more strongly than $H(Y)$ since by definition the sample used to

estimate $H(Y)$ is larger than the sample used to estimate $H(Y/X)$ (the sample for Y is the aggregate of all Y/X samples). Thus, from the equation $I(X;Y) = H(Y) - H(Y/X)$ it is evident that estimated mutual information will be positively biased, although again this bias will diminish as the sample size increases.