Appunti di Teoria dell'Informazione

Giovanni Bindi
Università degli Studi di Firenze

Indice

È stato un ingegnere e matematico statunitense, **Claude Elwood Shannon** (1916 – 2001), che per primo ha fatto diventare l'informazione qualcosa di ben definito e misurabile fornendo delle basi fondamentali per i sistemi di comunicazioni.
Shannon ha lavorato nei laboratori Bell dal '41 al '72 e nel 1948 ha pubblicato "*A Mathematical Theory of Communication*" su The Bell System Technical Journal, una relazione tecnica che ora è alla base della Teoria dell'Informazione.
Oltre a definire e dare un'unità di misura all'informazione, la teoria di Shannon permette di rispondere anche a due domande fondamentali:

- Quale è la massima compressione dei dati informativi senza perdita che si può ottenere.
- Quale è il massimo rate di trasmissione che si può avere per comunicazioni affidabili.

Gli studi di Shannon intendevano infatti migliorare l'efficienza della trasmissione dell'informazione:

"Il problema fondamentale della comunicazione consiste nel riprodurre in un punto, esattamente o approssimativamente, un messaggio selezionato in un altro punto."

L'importanza dei risultati degli studi di Shannon sta anche nel fatto che permettono di ridurre a forme analitiche abbastanza semplici, problemi in realtà molto complessi e generali.
Dato un messaggio prodotto da una sorgente informativa, l'obiettivo della teoria dell'informazione è capire come si deve rappresentare tale messaggio per ottenere una trasmissione efficiente dell'informazione in esso contenuta su di un canale di comunicazione reale, ovvero soggetto a inevitabili limitazioni fisiche.
Questo obiettivo viene perseguito attraverso quattro passi fondamentali:

1. Definizione e misura dell'informazione di una sorgente.
2. Capire, data una sorgente informativa, *come* e *quanto* è possibile ridurre il suo rate di trasmissione (*Codifica di Sorgente*).
3. Definire cosa sia la capacità di comunicazione di un canale e sotto quali condizioni i dati provenienti da una sorgente informativa possono essere trasmessi in modo affidabile (*Rate Distortion Theory*).
4. Come si può sfruttare al massimo la capacità di trasmissione di un canale rimuovendo (o rendendo trascurabili) gli effetti del canale di comunicazione (*Codifica di Canale*).

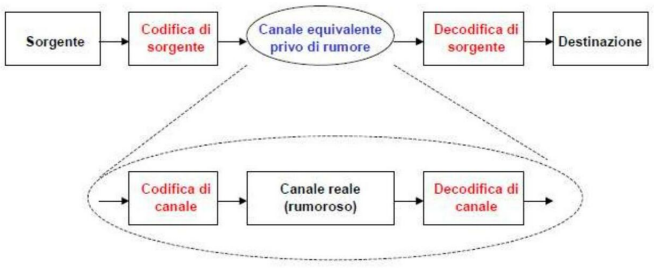


Figura 1: Percorso del messaggio, dalla sorgente al destinatario.

Come raffigurato in Figura 1 la codifica (sia di sorgente che di canale) si deve adattare alla sorgente e al canale in modo da avere la massima efficienza possibile nel trasferimento dell'informazione:

- La *codifica di sorgente* adatta la sorgente alla trasmissione su di un opportuno canale equivalente privo di rumore.
- La *codifica di canale* permette di trasmettere l'informazione emessa dalla sorgente (opportunamente trattata mediante la codifica di sorgente) in maniera affidabile su un canale reale caratterizzato da limitazioni fisiche.

* * *

Questi appunti sono stati elaborati principalmente sulla base del materiale pubblicato dal titolare del corso, a cui vanno tutti i crediti. I riferimenti principali oltre a questo sono stati:

1. *Elements of Information Theory* - *Thomas M. Cover & Joy A. Thomas.*
2. *Lecture notes on Information Theory and Coding* - *Mauro Barni & Benedetta Tondi.*
3. *Information Theory, Pattern Recognition, and Neural Networks* - *David MacKay.*

Chiunque volesse contribuire a questo materiale, segnalando degli errori o estendendo il contenuto, mi può contattare a gio.bindi@pm.me.

Novembre 2019.

1 Definizioni fondamentali e primi risultati

1.1 Sorgenti

In generale si dice sorgente S un oggetto che emette[1] simboli x appartenenti ad un alfabeto X. La sorgente è nota quando conosciamo la probabilità con cui viene emesso ciascun simbolo. Possiamo distinguere tra:

- Sorgenti **discrete**, in cui l'alfabeto è un insieme finito.
- Sorgenti **continue**, in cui l'alfabeto è un insieme di simboli infinito numerabile.

Sappiamo, comunque, che qualsiasi sorgente continua può essere trasformata in una sorgente numerica (in particolare digitale) grazie alle operazioni di campionamento e quantizzazione. Le sorgenti con cui abbiamo a che fare nei moderni sistemi informativi sono infatti sostanzialmente sempre sorgenti discrete. Per questo motivo la nostra attenzione sarà principalmente rivolta alle sorgenti discrete anche se poi i risultati verranno generalizzati al caso continuo perché ci sono contesti (in particolare il canale di comunicazione) in cui abbiamo a che fare con segnali/sorgenti continui.
I simboli emessi dalla sorgente sono delle **variabili aleatorie** (altrimenti *non ci sarebbe informazione da trasmettere*) e quindi ci si riferisce sempre a variabili e processi aleatori. Quando, infatti, si considera l'emissione successiva di simboli nel tempo si ha un **processo stocastico**, ovvero una sequenza di variabili aleatorie indicizzate nel tempo.

Un'altra distinzione che possiamo fare nelle sorgenti dipende dal legame tra simboli successivi (emessi ad istanti successivi dalla sorgente). Se questi sono dipendenti gli uni dagli altri la sorgente si dice **con memoria**, altrimenti si dice **senza memoria** e le variabili aleatorie sono indipendenti e identicamente distribuite.
Le sorgenti senza memoria sono più semplici ma più rare. Un lancio di monete è una sorgente senza memoria: in questo caso basta conoscere la probabilità di ogni singolo simbolo perché la probabilità congiunta è il prodotto delle singole probabilità. Le sorgenti con memoria invece sono più complesse ma sono anche quelle che di solito si trovano in pratica, e sono caratterizzate dalle probabilità congiunte. Ad esempio un testo scritto: ci sono ovviamente dei legami tra le lettere che escono, ad esempio una q è seguita da una u, perché devono rispettare una sintassi.

1.2 Informazione ed Entropia

Prima di tutto il problema della teoria dell'informazione è definire, rappresentare matematicamente e quantificare l'informazione che viene prodotta da una sorgente. L'informazione può essere di tipo:

- **Semantico**, ovvero riguardare il significato del messaggio.
- **Sintattico**, ovvero riguardare i simboli che si usano e come questi sono relazionati tra loro - come è costruito il messaggio.

[1] In questi appunti spesso si userà lo stesso simbolo sia per la sorgente che per l'alfabeto, anche se ci sono testi in cui si fa differenza, come in Information Theory, Inference, and Learning Algorithms.

La teoria dell'informazione si occupa *solo dell'aspetto* ***sintattico*** (o simbolico): a livello di sistema di comunicazione non ci interessa la semantica, ovvero cosa significa un dato messaggio (dal momento che incontriamo il problema della soggettività dell'informazione) ma la *quantità* di informazione che questo porta.
Shannon sviluppò l'idea di definire l'informazione legata ad un evento x solo in relazione alla probabilità $p(x)$ che quell'evento avvenga, in particolare ebbe l'intuizione di imporre che il contenuto informativo dell'evento fosse tanto maggiore quanto più bassa fosse la probabilità dell'evento associato. Le proprietà che la definizione di informazione $I(\cdot)$ deve soddisfare sono:

- Deve essere una funzione (continua) della probabilità: $I(x) = f(p(x))$ e deve essere $f(p(x)) \in (0, 1]$ (un evento che non può avvenire non è di interesse).
- Deve essere $I(1) = 0$, dal momento che un evento certo non porta informazione.
- Deve essere $p(x) \to 0^+ \implies I(x) \to \infty$, ovvero che un evento raro porti molta informazione, e quindi che la funzione f sia decrescente.

Sia S una sorgente (una variabile aleatoria discreta) che emette simboli su un alfabeto $X = \{x_1, x_2, \dots, x_M\}$ con una distribuzione $p(x) = Pr\{S = x\}, x \in X$.

Definizione 1. *Informazione*: L'informazione associata al simbolo x è definita[2] come

$$I(x) \coloneqq \log \frac{1}{p(x)} \tag{1.2.1}$$

Se si hanno più eventi *indipendenti*, essendo la probabilità congiunta degli eventi il prodotto delle probabilità, l'informazione complessiva è la somma delle singole informazioni:

$$I(x, y) = \log \frac{1}{p(x, y)} = \log \frac{1}{p(x)p(y)} = \log \frac{1}{p(x)} + \log \frac{1}{p(y)} = I(x) + I(y) \tag{1.2.2}$$

È importante sottolineare come l'informazione sia solo legata alla probabilità che un evento accada, non è in alcun modo legata alla natura dell'evento.
Se consideriamo una sorgente S discreta senza memoria (DMS): si ha che l'informazione media della sorgente è data dal valore atteso dell'informazione:

Definizione 2. *Entropia*: L'entropia associata alla sorgente DMS è definita come

$$H(S) \coloneqq \mathbb{E}_p[I(x)] = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log p(x) \tag{1.2.3}$$

e si misura in *bit/simbolo*. Vedremo più avanti come l'entropia dia una misura del "costo minimo" per rappresentare un'informazione, ovvero il *minimo numero medio di bit che servono per rappresentare le informazioni inviate da una sorgente.*
Quando si ha $p(x) = 0$ si adotta la convenzione $0 \times \log \frac{1}{p(0)} \coloneqq 0$ dal momento che

$$\lim_{x \to 0^+} x \log \frac{1}{x} = 0 \tag{1.2.4}$$

[2] Il logaritmo si intende implicitamente in base 2 anche se può essere ovviamente operato un cambio di base: $\log_b p = \log_b a \log_a p$. In questo caso cambia solamente l'unità con cui si misura l'informazione. Per $b = 2$ si ha il *bit*, per $b = 3$ il *trit*, per $b = e$ il *neper* e così via. Si veda la Figura 2b per l'andamento della funzione di informazione al variare della base b.

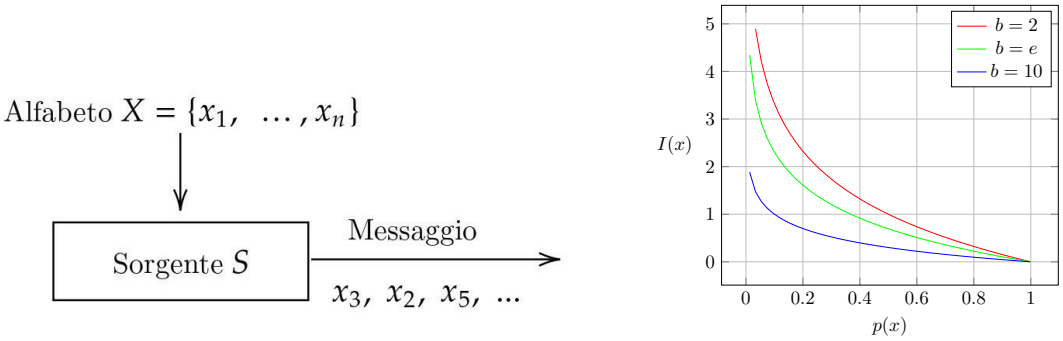


Figura 2: Rappresentazione schematica di una sorgente e grafico della funzione di informazione.

Esempio 1 : Sorgente Binaria

Consideriamo una sorgente S che emette due soli simboli: $X = \{x, y\}$ con probabilità $p(x) = q$ e $p(y) = 1 - q$. Si ha che

$$H(S) = H(q) = q \log \frac{1}{q} + (1 - q) \log \frac{1}{1 - q} \tag{1.2.5}$$

Ad esempio se $q = 0.7$ si avrebbe $H(q) \approx 0.88$ *bit/simbolo*. Quand'è che l'entropia di una sorgente binaria è massima? Si vede facilmente che il massimo si ottiene in:

$$\begin{aligned}
&\frac{d}{dq} H(q) = \frac{d}{dq} \Big\{ - q \log q - (1 - q) \log 1 - q \Big\} = \\
&= - \Big\{ \log q + q \frac{1}{q} \log e - \log(1 - q) - (1 - q) \frac{1}{1 - q} \log e \Big\} = \\
&= \log (1 - q) - \log q = 0
\end{aligned}$$

ovvero quando si ha

$$1 - q = q \implies q = \frac{1}{2}$$

Quindi si ha che l'entropia di una sorgente binaria è massima quando i due simboli sono equiprobabili e vale $H(q) = 1$.

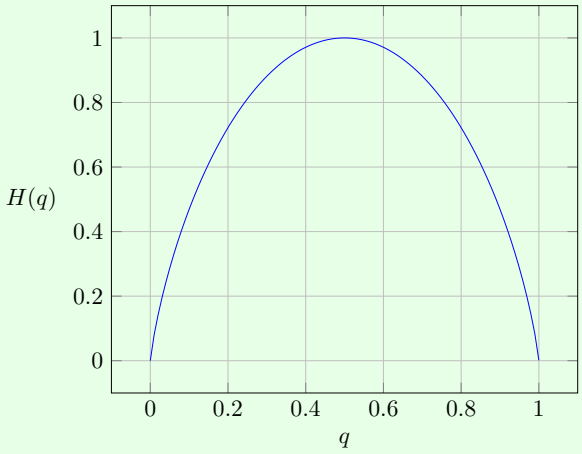


Figura 3: Entropia di una sorgente binaria, al variare di q.

Si hanno poi queste due proprietà, in cui la seconda è una generalizzazione di quanto appena visto per una sorgente binaria:

1. $H(S) \geq 0$ dal momento che $0 \leq p(x) \leq 1 \implies \log \frac{1}{p(x)} \geq 0$.
2. $H(S) \leq \log M$ con $H(S) = \log M \iff$ i simboli sono equiprobabili.

Vediamo la dimostrazione per questa seconda proprietà:

Dim: Vogliamo provare che $H(S) - \log M \leq 0$. Si ha che

$$\begin{aligned}
&H(S) - \log M = \\
&= \sum_{x \in X} p(x) \log \frac{1}{p(x)} - \log M \overset{\alpha}{=} \\
&= \sum_{x \in X} p(x) \log \frac{1}{p(x)} - \sum_{x \in X} p(x) \log M = \\
&= \sum_{x \in X} p(x) \Big(\log \frac{1}{Mp(x)} \Big) = \\
&= \log e \Big[\sum_{x \in X} p(x) \Big(\ln \frac{1}{Mp(x)} \Big) \Big] \overset{\beta}{\leq} \\
&\leq \log e \Big[\sum_{x \in X} p(x) \Big(\frac{1}{Mp(x)} - 1 \Big) \Big] = \\
&= \log e \Big[\sum_{x \in X} \Big(\frac{1}{M} - p(x) \Big) \Big] = \\
&= \log e \Big[1 - \sum_{x \in X} p(x) \Big] \overset{\alpha}{=} 0
\end{aligned}$$

Si ha poi che la disuguaglianza vale con l'uguale quando, $\forall x \in X$:

$$\ln \frac{1}{Mp(x)} = \frac{1}{Mp(x)} - 1 \iff \frac{1}{Mp(x)} = 1 \iff p(x) = \frac{1}{M}$$

ovvero quando tutti i simboli x sono equiprobabili. □

Dette p, q due distribuzioni sullo stesso alfabeto X si definisce una quantità importante:

Definizione 3. *Entropia Relativa*: L'entropia relativa, detta anche *divergenza di Kullback-Leibler* è una misura non simmetrica[3] della differenza tra due distribuzioni di probabilità. Misura infatti l'inefficienza (la perdita di informazione che si ha) nell'assumere che la distribuzione di probabilità

[3] Anche se è spesso pensata come una distanza, la divergenza KL non è una vera e propria metrica - per esempio, infatti, non è simmetrica: la KL da p a q non è in genere la stessa KL da q a p. Tuttavia, la sua forma infinitesimale, in particolare la sua matrice Hessiana, è un tensore metrico: è l'informazione metrica di Fisher. Oltre a non essere simmetrica non può essere una distanza dal momento che non vale la disuguaglianza triangolare.

sia q quando quella reale è p.

$$D(p\|q) \coloneqq \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \tag{1.2.6}$$

Anche per l'entropia relativa si può mostrare un'importante proprietà, quella della non-negatività:

- $D(p\|q) \geq 0$ con $D(p\|q) = 0 \iff p = q$

Dim: Vogliamo mostrare che $-D(p\|q) \leq 0$. Si ha:

$$-D(p\|q) = \sum_{x\in X} p(x)\log\frac{q(x)}{p(x)} = \log e \sum_{x\in X} p(x)\ln\frac{q(x)}{p(x)} \overset{\beta}{\leq}$$

$$\leq \log e \sum_{x\in X} p(x)\Big(\frac{q(x)}{p(x)} - 1\Big) = \log e\Big(\sum_{x\in X} q(x) - \sum_{x\in X} p(x)\Big) = 0$$

□

L'entropia relativa è importante in molti ambiti diversi della teoria dell'informazione (e non solo) ed ha una interpretazione legata alla costruzione dei codici: se abbiamo una sorgente S che emette simboli con distribuzione p per rappresentarla possiamo costruire un codice con lunghezza $H(p)$ bit/simbolo, se però usiamo un codice costruito per una distribuzione q c'è bisogno in media di una lunghezza $H(p) + D(p\|q)$ per rappresentarla.
Data una variabile aleatoria congiunta (X, Y) con distribuzione di probabilità congiunta $p(x, y)$ si definisce l'entropia congiunta delle variabili X e Y come

Definizione 4. *Entropia Congiunta*:
Sia $x \in X = \{x_1, \ldots, x_M\}$ e $y \in Y = \{y_1, \ldots, y_Q\}$ una coppia di variabili aleatorie con distribuzione congiunta $p(x, y)$. Si ha che l'entropia congiunta tra X e Y è definita come:

$$H(X, Y) \coloneqq \sum_{x\in X}\sum_{y\in Y} p(x, y)\log\frac{1}{p(x, y)} \tag{1.2.7}$$

Una quantità strettamente correlata all'entropia congiunta è la:

Definizione 5. *Entropia Condizionata*: L'entropia condizionata è definita come:

$$H(X|Y) = \sum_{x\in X}\sum_{y\in Y} p(x, y)\log\frac{1}{p(x|y)} \tag{1.2.8}$$

Per quanto riguarda il rapporto tra entropia congiunta e condizionata si può vedere che:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{1.2.9}$$

che prende il nome di *chain rule* per l'entropia.

Dim: Mostriamo che $H(X,Y) = H(X) + H(Y|X)$, vale l'analogo per l'altro caso.

$$H(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x,y)} \overset{\gamma}{=} \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)p(x)} =$$
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\Big[\log\frac{1}{p(x)} + \log\frac{1}{p(y|x)}\Big] =$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x)} + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)} =$$
$$\overset{\delta}{=} \sum_{x\in X} p(x)\log\frac{1}{p(x)} + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)} =$$
$$= H(X) + H(Y|X) \qquad \square$$

La chain rule per l'entropia può essere generalizzata a m variabili aleatorie $X_1, X_2, \ldots, X_m$:

$$H(X_1, X_2, \ldots, X_m) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_m|X_1, \ldots, X_{m-1}) \quad (1.2.10)$$

ricordando che la chain rule per la probabilità è data da:

$$p(x_1, x_2, \ldots, x_m) = p(x_1)\cdot p(x_2|x_1)\cdot p(x_3|x_1, x_2)\cdot\cdots\cdot p(x_m|x_1, \ldots, x_{m-1}) \quad (1.2.11)$$

Una proprietà importante dell'entropia condizionata è che, in generale, si ha

$$H(X|Y) \leq H(X) \quad (1.2.12)$$

ovvero che il condizionamento non può far aumentare l'entropia (*information can't hurt*).

Dim: Mostriamo che $H(X|Y) - H(X) \leq 0$

$$H(X|Y) - H(X) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x|y)} - \sum_{x\in X} p(x)\log\frac{1}{p(x)} \overset{\delta}{=}$$
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x|y)} - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x)} =$$
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x)}{p(x|y)} = \log e\sum_{x\in X}\sum_{y\in Y} p(x,y)\ln\frac{p(x)}{p(x|y)} \overset{\beta}{\leq}$$
$$\leq \log e\sum_{x\in X}\sum_{y\in Y} p(x,y)\Big[\frac{p(x)}{p(x|y)} - 1\Big] = \log e\sum_{x\in X}\sum_{y\in Y} p(x,y)\Big[\frac{p(x)p(y)}{p(x,y)} - 1\Big] =$$
$$= \log e\Big[\sum_{x\in X} p(x)\sum_{y\in Y} p(y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\Big] = 0 \qquad \square$$

Ovviamente vale anche $H(Y|X) \leq H(Y)$ per cui, per l'entropia congiunta, vale anche questa disuguaglianza:

$$H(X,Y) \leq H(X) + H(Y) \quad (1.2.13)$$

1.3 Informazione Mutua

Un'altra grandezza fondamentale è l'informazione mutua tra due sorgenti. Questa misura in un certo senso il grado di dipendenza di due variabili aleatorie, ovvero misura *quanta informazione di una variabile aleatoria è contenuta nell'altra*. Può essere pensata come la quantità di riduzione dell'incertezza su una variabile aleatoria quando si osserva l'altra. È definita come

Definizione 6. *Informazione mutua:* Date due variabili aleatorie X, Y si ha che l'informazione mutua è definita come:

$$I(X;Y) \coloneqq \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1.3.1}$$

Data questa definizione si vede subito che

$$I(X;Y) = D\Big(p(x,y)\|p(x)p(y)\Big) \tag{1.3.2}$$

ovvero che l'informazione mutua è *l'entropia relativa tra la distribuzione congiunta e il prodotto delle marginali*. Da questa considerazione si ha subito una prima proprietà per l'informazione mutua:

$$I(X;Y) \geq 0 \tag{1.3.3}$$

Si ha inoltre che, se le due variabili aleatorie sono indipendenti, $p(x,y) = p(x)p(y)$, si ha $I(X;Y) = 0$ ovvero non c'è distanza tra le due distribuzioni. Viceversa più le variabili aleatorie sono tra loro legate più l'informazione mutua cresce, difatti la differenza tra la distribuzione congiunta e il prodotto delle due aumenta.
Se X e Y sono indipendenti, allora la conoscenza di X non dà alcuna informazione riguardo a Y e viceversa, perciò la loro mutua informazione è zero. All'altro estremo, se X e Y sono identiche allora tutte le informazioni trasmesse da X sono condivise con Y: la conoscenza di X determina il valore di Y e viceversa.
L'informazione mutua è strettamente collegata all'entropia, infatti:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{1.3.4}$$

Dim: Mostriamo che $I(X;Y) = H(X) - H(X|Y)$, l'altro caso è equivalente.

$$\begin{aligned}
I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \sum_{x\in X}\sum_{y\in Y} p(x,y)\frac{p(x|y)p(y)}{p(x)p(y)} = \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x|y)}{p(x)} = \sum_{x\in X}\sum_{y\in Y} p(x,y)\Big[\log\frac{1}{p(x)} - \log\frac{1}{p(x|y)}\Big] = \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x)} - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x|y)} = \\
&= \sum_{x\in X} p(x)\log\frac{1}{p(x)} - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x|y)} = H(X) - H(X|Y) \qquad \square
\end{aligned}$$

In sostanza quindi l'informazione mutua è una differenza tra entropie: l'entropia di una variabile aleatoria meno l'incertezza di quella variabile aleatoria una volta che ho conosciuto l'altra. Si ha inoltre che

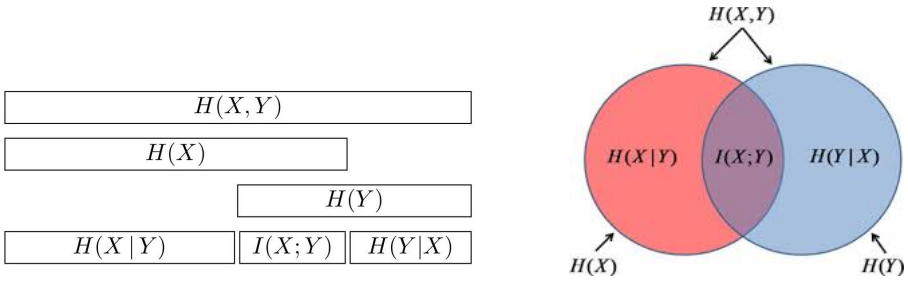


Figura 4: Due rappresentazioni equivalenti delle relazioni tra entropia e informazione mutua.

- L'informazione mutua è simmetrica: $I(X;Y) = I(Y;X)$.
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$ (dalla chain rule per l'entropia).
- Nel caso discreto vale $I(X;X) = H(X)$ (dal fatto che $H(X|X) = 0$) per cui si ha $I(X;X) \geq I(X;Y)$, ovvero che una variabile X contiene almeno tanta informazione riguardo a sé stessa di quanta ne può fornire una qualsiasi altra variabile Y.

1.4 Estensione di una sorgente

Siano $X_1, X_2, \ldots, X_n$ un numero n di variabili aleatorie identicamente distribuite (*i.d.*) (come ad esempio nella trasimissione di un messaggio in cui ogni simbolo emesso appartiene allo stesso alfabeto). L'entropia congiunta di queste n variabili aleatorie rappresenta una grandezza importante:

$$H(X_1, X_2, \ldots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1, x_2, \ldots, x_n) \log \frac{1}{p(x_1, x_2, \ldots, x_n)} \tag{1.4.1}$$

Definizione 7. *Estensione della sorgente:* La quantità $(X_1, X_2, \ldots, X_n)$ prende il nome di estensione della sorgente e si indica con X^n. Dal momento che queste n variabili aleatorie sono $i.d$ si ha che $H(X) = H(X_1) = H(X_2) = \cdots = H(X_n)$. Nel caso in cui non si ha memoria, oveero quando queste n variabili aleatorie sono anche indipendenti (sono quindi $i.i.d$) si ha che

$$H(X^n) \coloneqq H(X_1, X_2, \ldots, X_n) = nH(X) \tag{1.4.2}$$

Dim:

$$\begin{aligned}
H(X_1, X_2, \ldots, X_n) &= \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1, x_2, \ldots, x_n) \log \frac{1}{p(x_1, x_2, \ldots, x_n)} = \\
&= \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1)p(x_2)\ldots p(x_n) \log \frac{1}{p(x_1)p(x_2)\ldots p(x_n)} = \\
&= \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p(x_1)p(x_2)\ldots p(x_n) \Big[\log \frac{1}{p(x_1)} + \log \frac{1}{p(x_2)} \cdots + \log \frac{1}{p(x_n)} \Big]
\end{aligned}$$

Si ha quindi la somma di n termini. Analizziamo il generico elemento k di questa somma:

$$\sum_{x_1 \in X_1} \cdots \sum_{x_k \in X_k} \cdots \sum_{x_n \in X_n} p(x_1) \dots p(x_k) \dots p(x_n) \log \frac{1}{p(x_k)} =$$
$$\underbrace{\sum_{x_1 \in X_1} \cdots \sum_{x_n \in X_n} p(x_1) \dots p(x_n)}_{=1} \sum_{x_k \in X_k} p(x_k) \log \frac{1}{p(x_k)} = H(X_k)$$

Quindi, considerando tutti gli n termini $i.i.d$ vale:

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2) = \cdots + H(X_n) = nH(X) \qquad \square$$

1.5 Entropy Rates

Consideriamo ora sorgenti in cui l'emissione di un simbolo *non è indipendente* dai simboli emessi in precedenza dalla sorgente. Non si può più applicare quindi la definizione di entropia vista per una sorgente DMS, perché in questo caso l'informazione media portata da un simbolo della sorgente deve tenere in considerazione lo **stato** in cui si trova la sorgente (ovvero le relazioni del simbolo con quelli precedenti). Una sorgente con memoria è caratterizzata da uno stato che è rappresentato dai simboli emessi in precedenza. In questo caso vale quindi

$$H(X_1, X_2, \dots, X_n) < H(X_1) + H(X_2) + \dots H(X_n) = nH(X) \tag{1.5.1}$$

a causa della dipendenza dalle variabili.

Per definire l'entropia di una sorgente con memoria si deve tenere in considerazione la dipendenza tra simboli successivi. Se volessimo definire l'informazione media portata da un simbolo potremmo distinguere due casi:

- L'entropia congiunta di tutti i simboli emessi successivamente dalla sorgente e poi divisa per n, per avere l'informazione media per simbolo con $n \to \infty$ (asintoticamente):
$$\lim_{n \to \infty} \frac{H(X_1, X_2, \dots, X_n)}{n} \tag{1.5.2}$$
- L'informazione che viene portata da un simbolo, condizionata a tutti i simboli precedentemente emessi, quando $n \to \infty$, ovvero tenendo conto di infinite relazioni con i simboli precedenti
$$\lim_{n \to \infty} \frac{H(X_n | X_1, \dots, X_{n-1})}{n} \tag{1.5.3}$$

Questi due limiti prendono il nome di **entropy rates**.

È possibile dimostrare che nel caso di processi stocastici stazionari questi due limiti esistono e coincidono ma, in generale, per il limite di $n \to \infty$, calcolare le probabilità condizionate diventa impossibile. Solo in alcuni casi risulta fattibile e noi in particolare noi analizzeremo il caso delle sorgenti di Markov.

1.6 Sorgenti di Markov

Definizione 8. *Sorgente di Markov di ordine k:* Una sorgente di Markov di ordine k è una sorgente in cui l'emissione di un simbolo ad un certo istante di tempo dipende dai k simboli emessi in precedenza.

In generale si definisce processo di Markov (o anche *catena di Markov*), un processo stocastico in cui la probabilità di transizione che determina il passaggio a uno stato di sistema dipende solo dallo stato del sistema immediatamente precedente (proprietà di Markov) e non da come si è giunti a questo stato.

Nel caso particolare di una sorgente di Markov di ordine 1 l'insieme degli stati coincide con l'insieme dei simboli e la distribuzione di probabilità degli stati equivale alla distribuzione di probabilità dei simboli. In una sorgente di ordine k la probabilità di emettere il simbolo x_i è condizionata dallo stato precedente, formato dai $\{x_{i-1}, \dots, x_{i-k}\}$ simboli. Se supponiamo di conoscere lo stato allora abbiamo che l'informazione del simbolo x_i è data da

$$I(x_i|x_{i-1}, \dots, x_{i-k}) = \log \frac{1}{p(x_i|x_{i-1}, \dots, x_{i-k})} \tag{1.6.1}$$

da cui abbiamo che l'informazione media per simbolo della sorgente quando si trova in quello stato è

$$H(X|x_{i-1}, \dots, x_{i-k}) = \sum_{x_i \in X} p(x_i|x_{i-1}, \dots, x_{i-k}) \log \frac{1}{p(x_i|x_{i-1}, \dots, x_{i-k})} \tag{1.6.2}$$

Per cui per ottenere l'entropia della sorgente dobbiamo mediare su tutti i possibili stati. Chiamiamo σ_k l'insieme dei simboli dello stato $\sigma_k = \{x_{i-1}, \dots, x_{i-k}\}$: si ha che questo elemento appartiene all'estensione k-esima della sorgente $\sigma_k \in X^k$ per cui

$$\begin{aligned} H(X) &= \sum_{\sigma_k \in X^k} p(\sigma_k) \sum_{x_i \in X} p(x_i|\sigma_k) \log \frac{1}{(x_i|\sigma_k)} = \\ &= \sum_{\sigma_{k+1} \in X^{k+1}} p(\sigma_{k+1}) \log \frac{1}{p(x_i|\sigma_k)} \end{aligned} \tag{1.6.3}$$

dove $\sigma_{k+1} = \{x_i, x_{i-1}, \dots, x_{i-k}\}$. Se la sorgente fosse senza memoria ($k = 0$) si otterrebbe la definizione usuale di entropia ($\sigma_1 = \{x_i\}$), se fosse di ordine $k = 1$ si avrebbe

$$H(X) = \sum_{\{x_i, x_{i-1}\} \in X^2} p(x_i, x_{i-1}) \log \frac{1}{p(x_i|x_{i-1})} \tag{1.6.4}$$

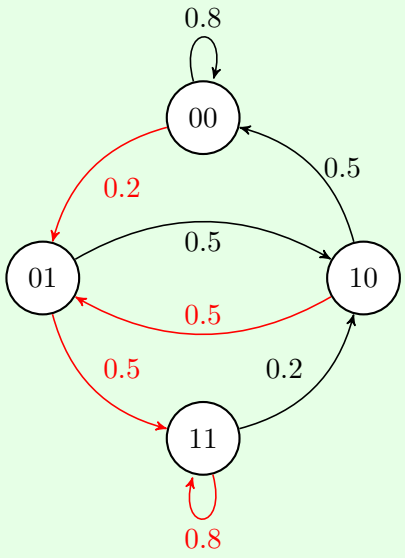
Esempio 2 : Sorgente di Markov di ordine 2

Consideriamo una sorgente di Markov di ordine $k = 2$ sull'alfabeto binario $X = \{0, 1\}$. Seguendo la notazione assumiamo che le probabilità condizionali dei simboli siano:

$$\begin{cases} p(0|00) = p(1|11) = 0.8 \\ p(1|00) = p(0|11) = 0.2 \\ p(0|01) = p(0|10) = p(1|01) = p(1|10) = 0.5 \end{cases}$$

Ovvero che, ad esempio, $p(x_i = 0|x_{i-1} = 0, x_{i-2} = 0) = 0.8$. Dal momento che abbiamo una

sorgente di ordine 2 su un alfabeto binario si hanno 4 stati: $\{00, 01, 10, 11\}$. Il diagramma a stati di questa sorgente è dato da:



in cui in rosso sono etichettate le transizioni con 1 e in nero quelle con 0. Il diagramma può anche essere equivalentemente scritto attraverso la matrice di transizione P

$$P = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \end{bmatrix}$$

da cui si può ricavare la distribuzione stazionaria π come $\pi P = \pi$, ottenendo

$$\pi = \begin{bmatrix} \frac{5}{14} & \frac{2}{14} & \frac{2}{14} & \frac{5}{14} \end{bmatrix}$$

ovvero che

$$\begin{cases} p(00) = p(11) = \frac{5}{14} \\ p(01) = p(10) = \frac{2}{14} \end{cases}$$

Infine possiamo scrivere la tabella delle probabilità della sorgente

x_{i-1}, x_{i-2}, x_i			$p(x_i \mid x_{i-1}, x_{i-2})$	$p(x_{i-1}, x_{i-2})$	$p(x_{i-1}, x_{i-2}, x_i)$
0	0	0	0.8	5/14	4/14
0	0	1	0.2	5/14	1/14
0	1	0	0.5	2/14	1/14
0	1	1	0.5	2/14	1/14
1	0	0	0.5	2/14	1/14
1	0	1	0.5	2/14	1/14
1	1	0	0.2	5/14	1/14
1	1	1	0.8	5/14	4/14

da cui calcolare l'entropia, seguendo 1.6.3, come

$$H(X) = \sum_{X^3} p(x_i, x_{i-1}, x_{i-2}) \log \frac{1}{p(x_i|x_{i-1}, x_{i-2})} =$$
$$= 2 \times \frac{4}{14} \log \frac{1}{0.8} + 2 \times \frac{1}{14} \log \frac{1}{0.2} + 4 \times \frac{1}{14} \log \frac{1}{0.5} \approx 0.80 \quad bit/simbolo$$

1.6.1 Sorgente estesa di una sorgente di Markov

Anche per le sorgenti con memoria si possono definire le sorgenti estese X^n, ovvero che emettono n simboli consecutivi, con la differenza che questi simboli hanno delle relazioni di dipendenza con i simboli precedenti. Il messaggio che la sorgente emette è $\sigma = \{x_1, \dots, x_n\}$ e, essendo una sorgente con generica memoria k si ha che per caratterizzare l'informazione media dobbiamo andare a considerare le probabilità condizionate

$$p(x_1, \dots, x_n|x_{n-1}, \dots, x_{n-k})$$

Inoltre se k è la memoria della sorgente di Markov e n è l'estensione del messaggio si ha che

$$\mu \coloneqq \left\lceil \frac{k}{n} \right\rceil \tag{1.6.5}$$

è la memoria della sorgente estesa X^n (si ha quindi che se $k \leq n$ vale sempre $\mu = 1$). Anche per le sorgenti con memoria di Markov vale la stessa relazione per quelle senza memoria, ovvero

$$H(X^n) = nH(X) \tag{1.6.6}$$

Dim: Dimostriamolo per una sorgente di ordine $k = 1$ e per la sua n-esima estensione X^n. Essendo $n \geq k$ allora anche X^n ha memoria $\mu = 1$.
Chiamiamo σ_n il messaggio $\sigma_n = \{x_1, \dots, x_n\}$ e σ_n^m il messaggio precedente, ovvero la sua memoria $\sigma_n^m = \{x_1^m, \dots, x_n^m\} = x_n^m$ dal momento che la sorgente è di ordine 1. Vale quindi

$$H(X^n) = \sum_{\sigma_n \in X^n} \sum_{x_n^m \in X} p(\sigma_n, x_n^m) \log \frac{1}{p(\sigma_n|x_n^m)}$$

Inoltre $p(\sigma_n|x_n^m) = p(x_1, \dots, x_n|x_n^m) = p(x_1|x_n^m)p(x_2|x_1) \dots p(x_n|x_{n-1})$ per la proprietà di Markov per cui vale

$$H(X^n) = \sum_{\sigma_n \in X^n} \sum_{x_n^m \in X} p(\sigma_n, x_n^m) \log \frac{1}{p(x_1|x_n^m)p(x_2|x_1) \dots p(x_n|x_{n-1})} =$$
$$= \sum_{\sigma_n \in X^n} \sum_{x_n^m \in X} p(\sigma_n, x_n^m) \left[\log \frac{1}{p(x_1|x_n^m)} + \dots \log \frac{1}{p(x_n|x_{n-1})} \right]$$

che è la somma di n termini tutti della stessa forma. Vediamo il primo:

$$\sum_{\sigma_n \in X^n} \sum_{x_n^m \in X} p(\sigma_n, x_n^m) \log \frac{1}{p(x_1|x_n^m)} = \sum_{\{x_1,\dots,x_n\} \in X^n} \sum_{x_n^m \in X} p(x_1, \dots, x_n, x_n^m) \log \frac{1}{p(x_1|x_n^m)} =$$
$$= \left[\sum_{x_1 \in X} \sum_{x_n^m \in X} p(x_1, x_n^m) \log \frac{1}{p(x_1|x_n^m)} \right] \left[\sum_{\{x_2,\dots,x_n\} \in X^{n-1}} \sum_{x_n^m \in X} p(x_2, \dots, x_n^m) \right] =$$
$$= \sum_{x_1 \in X} \sum_{x_n^m \in X} p(x_1, x_n^m) \log \frac{1}{p(x_1|x_n^m)} = H(X)$$

dove l'ultima uguaglianza vale perchè è la definizione di entropia di una sorgente di Markov di ordine 1 (anche se in una notazione differente, si veda 1.6.4). Da questa considerazione si ha

$$H(X^n) = nH(X)$$

□

1.6.2 Sorgente aggiunta di una sorgente di Markov

Sia, come prima, $X = \{x_1, \dots, x_M\}$ l'alfabeto di una sorgente di Markov di ordine k e siano $p(x_1), \dots, p(x_M)$ le probabilità (incondizionate) di emissione dei rispettivi simboli. La **sorgente aggiunta** di X, indicata con $\bar{X}$ è una sorgente *senza memoria* con lo stesso alfabeto di X. Si ha che vale sempre

$$H(X) \leq H(\bar{X}) \tag{1.6.7}$$

ovvero che *i legami tra i simboli riducono l'entropia della sorgente.*

Dim: Si dimostra nel caso di memoria $k = 1$, sfruttando la proprietà di positività dell'entropia relativa: $-D(\cdot||\cdot) \leq 0$. Detti $x_i, x_{i-1} \in X$ due simboli emessi dalla sorgente di Markov si ha

$$- D(p(x_i, x_{i-1})||p(x_i)p(x_{i-1})) = \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \log \frac{p(x_i)p(x_{i-1})}{p(x_i, x_{i-1})} =$$
$$= \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \log \frac{p(x_i)p(x_{i-1})}{p(x_i|x_{i-1})p(x_{i-1})} = \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \log \frac{p(x_i)}{p(x_i|x_{i-1})} =$$
$$= \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \left[\log p(x_i) + \log \frac{1}{p(x_i|x_{i-1})} \right] =$$
$$= \sum_{x_i} p(x_i) \log p(x_i) + \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \log \frac{1}{p(x_i|x_{i-1})} =$$
$$= - \sum_{x_i} p(x_i) \log \frac{1}{p(x_i)} + \sum_{x_i} \sum_{x_{i-1}} p(x_i, x_{i-1}) \log \frac{1}{p(x_i|x_{i-1})} =$$
$$= -H(\bar{X}) + H(X) \leq 0$$

Dove l'ultima uguaglianza vale perchè il termine

$$\sum_{x\in X} p(x)\log\frac{1}{p(x)} = H(\bar{X})$$

corrisponde proprio all'entropia della sorgente di Markov nel caso incondizionato (senza memoria) e l'altro termine per definizione (1.6.4) corrisponde all'entropia di una sorgente di Markov di ordine 1.
□

Seguendo l'esempio 2 appena svolto si avrebbe quindi che, calcolando le marginali,

$$\begin{cases} p(x_i=0) = 4/14+1/14+1/14+1/14 = 7/14 = 1/2 \\ p(x_i=1) = 1/14+1/14+1/14+4/14 = 7/14 = 1/2 \end{cases} \implies H(\bar{X}) = 1 \quad bit$$

che in effetti risulta maggiore dell'entropia precedentemente calcolata $H(\bar{X}) = 1 > 0.80 = H(X)$.

1.6.3 Sorgente aggiunta di una sorgente estesa di Markov

Si considera ora $\overline{X^n}$: la sorgente estesa dell'estensione n-esima di una sorgente di Markov di ordine $k=1$. Questa sorgente emette messaggi $\sigma_n = \{x_1,\dots,x_n\} \in X^n$ con probabilità indipendenti dal passato, senza memoria. L'entropia di questa sorgente è definita come

$$H(\overline{X^n}) = \sum_{\sigma_n\in X^n} p(\sigma_n)\log\frac{1}{p(\sigma_n)} = \sum_{X^n} p(x_1,\dots,x_n)\log\frac{1}{p(x_1,\dots,x_n)} \tag{1.6.8}$$

e vale

$$H(\overline{X^n}) = H(\bar{X}) + (n-1)H(X) = nH(X) + [H(\bar{X}) - H(X)] \tag{1.6.9}$$

Dim: Dal Teorema di Bayes si ha che

$$p(x_1,\dots,x_n) = p(x_n|x_1,\dots,x_{n-1})p(x_{n-1}|x_1,\dots,x_{n-2})\dots p(x_2|x_1)p(x_1)$$

ed essendo X una sorgente del primo ordine vale

$$p(x_1,\dots,x_n) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2})\dots p(x_2|x_1)p(x_1)$$

allora

$$\begin{aligned} H(\overline{X^n}) &= \sum_{X^n} p(x_1,\dots,x_n)\log\frac{1}{p(x_1,\dots,x_n)} = \\ &= \sum_{X^n} p(x_1,\dots,x_n)\log\frac{1}{p(x_1)p(x_2|x_1)p(x_3|x_2)\dots p(x_n|x_{n-1})} = \\ &= \sum_{X^n} p(x_1,\dots,x_n)\Big[\log\frac{1}{p(x_2|x_1)} + \dots + \log\frac{1}{p(x_1)}\Big] = \\ &= \sum_{X^n} p(x_1,\dots,x_n)\Big[\log\frac{1}{p(x_2|x_1)} + \dots + \log\frac{1}{p(x_n|x_{n-1})}\Big] + \sum_{X^n} p(x_1,\dots,x_n)\log\frac{1}{p(x_1)} = \end{aligned}$$

$$= \sum_{x_1,x_2} p(x_1,x_2)\log\frac{1}{p(x_2|x_1)} + \cdots + \sum_{x_n,x_{n-1}} p(x_n,x_{n-1})\log\frac{1}{p(x_n|x_{n-1})} + \sum_{x_1} p(x_1)\log\frac{1}{p(x_1)}$$

in cui tutti i primi $(n-1)$ termini equivalgono all'entropia della sorgente X di ordine 1 con memoria mentre l'ultimo termine equivale all'entropia della sorgente $\bar{X}$ senza memoria:

$$H(\overline{X^n}) = \underbrace{H(X) + \cdots + H(X)}_{n-1} + H(\bar{X}) = (n-1)H(X) + H(\bar{X}) \qquad \square$$

In generale si può dimostrare che per sorgenti di ordine k qualsiasi, detto $\epsilon_k > 0$ una costante che, se $n > k$, dipende unicamente dalla statistica della sorgente, si ha

$$H(\overline{X^n}) = nH(X) + \epsilon_k \tag{1.6.10}$$

per cui

$$\frac{H(\overline{X^n})}{n} = H(X) + \frac{\epsilon_k}{n} \tag{1.6.11}$$

che, all'aumentare della lunghezza del messaggio n, mostra come i ***vincoli di memoria perdano peso***

$$\lim_{n\to\infty} \frac{H(\overline{X^n})}{n} = H(X) \tag{1.6.12}$$

Nota bene: Si noti come l'aggiunta dell'estensione non sia equivalente all'estensione dell'aggiunta

$$H(\bar{X}^n) \neq H(\overline{X^n}) \tag{1.6.13}$$

dal momento che $\bar{X}$ è una sorgente senza memoria si ha $H(\bar{X}^n) = nH(\bar{X})$ mentre dalla 1.6.7 si ha

$$H(\overline{X^n}) \geq H(X^n) = nH(X) \tag{1.6.14}$$

Riprendendo l'esempio 2, in cui si è calcolato $H(X) = 0.80, H(\bar{X}) = 1\ bit$, possiamo calcolare ora $H(X^2) = 2H(X) = 1.6\ bit$ e

$$H(\overline{X^2}) = -2 \times \frac{5}{14}\log\frac{5}{14} - 2 \times \frac{2}{14}\log\frac{2}{14} \approx 1.86\ bit$$

mentre $H(\bar{X}^2)$ può essere calcolato prendendo la sorgente senza memoria $\bar{X}$ e ricavando la congiunta come $p(x_i, x_{i-1}) = p(x_i)p(x_{i-1})$ da cui

$$H(\bar{X}^2) = -4 \times \frac{1}{4}\log\frac{1}{4} = 2\ bit = 2H(\bar{X})$$

Si può poi calcolare $H(\overline{X^3})$ come

$$H(\overline{X^3}) = -2 \times \frac{4}{14}\log\frac{4}{14} - 6 \times \frac{1}{14}\log\frac{1}{14} \approx 2.66\ bit$$

Dalle proprietà delle catene di Markov si può poi ottenere $H(\overline{X^4}) \approx 3.47\ bit$ calcolando $P^2 = PP$ e, consequentemente, la distribuzione congiunta con il Teorema di Bayes. Si vede quindi che la

successione

$$\left\{H(\bar{X}), \frac{H(\overline{X^2})}{2}, \frac{H(\overline{X^3})}{3}, \frac{H(\overline{X^4})}{4}\right\} = \{1, 0.93, 0.89, 0.87\}$$

già per $n = 4$ si stà avvicinando a $H(X)$.

1.7 Sorgenti continue

Sia X una sorgente continua con pdf (probability density funcion) $f_X(x)$.

Definizione 9. *Entropia differenziale:* Si definisce l'entropia differenziale di X come

$$h(x) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx \tag{1.7.1}$$

se tale integrale esiste. Se l'entropia nel caso di una sorgente discreta è una quantità sempre non-negativa nel caso di una sorgente continua questo non è più vero, si ha infatti che $h(x) \in \mathbb{R}$.

Esempio 1 : Sorgente uniforme

Sia X una sorgente scalare con pdf uniforme nell'intervallo $[a, b] : f(x) \sim \mathcal{U}([a, b])$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{altrimenti} \end{cases}$$

allora

$$h(x) = \frac{1}{b-a} \log(b-a) \int_a^b dx = \log(b-a)$$

da cui si vede che, quando $b - a < 1 \implies h(x) < 0$

Esempio 2 : Sorgente Gaussiana

Sia X una sorgente Gaussiana scalare con media nulla e varianza σ^2:

$$f_X(x) \sim \mathcal{N}(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

allora

$$\begin{aligned} h(x) &= -\int_{-\infty}^{\infty} f_X(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \\ &= -\int_{-\infty}^{\infty} f_X(x) \Big(\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{x^2}{2\sigma^2}} \Big) dx = \\ &= -\Big[\log \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{\int_{-\infty}^{\infty} f_X(x) dx}_{=1} - \frac{\log e}{2\sigma^2} \underbrace{\int_{-\infty}^{\infty} f_X(x) x^2 dx}_{=\sigma^2} \Big] = \\ &= -\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2} \log e = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \log e = \\ &= \frac{1}{2} \log 2\pi\sigma^2 e \end{aligned}$$

Si può dimostrare che, data una sorgente X con pdf $f_X(x)$ e varianza σ^2, la sua entropia differenziale $h(x)$ è limitata superiormente dall'entropia differenziale di una sorgente con pdf gaussiana con la stessa varianza σ^2. In altre parole:

$$h(x) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx \leq \frac{1}{2} \log 2\pi\sigma^2 e \tag{1.7.2}$$

Dim: La dimostrazione si basa sull'uso della proprietà di non-negatività dell'entropia relativa, nella sua versione differenziale:

$$D(f_X(x)||g_X(x)) \coloneqq \int_{-\infty}^{\infty} f_X(x) \log \frac{f_X(x)}{g_X(x)} dx \geq 0 \tag{1.7.3}$$

Se prendiamo $g_X(x) \sim \mathcal{N}(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ si ha

$$\begin{aligned}
-D(f_X(x)||g_X(x)) &= \int_{-\infty}^{\infty} f_X(x) \log \frac{g_X(x)}{f_X(x)} dx = \\
&= \int_{-\infty}^{\infty} f_X(x) \log g_X(x) dx - \overbrace{\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx}^{=-h(x)} = \\
&= h(x) + \int_{-\infty}^{\infty} f_X(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \\
&= h(x) + \int_{-\infty}^{\infty} f_X(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx + \int_{-\infty}^{\infty} f_X(x) \log e^{-\frac{x^2}{2\sigma^2}} dx = \\
&= h(x) + \log \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} f_X(x) dx - \frac{\log e}{2\sigma^2} \int_{-\infty}^{\infty} f_X(x) x^2 dx = \\
&= h(x) + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \log e = \\
&= h(x) - \frac{1}{2} \log 2\pi\sigma^2 e \leq 0
\end{aligned}$$

□

Si ha quindi che *a parità di varianza le sorgenti gaussiane generano la massima informazione media.*

2 Compressione di Sorgente

Il nostro obiettivo è quello di associare ad ogni simbolo informativo una sequenza di codice (di solito una sequenza binaria). Nel fare questo consideriamo sempre una sorgente discreta di informazione, ovvero un qualsiasi elemento del sistema in grado di produrre una successione di simboli informativi ed un canale ideale, totalmente affidabile, che quindi non altera l'informazione trasmessa (a ciascun simbolo inviato corrisponderà esattamente lo stesso simbolo ricevuto). Di conseguenza l'operazione di decodifica ricostruirà esattamente i simboli della sorgente discreta originaria. Esistono due tipi di codifiche di sorgente:

- **Codifica lossy** (con perdita): si applica soprattutto a sorgenti audio e video, si sfruttano fenomeni percettivi per ridurre la quantità di bit da trasmettere. La cascata dei processi di compressione/decompressione porta in uscita ad un flusso digitale diverso da quello in ingresso. Tuttavia, si fa in modo che la perdita sia tollerabile (o addirittura irrilevante) per l'utente finale. La compressione con perdite ha il vantaggio che può ottenere tassi di compressione molto più elevati rispetto a quella senza perdite.
- **Codifica lossless** (senza perdita): si applica *solo a sorgenti discrete* e lo scopo della codifica di sorgente è quello di permettere la ricostruzione *integrale* di quanto trasmesso, che dunque viene detto senza perdita di informazione. La cascata dei processi di compressione/decompressione porta esattamente allo stesso flusso digitale che si ha in ingresso.

Noi ci occuperemo della codifica lossless.

2.1 Tipi di codice

Un codice è una mappatura di una sequenza di simboli di una sorgente S con alfabeto $X = \{x_1, \ldots, x_n\}$ in una sequenza di simboli appartenenti ad un altro alfabeto $B = \{b_1, \ldots, b_m\}$ di qualsiasi lunghezza.

Definizione 10. *Codice:* Un codice è una funzione $C(\cdot)$ che associa ad ogni simbolo della sorgente una stringa di elementi appartenenti ad un alfabeto B:

$$C : X \to B^* \tag{2.1.1}$$

Dove B^* rappresenta l'insieme di tutte le stringhe sull'alfabeto B. Dato un simbolo $x \in X$ la parola di codice (*codeword*) associata si denota con $C(x)$.

Ovviamente non tutti i codici vanno bene, ce ne sono alcuni che sono preferibili, ed in particolare ciò che ci interessa di un codice è:

- La **non ambiguità**: ovvero la possibilità del ricevitore di ricostruire la successione di simboli trasmessi della sorgente S, senza ambiguità.
- L'**efficienza**: minore è la lunghezza media del codice utilizzato meglio è, dovendo trasmettere meno bit.

Definizione 11. *Lunghezza media:* La lunghezza media di un codice, misurata in [*bit/simbolo*] è definita come:

$$L = \mathbb{E}_p[l(x)] = \sum_{x \in X} p(x) l(x) \tag{2.1.2}$$

dove $l(x)$ è la lunghezza della parola di codice associata ad x. La media viene ovviamente fatta sulla distribuzione della sorgente. Vedremo in seguito come questa quantità sia legata all'efficienza di un codice.
Oltre all'efficienza un codice deve garantire la non ambiguità: non tutti i codici però la garantiscono[4] e quindi dobbiamo andare a considerare solo il sottoinsieme di codici per cui si verifica questa proprietà. L'ambiguità è data dalla possibilità di associare la stessa parola di codice a due simboli diversi. Rendendo quindi necessaria l'imposizione che il codice sia una funzione iniettiva, si definiscono i **codici non-singolari**:

Definizione 12. *Codice non-singolare:* Un codice viene detto non-singolare quando:

$$\forall x_i, x_j \in X : x_i \neq x_j \implies C(x_i) \neq C(x_j) \tag{2.1.3}$$

Tuttavia avere un codice non singolare non è sufficiente a garantire la non ambiguità, dovendo anche tenere in conto delle concatenazioni definiamo l'estensione k-esima del codice C:

Definizione 13. *Estensione k-esima del codice C:* C^k È una funzione dall'insieme delle stringhe di k simboli della sorgente all'insieme delle parole sull'alfabeto B, data dalla concatenazione delle singole parole di codice:

$$C^k : X^k \to B^*, \quad C^k(x_1 x_2 \dots x_k) = C(x_1)C(x_2)\dots C(x_k) \tag{2.1.4}$$

Quindi l'estensione k-esima del codice mappa una sequenza di k simboli della sorgente S in una sequenza di k parole di codice.

Definizione 14. *Codice univocamente decodificabile (UD):* Un codice si dice **univocamente decodificabile** se l'estensione k-esima del codice è non-singolare, per ogni valore di k.

È possibile verificare se un codice è univocamente decodificabile controllando "prefissi" e "suffissi". Prendendo una generica parola di codice $C(x) = (b_1 b_2 \dots b_p)$ qualsiasi sequenza $(b_1 \dots b_k)$ di $k < p$ simboli è un *prefisso* mentre la sequenza $(b_{k+1} \dots b_p)$ prende il nome di *suffisso*. La procedura per determinare se il codice è UD è la seguente:
Si prendono *tutte le possibili coppie di parole di codice* e si controlla se qualcuna è il prefisso di un'altra. Nel caso si aggiunge il suffisso alle parole di codice. Si ripete fino a che:

- Uno dei suffissi aggiunti è una parola di codice: Il codice **non è UD**.
- Non ci sono più suffissi da aggiungere: Il codice **è UD**.

[4] Si pensi al caso banale in cui $X = \{a, b, c\}$ e $C(a) = C(b) = C(c) = 0$: in ogni caso si riceve 0 e non è possibile risalire al simbolo inviato.

Esempio 1 : Codici UD

Vediamo due codici $C : \{a, b, c\} \to \{0, 1\}^*$

- $C_1 = \{0, 01, 11\}$
  1. 0 è prefisso per 01, dobbiamo quindi aggiungere il suffisso $1 : C_1 = \{0, 01, 11, \mathit{1}\}$.
  2. 1 è prefisso per 11, dobbiamo quindi aggiungere il suffisso $\mathit{1}$, che è già presente.
  3. Non ci sono più suffissi da aggiungere quindi il codice è UD.
- $C_2 = \{0, 01, 10\}$
  1. 0 è prefisso per 01, dobbiamo quindi aggiungere il suffisso $1 : C_2 = \{0, 01, 10, \mathit{1}\}$.
  2. 1 è prefisso per 10, dobbiamo quindi aggiungere il suffisso $0 : C_2 = \{0, 01, 10, \mathit{1}, \mathit{0}\}$.
  3. Il suffisso $\mathit{0}$ è uguale ad una parola di codice, quindi il codice non è UD.

C'è però un'altra caratteristica che ci interessa, ovvero la velocità di decodifica. Esistono infatti codici non singolari e UD che però hanno il difetto che per poter rilevare quello che è stato trasmesso senza ambiguità richiedono di attendere la ricezione di un certo numero di simboli consecutivi. Questo *introduce un ritardo di decodifica.*
Un codice si dice **istantaneo** (*prefix-free*) se è possibile decodificare qualunque parola di codice della sequenza senza fare riferimento alle successive parole di codice. Esiste una condizione necessaria e sufficiente che esprime questa definizione:

Definizione 15. *Codice istantaneo:* Un codice si dice istantaneo quando nessuna parola di codice è il prefisso di un'altra. Nel codice istantaneo nessuna parola è prefisso di un'altra quindi non ci sono suffissi che possono essere parole di codice (dal momento che non esistono suffissi).

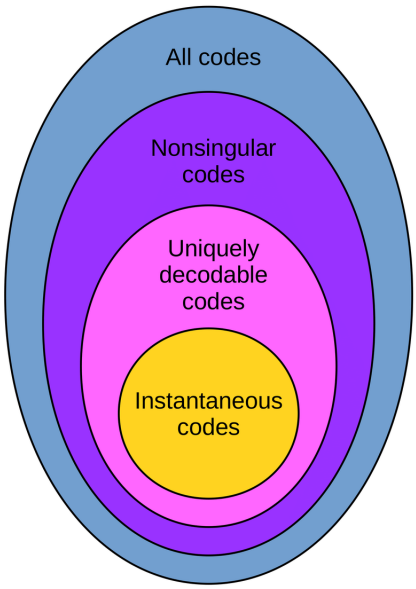


Figura 5: Gerarchia dei codici.

Supponendo di avere una sorgente S che emette simboli sull'alfabeto $X = \{a, b, c, d\}$ esaminiamo quattro codici $C : X \to \{0, 1\}^*$ e le loro proprietà.

X	Singolare	Non-singolare	UD	Istantaneo
a	0	0	10	0
b	0	010	00	10
c	0	01	11	110
d	0	10	110	111
	C_1	C_2	C_3	C_4

In particolare C_3 è UD ma non istantaneo perché non rispetta la regola del prefisso: se si riceve 11 si deve aspettare i bit successivi per sapere se si è effettivamente ricevuto un c.

2.2 Disuguaglianza di Kraft-McMillan

Abbiamo detto che per la codifica di sorgente si è interessati solo ai codici istantanei, ma per ottenerli si devono rispettare dei vincoli sulla lunghezza dei codici. Non è sempre possibile costruire codici UD e istantanei, ci sono delle condizioni che devono essere verificate.
Data una sorgente S con alfabeto $X = \{x_1, x_2, \dots, x_N\}$ e un codice C su un alfabeto[5] $B = \{b_1, b_2, \dots, b_R\}$ che associa ad ogni simbolo x_i un codice $C(x_i)$ di lunghezza $l_i = l(x_i)$ si ha una

Condizione Necessaria e Sufficiente (***Disuguaglianza di Kraft-McMillan***):
Affinchè esista un codice UD e istantaneo con parole di codice lunghe $l_1, l_2, \dots, l_N$ deve essere:

$$\sum_{i=1}^{N} R^{-l_i} \leq 1 \tag{2.2.1}$$

Si vede che tale disuguaglianza *considera solo le lunghezze dei codici e niente altro*. Inoltre, è importante sottolineare che questo teorema dice solo che *se* questa condizione è soddisfatta sicuramente *esiste* un codice univocamente decodificabile ed istantaneo con queste lunghezze, ma *non dice* che *tutti* i codici con queste lunghezze sono univocamente decodificabili ed istantanei, dice solo che queste lunghezze sono corrette per costruire un univocamente decodificabile ed istantaneo.

Dim:
Dimostriamo prima la *condizione sufficiente*, ovvero che, dati $R, l_1, l_2, \dots, l_N$:

$$\sum_{i=1}^{N} R^{-l_i} \leq 1 \implies \exists \text{ un codice istantaneo con queste lunghezze.}$$

Possiamo pensare, senza perdita di generalità, di ordinare le lunghezze l_i in ordine crescente e di chiamare l la lunghezza massima $l \coloneqq \max\{l_1, l_2, \dots, l_N\}$ con $l > 0$. Quindi si avranno:

- n_1 parole di codice lunghe 1, con $n_1 \geq 0$.
- n_2 parole di codice lunghe 2, con $n_2 \geq 0$.
- ...
- n_{l-1} parole di codice lunghe $l-1$ con $n_{l-1} \geq 0$.
- n_l parole di codice lunghe l, con $n_l > 0$.

Quindi possiamo riscrivere la disuguaglianza come:

$$\sum_{i=1}^{N} R^{-l_i} = \sum_{i=1}^{l} n_i R^{-i} \leq 1$$

Moltiplicando entrambi i membri per R^l si ha che

[5] Nel caso in cui l'alfabeto di codifica sia binario $B = \{0, 1\}$ e $|B| = R = 2$.

$$R^l \sum_{i=1}^{l} n_i R^{-i} = \sum_{i=1}^{l} n_i R^{l-i} \leq R^l$$

decomponendo la somma si ha

$$n_1 R^{l-1} + n_2 R^{l-2} + \cdots + n_{l-2} R^2 + n_{l-1} R + n_l \leq R^l$$

isolando n_l si ottiene

$$n_l \leq R^l - n_1 R^{l-1} - n_2 R^{l-2} - \cdots - n_{l-1} R$$

ma, essendo $n_l > 0$, anche

$$0 < R^l - n_1 R^{l-1} - n_2 R^{l-2} - \cdots - n_{l-1} R$$

da cui

$$\begin{aligned} n_{l-1} R &< R^l - n_1 R^{l-1} - n_2 R^{l-2} - \cdots - n_{l-2} R^2 \\ n_{l-1} &< R^{l-1} - n_1 R^{l-2} - n_2 R^{l-3} - \cdots - n_{l-2} R \end{aligned}$$

Ricordando che $n_{l-1} \geq 0$ si ha

$$\begin{aligned} 0 &< R^{l-1} - n_1 R^{l-2} - n_2 R^{l-3} - \cdots - n_{l-2} R \\ n_{l-2} R &< R^{l-1} - n_1 R^{l-2} - n_2 R^{l-3} - \cdots - n_{l-3} R^2 \\ n_{l-2} &< R^{l-2} - n_1 R^{l-3} - n_2 R^{l-4} - \ldots n_{l-3} R \end{aligned}$$

Proseguendo in questo modo fino a n_1 si ottiene

$$\begin{cases} n_l \leq R^l - n_1 R^{l-1} - n_2 R^{l-2} - \cdots - n_{l-1} R \\ n_{l-1} < R^{l-1} - n_1 R^{l-2} - n_2 R^{l-3} - \cdots - n_{l-2} R \\ n_{l-2} < R^{l-2} - n_1 R^{l-3} - n_2 R^{l-4} - \ldots n_{l-3} R \\ \cdots \\ n_3 < R^3 - n_1 R^2 - n_2 R \\ n_2 < R^2 - n_1 R \\ n_1 < R \end{cases}$$

Quindi, ricapitolando, se vale la condizione di Kraft-McMillan sulle lunghezze l_i si ha che valgono queste disequazioni, che altro non sono che la definizione alternativa di un codice istantaneo.
Dimostriamo ora la *condizione necessaria*, ovvero che per un codice univocamente decodificabile (quindi anche per uno istantaneo[5]) con lunghezze $\{l_1, l_2, \ldots, l_N\} \implies \sum_{i=1}^{N} R^{-l_i} \leq 1$. Si ha

$$\Big(\sum_{i=1}^{N} R^{-l_i} \Big)^n = \underbrace{\sum_{i=1}^{N} R^{-l_i} \sum_{j=1}^{N} R^{-l_j} \cdots \sum_{k=1}^{N} R^{-l_k}}_{n} = \sum_{i=1}^{N} \sum_{j=1}^{N} \cdots \sum_{k=1}^{N} R^{-(l_i + l_j + \cdots + l_k)}$$

Si ha che $p \coloneqq l_i + l_j + \dots l_k$ non è altro che la somma delle lunghezze di n parole di codice. Chiamando come prima $l \coloneqq \max\{l_1, l_2, \dots, l_N\}$ si ha che $p \in [n, nl]$ (ovvero è compreso tra la somma di n parole lunghe 1 ed n parole lunghe l) da cui, chiamando L_p il numero di concatenazioni di n parole di codice le cui lunghezze sommate danno p, si ha:

$$\Big(\sum_{i=1}^{N} R^{-l_i}\Big)^n = \sum_{i=1}^{N}\sum_{j=1}^{N}\cdots\sum_{k=1}^{N} R^{-(l_i+l_j+\cdots+l_k)} = \sum_{p=n}^{nl} L_p R^{-p}$$

Ma la concatenazione di n parole di codice non è altro che la parola di codice di un messaggio di n simboli, cioè un codice dell'estensione n-esima della sorgente S^n. Per definizione di univocamente decodificabile, quindi, qualunque sia l'estensione della sorgente $(\forall n)$ si ha che presi due codici questi devono essere diversi.
Questo implica che, se abbiamo L_p codici di lunghezza p, per fare in modo che siano diversi tra loro deve valere

$$L_p \leq R^p$$

Da cui:

$$\Big(\sum_{i=1}^{N} R^{-l_i}\Big)^n = \sum_{p=n}^{nl} L_p R^{-p} \leq \sum_{p=n}^{nl} R^p R^{-p} = \sum_{p=n}^{nl} 1 = nl - n + 1 \leq nl$$

quindi

$$\Big(\sum_{i=1}^{N} R^{-l_i}\Big)^n \leq nl \implies \sum_{i=1}^{N} R^{-l_i} \leq 1$$

Infatti se fosse $\sum_{i=1}^{N} R^{-l_i} > 1$ (quindi $\sum_{i=1}^{N} R^{-l_i} = 1 + \epsilon$) si avrebbe che, ad esempio

$$\lim_{n\to\infty} \frac{(1+\epsilon)^n}{n} = \infty \not\leq l$$

□

2.3 Codici compatti

Siamo adesso interessati a definire codici *efficienti*. Iniziamo con una definizione:

Definizione 16. *Codice compatto:* Un codice $\mathcal{C}$ si definisce compatto se:

1. È univocamente decodificabile.
2. La sua lunghezza media $L_{\mathcal{C}}$ è la minore tra tutti gli altri codici UD per la sorgente S sull'alfabeto B.
$$L_{\mathcal{C}} \leq L_C, \quad \forall C : X \to B^*$$

Per trovare codici compatti dobbiamo capire quale è la *minima lunghezza possibile* di un codice. Assumendo momentaneamente che la sorgente S sia una DMS (rilasseremo poi quest'ipotesi alle

sorgenti di Markov) abbiamo un risultato molto importante. Si ha che, detta $H_b(S)$ l'entropia della sorgente S in base b,

$$\frac{H(S)}{\log b} = H_b(S) \leq L \tag{2.3.1}$$

ovvero che l'entropia della sorgente rappresenta un limite inferiore per la lunghezza media di un codice univocamente decodificabile. Questo è un primo importante risultato perché *lega la definizione di informazione/entropia con una quantità che non dipende dalla definizione stessa di informazione.*
Se, per un codice UD $\mathcal{C}$, vale il limite inferiore $H_b(S) = L_{\mathcal{C}}$ allora il codice è un **codice compatto**. Questo avviene quando $\forall x \in X, \exists l \in \mathbb{N}$:

$$p(x) = b^{-l}$$

nel caso in cui $b = 2$ la sorgente si dice *diadica.*

Dim: Dimostriamo che $H(S) - L \leq 0$, il caso generale non è molto diverso.

$$\begin{aligned}
H(S) - L &= \sum_{x\in X} p(x) \log \frac{1}{p(x)} - \sum_{x\in X} p(x) l(x) = \\
&= \sum_{x\in X} p(x) \log \frac{1}{p(x)} - \sum_{x\in X} p(x) \log 2^{l(x)} = \\
&= \sum_{x\in X} p(x) \log \frac{2^{-l(x)}}{p(x)} = \log e \sum_{x\in X} p(x) \ln \frac{2^{-l(x)}}{p(x)} \leq \\
&\leq \log e \sum_{x\in X} p(x) \Big(\frac{2^{-l(x)}}{p(x)} - 1 \Big) = \\
&= \log e \Bigg(\underbrace{\sum_{x\in X} 2^{-l(x)}}_{\leq 1} - \sum_{x\in X} p(x) \Bigg) \leq 0
\end{aligned}$$

Inoltre l'uguaglianza vale quando le due disuguaglianze valgono con l'uguaglianza, cioè quando:

$$\begin{cases} \frac{2^{-l(x)}}{p(x)} = 1 \iff p(x) = 2^{-l(x)}, \ \forall x \in X \\ \sum_{x\in X} 2^{-l(x)} = 1 \end{cases}$$

□

Quindi, per una DMS in cui le probabilità di emissione siano della forma (generale), $\forall x \in X$

$$p(x) = b^{-l(x)} = \Big(\frac{1}{b}\Big)^{l(x)}$$

si può avere una lunghezza di codice (istantaneo) *minima pari all'entropia della sorgente stessa.* Ovviamente avere probabilità di questa forma non è realistico. Consideriamo allora il caso di una sorgente DMS in cui i valori di $p(x)$ siano arbitrari, quindi non rispettano nessun vincolo. Questo implica che $H_b(S) < L$, ovvero non si può più avere un codice di lunghezza minima possibile.

Definizione 17. *Efficienza:* Si definisce l'efficienza η di un codice con lunghezza media L per una sorgente S come:

$$\eta \coloneqq \frac{H_b(S)}{L} \tag{2.3.2}$$

Quanto più questo numero (puro) si avvicina a 1 tanto maggiore è l'efficienza della codifica.

2.4 Primo Teorema di Shannon

Shannon ha studiato come cercare di avvicinarsi al limite minimo anche in caso in cui le probabilità di emissione dei simboli siano arbitrarie. Ha pensato, nel caso in cui i valori di $p(x)$ siano arbitrari, di approssimare $l(x)$ con l'intero maggiore più vicino a quello di una sorgente diadica[6], cioè, $\forall x \in X$:

$$l_S(x) \coloneqq \left\lceil \log_b \frac{1}{p(x)} \right\rceil \tag{2.4.1}$$

Si ha ovviamente che, $\forall x \in X$

$$l(x) \leq l_S(x) \leq l(x) + 1 \tag{2.4.2}$$

e che il codice costruito con queste lunghezze è istantaneo, soddisfando la disugaglianza di Kraft-McMillan. Considerando infatti la prima delle due disugaglianze $l(x) \leq l_S(x)$ vale

$$\begin{aligned} l(x) = \log_b \frac{1}{p(x)} \leq l_S(x) \implies \frac{1}{p(x)} \leq b^{l_S(x)} \implies \\ \implies p(x) \geq b^{-l_S(x)} \implies \sum_{x\in X} p(x) = 1 \geq \sum_{x\in X} b^{-l_S(x)} \end{aligned}$$

Inoltre si ha che

$$H_b(S) \leq L \leq H_b(S) + 1 \tag{2.4.3}$$

Dim: Dimostriamo nel caso $b = 2$. Si ha

$$\log \frac{1}{p(x)} \leq l_S(x) \leq \log \frac{1}{p(x)} + 1 \text{ da cui}$$

$$\underbrace{\sum_{x\in X} p(x) \log \frac{1}{p(x)}}_{=H(S)} \leq \underbrace{\sum_{x\in X} p(x) l_S(x)}_{=L} \leq \sum_{x\in X} p(x)\Big(\log \frac{1}{p(x)} + 1\Big) = H(S) + 1$$

□

2.4.1 Sorgenti discrete senza memoria

Shannon ha osservato che, estendendo la sorgente e codificando i messaggi invece dei singoli simboli la lunghezza media per codificare un simbolo si riduce. Basandoci su questa osservazione si consideri la sorgente estesa S^n in cui vengono inviati messaggi σ_i di n simboli appartenenti all'alfabeto $X = \{x_1, x_2, \ldots, x_M\}$:

$$\sigma_i = \{x_{i_1} x_{i_2} \ldots x_{i_n}\}$$

[6] In una sorgente diadica si ha, equivalentemente, $l(x) = \log_b p(x)^{-1} = -\log_b p(x)$

Si hanno quindi M^n possibili messaggi σ emittibili dalla sorgente estesa S^n caratterizzati da una probabilità $p(\sigma)$. Scegliendo, seguendo la codifica di Shannon, un codice di lunghezza

$$\lambda(\sigma) = \left\lceil \log \frac{1}{p(\sigma)} \right\rceil \tag{2.4.4}$$

si ha, come prima

$$\log \frac{1}{p(\sigma)} \leq \lambda(\sigma) \leq \log \frac{1}{p(\sigma)} + 1$$

da cui, mediando su tutti gli M^n messaggi $\sigma \in S^n$, si ottiene

$$\sum_{\sigma \in S^n} p(\sigma) \log \frac{1}{p(\sigma)} \leq \sum_{\sigma \in S^n} p(\sigma)\lambda(\sigma) \leq \sum_{\sigma \in S^n} p(\sigma)\Big(\log \frac{1}{p(\sigma)} + 1 \Big) \tag{2.4.5}$$

da cui

$$H(S^n) \leq L_n \leq H(S^n) + 1 \tag{2.4.6}$$

in cui si è definito

$$L_n \coloneqq \sum_{\sigma \in S^n} p(\sigma)\lambda(\sigma) \tag{2.4.7}$$

ovvero la lunghezza media del codice per simbolo esteso (il numero medio di simboli dell'alfabeto di codice) quando si mandano e si decodificano n simboli consecutivi. Ricordando poi che $H(S^n) = nH(S)$ si ottiene

$$H(S) \leq \frac{L_n}{n} \leq H(S) + \frac{1}{n} \tag{2.4.8}$$

Si ha quindi che L_n/n, cioè il numero medio di simboli del codice usati per codificare un blocco di n simboli, tende a $H(S)$ all'aumentare di n.
Si ha quindi il **Primo Teorema di Shannon per una DMS**: *possiamo codificare una sorgente senza memoria con una lunghezza media per simbolo vicina a piacere al limite minimo (l'entropia della sorgente) codificando più simboli insieme invece che uno solo.*
Il prezzo che si paga è la complessità, dovendo codificare e decodificare n simboli consecutivi lavorando con la sorgente estesa S^n.

2.4.2 Sorgenti di Markov

Sia ora S una sorgente discreta con memoria di Markov. Quanto visto precedentemente per le sorgenti DMS in merito alla minima lunghezza di codifica vale anche per le sorgenti con memoria. Se infatti consideriamo $\bar{S}$, la sorgente aggiunta di S, essendo questa senza memoria vale quando già visto ovvero $H(\bar{S}) \leq L$. Ma vale anche $H(S) \leq H(\bar{S})$ da cui

$$H(S) \leq L \tag{2.4.9}$$

Possiamo estendere il Primo Teorema di Shannon alle sorgenti con memoria di Markov.
Per una sorgente di Markov S di ordine 1 si ha

$$H(S) + \frac{H(\bar{S}) - H(S)}{n} \leq \frac{L_n}{n} \leq H(S) + \frac{H(\bar{S}) - H(S)}{n} + 1 \tag{2.4.10}$$

mentre in generale per una sorgente di Markov di ordine k vale, per un certo $\epsilon_k \in \mathbb{R}$

$$H(S) + \frac{\epsilon_k}{n} \leq \frac{L_n}{n} \leq H(S) + \frac{\epsilon_k}{n} + 1 \tag{2.4.11}$$

Dim: Dimostriamo nel caso di una sorgente di Markov di ordine 1.
Consideriamo l'aggiunta di S, ovvero la sorgente $\bar{S}$, si ha che, essendo questa una DMS, scelto $\forall x \in X$ un codice di Shannon dato da $l(x) = \lceil -\log p(x) \rceil$ (dove $p(x)$ è la probabilità incondizionata), vale

$$H(\bar{S}) \leq L \leq H(\bar{S}) + 1$$

Si può quindi prendere la sorgente aggiunta $\overline{S^n}$ della sorgente estesa e arrivare esattamente come prima a scrivere

$$H(\overline{S^n}) \leq L_n \leq H(\overline{S^n}) + 1$$

Ricordando poi, dalla 1.6.9, che per una sorgente di Markov del primo ordine si ha

$$H(\overline{S^n}) = nH(S) + [H(\bar{S}) - H(S)]$$

da cui

$$H(S) + \frac{H(\bar{S}) - H(S)}{n} \leq \frac{L_n}{n} \leq H(S) + \frac{H(\bar{S}) - H(S)}{n} + 1$$

Per le sorgenti di Markov di ordine superiore l'espressione è analoga perché vale

$$H(\overline{S^n}) = nH(S) + \epsilon_k, \qquad \square$$

3 Rate-Distortion Theory

Il primo teorema di Shannon si riferisce a codifica senza perdita e a sorgenti discrete. Quando si ha a che fare con la codifica lossless, non ci si deve preoccupare di come avverrà la ricostruzione dell'informazione dopo la decodifica perché sappiamo che il processo è completamente reversibile, quindi la ricostruzione in decodifica è identica all'originale. Tuttavia, Shannon ci dice che se vogliamo preservare tutta l'informazione della sorgente, la capacità di compressione ha un limite fondamentale che è dato dall'entropia.
In alcuni casi ed applicazioni la compressione lossless può andare bene, ma in altri può essere necessario aumentare il tasso di compressione accettando un certo livello di perdita di informazione, ovvero facendo una codifica *lossy*. Inoltre si deve tenere in considerazione che quando la sorgente informativa è analogica per trasformarla in numerica si effettua *sempre* una codifica di sorgente lossy.

Se nella codifica lossless l'unica metrica di interesse è il *rate di generazione* R, ovvero il **numero di bit per simbolo necessari a rappresentare la sorgente**, nella codifica lossy questo non basta, infatti se così fosse, la migliore forma di codifica sarebbe buttare via tutti i dati.
Quando si ha codifica lossy ci interessa quindi il rate ma anche la perdita di informazione, ovvero una misura della differenza tra l'informazione originaria e quella ricostruita. La perdita dell'informazione viene denominata **distorsione**.

La *distorsione* D è tanto maggiore tanto più i dati ricostruiti, a valle della compressione, distano in qualità dai dati originali. Il concetto di qualità (distorsione accettabile) non è un concetto assoluto, ma dipende necessariamente dall'*applicazione* dei dati, e cioè dall'impiego che si fa dei dati ricostruiti e da chi ne usufruisce.
Si hanno diversi modi di misurare la distorsione e la qualità di un segnale ricostruito e diversi livelli di distorsione accettabili. In ogni caso, comunque sia definita la misura della distorsione, ciò che si desidera è trovare una tecnica di codifica che per ogni fissato livello di distorsione D codifichi la sorgente al tasso R più piccolo possibile (o, al contrario, che per ogni fissato tasso di codifica R comporti la minima distorsione D).

La Rate-Distortion Theory è il ramo della teoria dell'informazione che descrive il *trade-off* tra il rate di trasmissione (bit/simbolo usati per la codifica) e la corrispondente distorsione e fornisce i limiti per la compressione con perdita. Si sottolinea come questa sia essenziale per sorgenti continue per le quali non è possibile avere una codifica lossless.
La Rate-Distortion Theory si occupa quindi di trovare le prestazioni limite teoriche, in termini di funzioni $R(D)$ e $D(R)$ per assegnate sorgenti e misure di distorsione. A livello pratico si vuole

- Definire un modo per misurare la distorsione D.
- Determinare il rate minimo (massima compressione) R a cui si può lavorare ammettendo una certa distorsione o, equivalentemente, la distorsione minima che si ha lavorando con un certo rate.

Come nel caso della codifica lossless la teoria dell'informazione fornisce dei limiti teorici che poi servono per la progettazione delle tecniche di codifica. Tuttavia in caso di codifica con perdita, anche per sorgenti piuttosto semplici, esistono pochi risultati in forma chiusa della rate-distortion theory, e si dovrà quindi ricorrere ad approssimazioni. Inoltre, anche quando le curve limite sono

note, i metodi di codifica esistenti forniscono prestazioni lontane da quelle ottime, soprattutto a causa dei vincoli di memoria e complessità computazionale ai quali si deve sottostare.

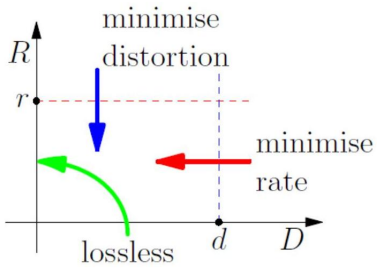


Figura 6: Obiettivo della RDT.

3.1 Distorsione

Dato che la valutazione della qualità della codifica dipende dall'uso che si deve fare dei dati codificati non è possibile definire un metodo di misura universalmente valido. Si deve scegliere allora il metodo che di volta in volta risulta più adatto all'applicazione. Sia $x \in X$ un elemento dell'alfabeto della sorgente e sia $\hat{x} \in \hat{X}$ l'elemento stesso ricostruito dal ricevitore. Dovendo quantificare la distorsione dovremo avere che questa sia, in qualche modo, funzione di una distanza $d(\cdot, \cdot)$ tra i due oggetti. Vediamo due metriche di distanza per variabili aleatorie:

1. Nel caso di sorgenti binarie una misura comunemente usata è la *Distanza di Hamming:*
$$d(x, \hat{x}) = x \oplus \hat{x} := \begin{cases} 0, & \text{se } x = \hat{x} \\ 1, & \text{se } x \neq \hat{x} \end{cases} \tag{3.1.1}$$

2. Nel caso di sorgenti continue il modo più naturale per vedere la fedeltà della ricostruzione è fare la differenza tra i valori iniziali e quelli ricostruiti. Il più popolare è l'*errore quadratico* (SE):
$$d(x, \hat{x}) = (x - \hat{x})^2$$

Ovviamente può essere definita anche in altro modo, ad esempio come la differenza assoluta tra due campioni. In generale, la cosa migliore di tutti sarebbe tener conto dell'effetto finale della distorsione, ovvero come questa viene percepita: tale valutazione non è però una misura oggettiva, ma dipendente dal contesto e quindi per poterla effettivamente quantificare si deve ricorrere a misure come l'MSE che non sono perfette ma semplici e oggettive. Ad esempio nel caso della trasmissione di segnali vocali anche se l'MSE è alto si può avere una buona percezione, ovvero che la distorsione percepita è bassa.

Definizione 18. *Distorsione:* La distorsione D è definita come
$$D = \mathbb{E}[d(X, \hat{X})] \tag{3.1.2}$$
dove la media viene fatta su tutti gli elementi dell'alfabeto X e su tutti i possibili simboli ricostruiti dell'alfabeto $\hat{X}$, quindi sulla distribuzione congiunta $p(x, \hat{x})$.
Nei due casi precedenti quindi si ha

- La probabilità di ricostruire in modo sbagliato $D = \mathbb{E}[X \oplus \hat{X}] \coloneqq P_e$
- Il Mean Squared Error (MSE) dato da $D = \mathbb{E}[(X - \hat{X})^2]$

La distorsione con la distanza di Hamming prende il nome di *probability of a reconstruction error* dal momento che $\mathbb{E}[X \oplus \hat{X}] = 0 \times Pr\{x = \hat{x}\} + 1 \times Pr\{x \neq \hat{x}\} = P_e$.
Estendendo alle sequenze di simboli $x^n, \hat{x}^n$ si ha

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{3.1.3}$$

che, nel caso di sorgenti stazionarie $(i.i.d)$, porta allo stesso valor medio

$$D = \mathbb{E}[d(X^n, \hat{X}^n) = \mathbb{E}[d(X, \hat{X})] \tag{3.1.4}$$

Si può definire anche un'altra importante grandezza: il rapporto segnalre-rumore (**SNR**). Il rapporto segnale-rumore è un numero puro o adimensionale, dato dal rapporto fra due grandezze omogenee, che esprime quanto il segnale sia *più potente* del rumore nel sistema considerato. È formalmente espresso dalla relazione:

$$SNR = \frac{\sigma_x^2}{\sigma_d^2} \tag{3.1.5}$$

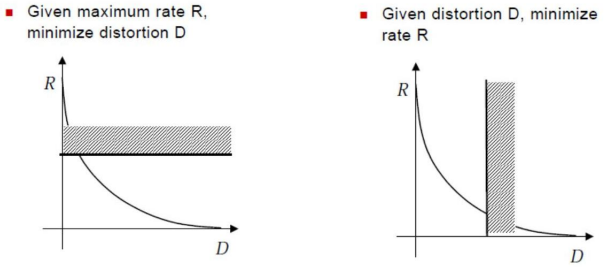
dove σ_x^2 rappresenta la potenza del segnale utile e σ_d^2 la potenza totale del rumore presente nel sistema (dato dalla distorsione). Queste vengono solitamente espresse in $Watt$ o dBm. Più basso è l'SNR, più sarà difficoltosa la decodifica del segnale ovvero più alta sarà la probabilità di errore.

3.2 Funzione di Rate-Distortion

Data una distribuzione statistica della sorgente e una misura della distorsione D, la domanda a cui si prova a rispondere è *quale è la minima distorsione D ottenibile quando viene fissata una velocità di trasmissione R?* La funzione di Rate-Distortion $R(D)$ fornisce il numero minimo di bit (cioè il rate minimo R) che si può usare per rappresentare una sorgente e che garantisce un errore di ricostruzione

$$\mathbb{E}[d(X, \hat{X})] \leq D \tag{3.2.1}$$

Il caso $D = 0$ significa che non viene accettata distorisione da cui $R(0)$ nel caso di sorgenti discrete coincide con l'entropia della sorgente $H(X)$ e, nel caso di sorgenti continue, $R(0) = \infty$.
$R(D)$ è una funzione monotona decrescente di D. Equivalentemente potremmo anche determinare l'inversa $D(R)$, rispondendo alla domanda *qual è la velocità minima (numero minimo di bit per la rappresentazione) che si può avere per garantire una data distorsione?* In questo caso la funzione inversa $D(R)$ (Distortion-Rate) fornisce il livello di distorsione del segnale riscostruito fissato il massimo numero di bit che si possono spendere per la rappresentazione.



Definizione 19. *Raggiungibilità:* Una coppia (R, D) si dice raggiungibile se esiste una codifica a rate R tale che

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \leq D \tag{3.2.2}$$

La **regione di rate-distortion** $\mathcal{R}$ rappresenta l'insieme dei punti (R, D) raggiungibili e la funzione di rate distortion è l'estremo inferiore dei rate R tali che $(R, D) \in \mathcal{R}$ per ciascun valore di D fissato. La funzione di rate-distortion $R(D)$ è quindi il *lower-bound del rate di trasmissione per un fissato valore di distorsione D:* Se $R > R(D)$ allora esiste una sequenza di codici con una distorsione media che si avvicina a D, altrimenti se il rate è nella zona inferiore, tali codici non esistono.

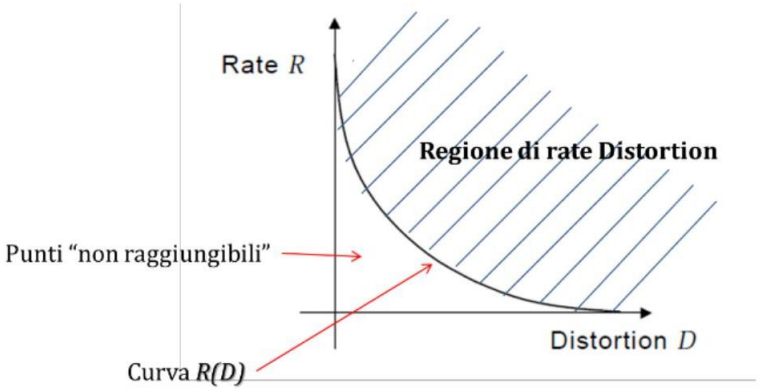


Figura 7: Regione di rate distortion.

Il principale Teorema della Rate Distortion Theory è dovuto a Shannon (1956) ed è noto come *Lossy Coding Theorem* (Teorema della Codifica con Perdita):

Teorema: Sia $(X, \hat{X}) \sim p(x, \hat{x})$ con allora la funzione $R(D)$ per la sorgente[7] X con simboli *i.i.d.* e funzione di distanza $d(x, \hat{x})$ è data da

$$R(D) = \min_{p(x|\hat{x})} I(X; \hat{X})$$
$$\text{s.t. } \mathbb{E}[d(X, \hat{X})] \leq D$$

Fissata un valore di distorsione D, il minimo rate R possibile si trova minimizzando l'informazione muta, tra tutte le coppie $(x, \hat{x})$ (e quindi tra tutte le possibili codifiche $\hat{x} = C(x)$) che soddisfano D. Si noti come il vero grado di libertà nella minimizzazione sia la distribuzione condizionata $p(\hat{x}|x)$, derivante dal fatto che la minimizzazione avvenga su $p(x, \hat{x}) = p(\hat{x}|x)p(x)$ in cui $p(x)$ dipende dalla sorgente mentre $p(\hat{x}|x)$ descrive la codifica ed è quello che ci permette di minimizzare $I(X; \hat{X})$.
L'informazione mutua tra due sorgenti ci dice quanto di una sorgente è contenuto nell'altra, in questo caso quanto dell'informazione originaria è contenuta in quella ricostruita, dopo la codifica/-decodifica con perdita.

Ricapitolando: più la codifica comprime (più si riducono i bit di rappresentazione R) e più si perde informazione. Si avrà quindi sempre meno informazione di X contenuta in $\hat{X}$ e $I(X; \hat{X})$ diminuirà.

[7] La stessa definizione vale nel caso continuo utilizzando le funzioni densità di probabilità.

Quest'ultima però può diminuire solo fino al limite in cui garantisce il vincolo sulla distorsione D, restringendo quindi l'insieme di coppie $(X, \hat{X})$ ammissibili.

Se si ha una sorgente continua e la si quantizza questa è una forma di codifica con perdita: fissata la massima distorsione che si vuole avere (che corrisponde all'errore di quantizzazione) si fissano i livelli di quantizzazione, e quindi i bit di rappresentazione, e quindi il rate.

Esempio 1 : R(D) per una sorgente gaussiana.

Sia $X \sim \mathcal{N}(0, \sigma^2)$. Per questo genere di sorgenti è ragionevole adottare la distanza a errore quadratico. La funzione di rate distortion $R(D)$ è data da

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & \text{se } 0 \leq D \leq \sigma^2 \\ 0 & \text{se } D > \sigma^2 \end{cases}$$

Dal momento che la funzione è invertibile in $[0, \sigma^2]$ si ha

$$D(R) = \sigma^2 2^{-2R}$$

infatti

$$2^{R(D)} = 2^{\frac{1}{2}\log(\frac{\sigma^2}{D})} = 2^{\log \frac{\sigma}{\sqrt{D}}} = \frac{\sigma}{\sqrt{D}} \implies \sqrt{D} = \frac{\sigma}{2^R} \implies D = \sigma^2 2^{-2R}$$

Inoltre il SNR associato alla distorsione è dato da

$$SNR = \frac{\sigma^2}{D} = 2^{2R}$$

da cui

$$SNR_{db} \approx 6R$$

ovvero, se si aumenta di 1 bit il rate (si rappresenta con un bit in più i simboli della sorgente in media) la distorsione si riduce di $\frac{1}{2^2}$ cioè di circa $6db$.

3.3 Rate-Distortion Bounds

Spesso non è possibile trovare la funzione di Rate-Distortion per sorgenti con distribuzione qualsiasi. In questi casi è quindi utile avere dei limiti (superiori ed inferiori) che forniscono comunque un range di valori "raggiungibili", che quindi possono essere presi a riferimento per la progettazione e per valutare margini di miglioramento. *In questo modo la funzione di rate-distortion gioca lo stesso ruolo per la compressione con perdita dell'entropia nel caso di quella senza perdita.*

Si può dimostrare che per una generica sorgente continua con pdf $f_X(x)$ a media nulla e varianza σ^2, se la distorsione è misurata come un MSE, allora vale

$$h(x) - \frac{1}{2}\log(2\pi e D) \leq R(D) \leq \frac{1}{2}\log\frac{\sigma^2}{D} \tag{3.3.1}$$

Per una variabile aleatoria gaussiana questi due bound coincidono.

3.4 Quantizzazione

Quando la sorgente d'informazione è continua, il processo di acquisizione dei dati è un processo con perdita di informazione: l'informazione deve essere discretizzata attraverso il del campionamento e i valori continui devono essere rappresentati con una stringa *finita* di simboli binari. La perdita di informazione dovuta alla quantizzazione è inevitabile quando si passa da sorgenti continue a sistemi digitali, però è una perdita controllata, e può essere usata per diminuire la quantità di dati da gestire (ovvero diminuire il rate del segnale).

> *"The question is: can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way?"*
> Shannon, A Mathematical Theory of Communication.

Per la progettazione di un quantizzatore quindi la questione è *come trovare la migliore possibile rappresentazione* di una sorgente continua dato un certo rate di trasmissione R. Esistono vari tipi di quantizzatori e ciascuno porta ad un certo valore di D.
In generale la quantizzazione è un processo molto semplice: si rappresentano tutti i possibili valori assunti dai simboli sorgente, con un ben più ristretto set di valori.

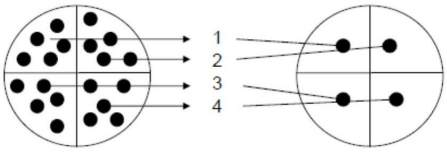


Figura 8: Perdita della iniettività.

Il quantizzatore divide il range di valori che una sorgente genera in un numero finito di intervalli ciascuno dei quali è rappresentato da un *valore di riferimento* e poi da una parola di codice (binaria): infiniti valori possono cadere nello stesso intervallo e quindi il processo è irreversibile. Conoscere la parola di codice permette di conoscere solo l'intervallo di appartenenza e non più quale degli infiniti valori iniziali fosse quello corretto. La costruzione degli intervalli di valori e di come questi sono rappresentati sono parte della progettazione del codificatore di sorgente.
È necessario quindi capire come dividere il range di ingresso in intervalli e come assegnare i codici binari a ciascun intervallo per avere un rate e/o una distorsione desiderati[8].
Supponiamo che la nostra sorgente sia caratterizzata da una pdf $f_X(x)$ e che la dinamica di ingresso venga divisa in m intervalli $[x_i, x_{i+1}), i = 1, \ldots, m$, ciascuno rappresentato dal valore di riferimento $\hat{x}_i$.

[8] Come detto prima un criterio può essere quello di prendere come misura di distorsione l'errore quadratico medio.

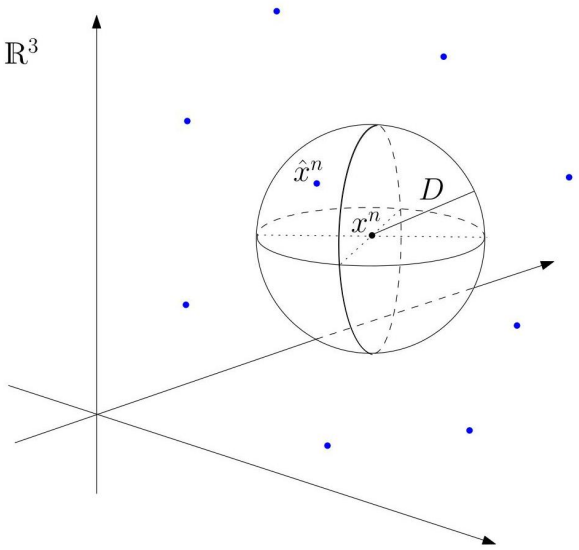


Figura 9: Rappresentazione tridimensionale della regione di quantizzazione.

L'operazione di **quantizzazione** è data da una certa funzione $Q(\cdot)$

$$\hat{x}_i = Q(x), \qquad \text{se } x_i \leq x \leq x_{i+1} \tag{3.4.1}$$

L'MSE è dato invece da

$$D = \mathbb{E}[(x - Q(x))^2] = \int_{-\infty}^{\infty} \Big(x - Q(x)\Big)^2 f_X(x)dx = \sum_{i=1}^{m} \int_{x_i}^{x_{i+1}} (x - \hat{x}_i) f_X(x)dx \tag{3.4.2}$$

La differenza tra il campione originario e quello quantizzato prende il nome di *distorsione di quantizzazione* o *rumore di quantizzazione*.

Ciascuno di questi intervalli dovrà essere poi codificato: se si usano per la codifica parole di codice della stessa lunghezza l ($l = \lceil \log m \rceil$) la progettazione del quantizzatore, dati $f_X(x)$ e m, si riduce a determinare gli intervalli x_i e i valori $\hat{x}_i$ che li rappresentano. Si potrebbero tuttavia utilizzare parole di lunghezza diversa l_i (come ad esempio nella codifica di Huffman) e allora la scelta degli intervalli *influenza anche il rate R della sorgente*:

$$R = \sum_{i=1}^{m} l_i p(\hat{x}_i) = \sum_{i=1}^{m} l_i \int_{x_i}^{x_{i+1}} f_X(x)dx \tag{3.4.3}$$

Possiamo concludere che:

- La distorsione D dipende da come vengono scelti gli intervalli e dai *valori* $\hat{x}_i$ scelti per rappresentarli.
- Il rate R dipende da come vengono scelti gli intervalli e dai *codici* usati per rappresentarli.

Si ha quindi che i problemi di trovare le migliori partizioni, rappresentazioni e codifica sono legati tra loro.

Quindi, ricapitolando:
Dato un limite di distorsione D^* (massimo tollerato) si devono determinare gli intervalli di quantizzazione $[x_i, x_{i+1})$ e i valori che li rappresentano $\hat{x}_i$ in modo che soddisfino:

$$\begin{cases} D \leq D^* \\ R = \sum_i l_i p(\hat{x}_i) = \sum_i l_i \int_{x_i}^{x_{i+1}} f_X(x)dx \end{cases} \tag{3.4.4}$$

oppure, equivalentemente, dato un limite massimo di rate R^* si devono trovare i valori degli intervalli di quantizzazione $[x_i, x_{i+1})$ e le parole di codice tali che soddisfino:

$$\begin{cases} R \leq R^* \\ D = \int_{-\infty}^{\infty} \Big(x - Q(x)\Big)^2 f_X(x)dx = \sum_i^m \int_{x_i}^{x_{i+1}} (x - \hat{x}_i) f_X(x)dx \end{cases} \tag{3.4.5}$$

3.4.1 Quantizzazione scalare

Un quantizzatore scalare, associa ad ogni valore continuo della sorgente un valore $\hat{x}_i$ appartenente ad un set discreto di m valori ciascuno dei quali rappresenta una particolare porzione dell'asse dei numeri reali $[x_i, x_{i+1})$ e $\hat{x}_i$ è un valore interno a tale intervallo. Un quantizzatore di largo impiego è quello **uniforme**, in cui tutti gli intervalli di quantizzazione sono uguali. È completamente caratterizzato dall'ampiezza Δ dell'intervallo di quantizzazione (lo step di quantizzazione) e dal numero di livelli $m = 2^R$.
Se la sorgente X è uniforme sull'intervallo $[-A, A]$ si può pensare che la quantizzazione uniforme sia la scelta più appropriata: fissato m vogliamo trovare il valore Δ che minimizza la distorsione D, ovvero l'errore di quantizzazione. Essendo la distribuzione uniforme si vede facilmente che

$$\Delta = \frac{2A}{m} \tag{3.4.6}$$

e che l'errore di quantizzazione $e_q \coloneqq |x_i - \hat{x}_i|$ è distribuito uniformemente sull intervallo $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$. La distorsione (MSE) in questo modo diventa

$$\begin{aligned} D = \mathbb{E}[(X - \hat{X})^2] &= \int_{-\infty}^{\infty} \Big(x - Q(x)\Big)^2 f_X(x)dx = \\ &= \sum_{i=1}^{m} \int_{x_i}^{x_{i+1}} (x - \hat{x}_i)^2 f_X(x)dx = \\ &= \sum_{i=1}^{m} \frac{1}{2A} \int_{x_i}^{x_{i+1}} (x - \hat{x}_i)^2 dx = \frac{m}{2A} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 dx = \\ &= \frac{1}{\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 dx = \frac{\Delta^2}{12} \end{aligned}$$

Ricordando che la varianza di una sorgente uniformemente distribuita su $[-A, A]$ è data da $A^2/3$ si ha che

$$SNR = \frac{A^2/3}{\Delta^2/12} = \frac{12A^2}{3\Delta^2} = \frac{4A^2}{\Delta^2} = m^2 = 2^{2R} \approx 6R\, db \tag{3.4.7}$$

Non sempre la quantizzazione uniforme da buoni risultati, infatti l'errore di quantizzazione che si compie dipende anche dalla statistica della sorgente. Il quantizzatore uniforme si comporta in

maniera ottima (minima potenza del rumore di quantizzazione) *solo se le ampiezze dei campioni del segnale sono caratterizzate da una distribuzione uniforme.* Se la sorgente non è uniformemente distribuita ci sono range di valori in cui è più probabile avere un campione della sorgente ed è quindi meglio aumentare la densità dei livelli di quantizzazione nelle zone più popolate, in modo da avere una maggiore «accuratezza» dove si hanno più campioni, diminuendo di conseguenza la distorsione.
Quest'ultima può essere quindi ridotta scegliendo gli intervalli di quantizzazione in funzione delle statistiche di X (tenendo conto della pdf della sorgente), il che ci porta a una **quantizzazione non uniforme**: *assegnare intervalli più stretti per le ampiezze più frequenti e intervalli di dimensioni crescenti per le ampiezze meno frequenti.* Facendo così si presta quindi più attenzione nella quantizzazione di valori che si presentano con maggiore probabilità diminuendo così l'errore di quantizzazione.
L'obiettivo della quantizzazione non uniforme è quindi trovare l'ampiezza degli intervalli ed i valori che li rappresentano tali che minimizzino la distorsione, fissato m.

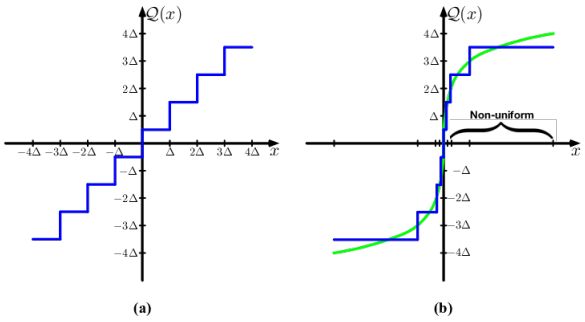


Figura 10: Quantizzazione uniforme e quantizzazione non uniforme a confronto.

3.4.2 Quantizzazione vettoriale

Il metodo più generale di quantizzazione è la **quantizzazione vettoriale**. Abbiamo già visto per la codifica lossless che codificare sequenze di simboli è più efficiente che codificare simboli isolati, nel caso della quantizzazione vettoriale si ha una sorta di dualità rispetto a questo fatto: così come la quantizzazione scalare prevede una quantizzazione separata campione per campione quella vettoriale si occupa di quantizzare blocchi di campioni. Quest'ultima può portare ad una distorsione ben più bassa rispetto a quella scalare ed è ancora più efficiente se i campioni sono statisticamente *dipendenti*.
Se n è la cardinalità del blocco emesso dalla sorgente (la lunghezza del messaggio) si ha che i livelli di quantizzazione sono vettori $\hat{x}_1^n, \hat{x}_2^n, \ldots, \hat{x}_m^n \in \mathbb{R}^n$ con $m = 2^{nR}$.
La quantizzazione scalare quantizza ogni singolo simbolo separatamente in un livello mentre quella vettoriale vede l'intero vettore come un simbolo unico e lo quantizza in un unico livello.
Nella quantizzazione vettoriale si classificano blocchi di dati in un numero discreto di categorie (celle) in modo da ottimizzare qualche criterio (ad esempio la distorsione quadratica media): Le celle sono le "regioni di quantizzazione", ovvero tutti i vettori in ingresso che cadono all'interno di una data cella sono associati alla stessa parola di codice. Il problema è definire le celle e i vettori di quantizzazione ad esse associate per poter effettuare questa sorta di *pattern recognition*. Anche in questo caso si possono trovare dei metodi per definire le regioni di quantizzazione in relazione alla distribuzione di probabilità della sorgente (non uniforme).

Se, ad esempio, avessimo $\mathbf{x} = (x_1, x_2)$ con $x_1, x_2 \in X$ si avrebbe $Q(\mathbf{x}) = (\hat{x}_1, \hat{x}_2) \in \mathbb{R}^2$ in cui i simboli x_1, x_2 verrebbero mappati sugli assi, inducendo una struttura a reticolo. Ad ogni cella viene associato un codice e l'unico grado di libertà è il livello di quantizzazione, lavorando in $\mathbb{R}$.

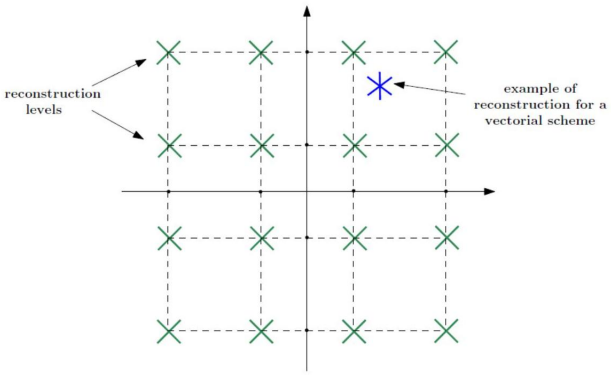


Figura 11: Quantizzazione a quadrati.

Con la quantizzazione vettoriale ci si svincola dalla forma quadrata delle celle: in questo caso, lavorando *direttamente* in $\mathbb{R}^2$, il vettore quantizzato può assumere qualsiasi valore.

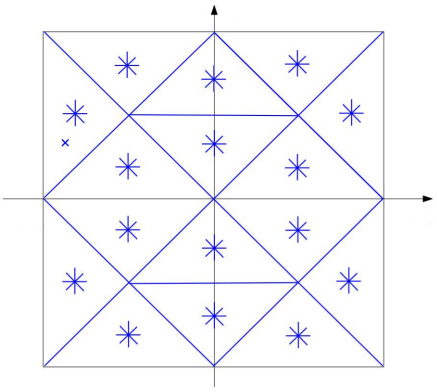


Figura 12: Tassellazione triangolare (esempio di quantizzazione vettoriale).

Nella quantizzazione vettoriale c'è più libertà nella definizione delle regioni di quantizzazione perchè non si è vincolati ad una griglia rigida, come nel caso scalare. Anche nel caso di distribuzione uniforme con la quantizzazione vettoriale si riduce la distorsione. Si può mostrare che utilizzando, ad esempio, regioni esagonali

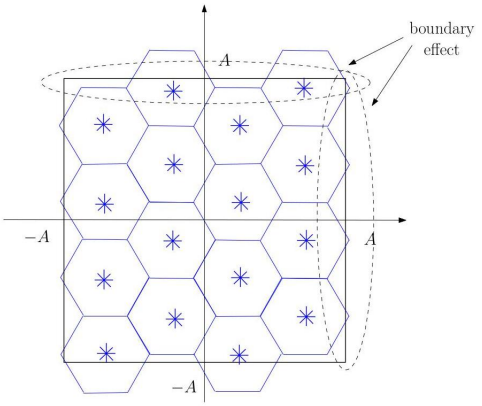


Figura 13: Tassellazione esagonale del dominio uniforme $[-A, A]^2$.

si può ridurre la distorsione. A parità di area il calcolo dei momenti centrali d'inerzia suggerisce di utilizzare figure con più lati possibili, dovendoci avvicinare al minimo teorico dato dalla sfera. Questo induce però dei problemi legati ad effetti di bordo, dati dal fatto che non è possibile ricoprire una regione quadrata con poligoni con più di 4 lati. Questo effetto diventa trascurabile all'aumentare del numero di intervalli di quantizzazione, ovvero eseguendo una *quantizzazione fine*.

Esempio 2 : Sorgenti con memoria.

Nelle sorgenti con memoria, il beneficio è ancora più evidente: siano X, Y due v.a. con pdf congiunta

$$f_{XY}(x,y) = \begin{cases} \frac{1}{ab}, & (x,y) \in \mathcal{A}, \\ 0, & \text{altrimenti.} \end{cases} \tag{3.4.8}$$

dove $\mathcal{A}$ è la regione rettangolare delimitata da due lati a e b come in Figura 14:

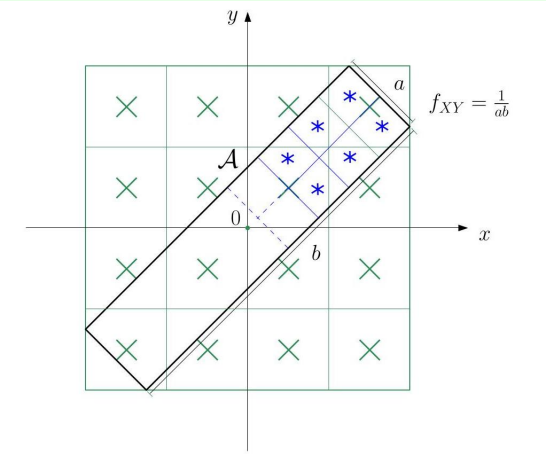


Figura 14: Esempio di due v.a. correlate X e Y. La quantizzazione vettoriale (stelline in blu) risulta necessaria dal momento che ogni struttura scalare (tassellazione in verde) porta ad una distribuzione non soddisfacente dei punti ricostruiti.

Con la quantizzazione scalare si ottiene un risultato inefficiente dal momento che ci sono molte aree che non sono mai interessate dai valori, visto che questa tiene conto solo delle marginali $f_X(x), f_Y(y)$, non rendendo la quantizzazione efficiente dal punto di vista dello spreco di risorse. La correlazione tra X e Y rende necessario il dover ricorrere ad una quantizzazione vettoriale.

4 Capacità di Canale

Si considera ora il caso di una trasmissione in cui sia presente un canale non ideale, ovvero in cui *non è certo che quello che viene inviato corrisponda a quello che viene ricevuto.* Il problema che ci poniamo è quindi l'**affidabilità** della trasmissione: la codifica di canale affronta il problema di trasmettere in maniera affidabile su un canale non affidabile. La codifica di canale è strettamente collegata alla capacità di canale, ovvero alla capacità del canale di "far passare" l'informazione, che, come vedremo, ci porterà al Secondo Teorema di Shannon. Codifica di canale e di sorgente sono duali ma separate:

- Con la codifica di sorgente si toglie la ridondanza per comprimere l'informazione.
- Con la codifica di canale si vuole aumentare l'affidabilità della trasmissione aggiungendo ridondanza al segnale trasmesso: si cerca un compromesso tra affidabilità ed efficienza realizzando un processo di codifica a controllo d'errore per ridurre gli effetti del rumore presente sul canale.

4.1 Canale

Il concetto di canale è piuttosto ampio: rappresenta ciò che succede ad un'informazione tra sorgente e destinatario. Può includere, a seconda delle necessità, solo il mezzo fisico (ad esempio il solo doppino telefonico o un nastro magnetico) o, ad esempio, anche tutto ciò che è compreso tra un microfono e un altoparlante. Alcuni esempi di canale possono essere

- Linea telefonica /ADSL (rumore termico, distorsioni, cross-talk, ...).
- Comunicazioni wireless (attenuazione atmosfera, rumore termico, interferenze, ...).
- Hard-disk[9] (errori di lettura/scrittura, materiali imperfetti, ...).

In ogni caso, in un canale, esiste *sempre* la presenza di rumore che compromette la trasmissione.

4.1.1 Canale discreto senza memoria

Definizione 20. *Canale discreto senza memoria tempo-invariante:* Si definisce canale discreto senza memoria tempo-invariante un canale in cui:

- I simboli in uscita dalla sorgente (quindi in ingresso al canale) e in uscita dal canale, appartengano ad un alfabeto finito.
- In uscita dal canale si ha una sequenza che è casuale ma ha una distribuzione statistica che dipende dalla sequenza in ingresso.

[9]Non necessariamente la comunicazione deve avvenire tra due oggetti distinti, si può pensare che la sorgente e il destinatario siano lo stesso oggetto ma a tempi differenti, come nel caso della memorizzazione dell'informazione.

- L'uscita in un determinato istante dipende solo dal simbolo in ingresso al canale in quell'istante e non dai simboli precedentemente trasmessi.
- Le proprietà del canale non variano nel tempo.

Un canale discreto senza memoria tempo-invariante è univocamente definito dall'alfabeto di ingresso A, l'alfabeto di uscita B e le *forward probability* $p(b|a)$.

Figura 15: Schema ad alto livello di un canale di comunicazione.

In ricezione, dalla sequenza in uscita, si prova a ricostruire la sequenza di ingresso. Siccome però due o più input possono portare alla stessa parola in uscita dal canale, la ricostruzione della sequenza originaria può essere **affetta da errori**: la comunicazione ha successo quando il ricevitore ed il trasmettitore *concordano* su ciò che è stato trasmesso. Un canale può essere rappresentato con un grafo o, equivalentemente, con una matrice di canale $\mathcal{P}$

Esempio 1 : Rappresentazioni di canale

Detto $X = \{x_1, x_2\}$ l'alfabeto d'ingresso e $Y = \{y_1, y_2, y_3\}$ l'alfabeto di uscita si hanno le seguenti rappresentazioni equivalenti

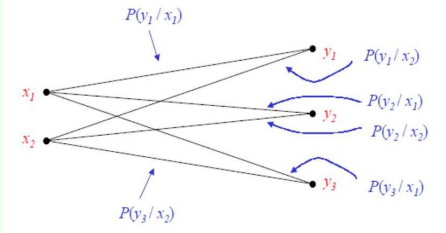


Figura 16: Rappresentazione a grafo di un canale.

$$\mathcal{P} = \begin{bmatrix} p(y_1|x_1) & p(y_2|x_1) & p(y_3|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & p(y_3|x_2) \end{bmatrix}$$

in cui si ha che le $\mathcal{P}_{ij} = p(b_j|a_i)$ e che le righe sommano ad 1[10]: $\sum_{j=1}^{s} p(b_j|a_i) = 1, \forall i = 1, \ldots, r$. Se in ingresso al canale invece dei singoli elementi dell'alfabeto abbiamo sequenze di n simboli dobbiamo fare riferimento all'estensione n-esima della sorgente A^n e del destinatario B^n. Il canale in questo caso è completamente definito dai due alfabeti A^n, B^n e dalla matrice di canale Π data dal prodotto di Kronecker n volte della matrice $\mathcal{P}$ del canale originario:

$$\Pi = \underbrace{\mathcal{P} \otimes \mathcal{P} \otimes \cdots \otimes \mathcal{P}}_{n} \tag{4.1.1}$$

[10]Questo serve a garantire che per ogni input a verrà effettivamente generato un output.

Chiariamo la **notazione**:

- $p(a)$ rappresenta la probabilità **a priori** dei simboli in ingresso, ovvero la probabilità che la sorgente emetta il simbolo a.
- $p(b)$ rappresenta la probabilità che a destinazione venga ricevuto il simbolo b.
- $p(a, b)$ rappresenta la probabilità congiunta che sia stato trasmesso il simbolo a e che venga ricevuto il simbolo b.
- $p(b|a)$ rappresenta la probabilità condizionata che sia stato ricevuto il simbolo b dato che è stato trasmesso il simbolo a, è quella che abbiamo definito **forward probability**.
- $p(a|b)$ rappresenta la probabilità condizionata che sia stato trasmesso il simbolo a dato che è stato ricevuto il simbolo b, prende il nome di **backward probability**, ed è la probabilità **a posteriori** dei simboli in ingresso.

Si possono quindi definire due entropie: l'entropia *a priori*

$$H(A) = \sum_{a \in A} p(a) \log \log \frac{1}{p(a)} \tag{4.1.2}$$

e l'entropia *a posteriori*

$$H(A|b) = \sum_{a \in A} p(a, b) \log \frac{1}{p(a|b)} \tag{4.1.3}$$

che, come abbiamo detto, rappresentano il numero medio di cifre binarie necessarie per rappresentare A, rispettivamente considerando solo la sorgente e considerando di poter osservare la particolare uscita b del canale.
Mediando su tutti i possibili simboli ricevuti

$$H(A|B) = \sum_{b \in B} H(A|b) = \sum_{b \in A} \sum_{a \in A} p(a, b) \log \frac{1}{p(a|b)} \tag{4.1.4}$$

si ha l'informazione media sulla sorgente conoscendo i simboli ricevuti, ovvero l'incertezza rimasta su A dopo aver conosciuto B. Questa quantità prende il nome di **equivocazione di canale** e rappresenta l'informazione aggiuntiva che serve in ricezione dopo aver osservato B, quindi **ciò che si è perso a causa del canale**.
Se ci mettiamo dalla parte del destinatario al tempo 0 non abbiamo visto arrivare alcun simbolo dal canale e la nostra "incertezza" sulla sorgente A è pertanto l'entropia $H(A)$. Dopo l'osservazione però, la nostra incertezza si riduce a $H(A|B)$. L'informazione che ha viaggiato sul canale è dunque

$$H(A) - H(A|B) = I(A; B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \frac{p(a, b)}{p(a)p(b)} \tag{4.1.5}$$

Infatti, ricordando che $I(A; B)$ è l'informazione contenuta sia in A che in B, si ha che questa rappresenta ciò che effettivamente ha attraversato il canale. Di conseguenza la *massima quantità di informazione che può viaggiare attraverso un canale è data dal massimo valore che può assumere l'informazione mutua.*

4.1.2 Capacità di Canale

Abbiamo detto quindi che la massima quantità di informazione che può viaggiare attraverso un canale è quindi data dal massimo valore che può assumere l'informazione mutua. Questa però dipende sia dalla probabilità a priori $p(a)$ che dalla forward probability $p(b|a)$ quindi, rispettivamente, dalla sorgente e dalla matrice di canale. Infatti quanta informazione arriva a destinazione dipende sia dal tipo di canale considerato che anche dall'uso che viene fatto del canale. Se vogliamo massimizzare l'informazione trasportata dal canale dobbiamo anche agire sulla sorgente.
Al fine di caratterizzare un canale discreto senza memoria **indipendentemente dalla sorgente in ingresso**, si definisce capacità del canale il valore massimo dell'informazione mutua rispetto a tutte le possibili distribuzioni delle probabilità dei simboli di ingresso:

Definizione 21. *Capacità di Canale:* Misurata in $[bit/simbolo]$ la capacità di canale $\mathcal{C}$ è definita come

$$\mathcal{C} \coloneqq \max_{p(a)} I(A;B) \tag{4.1.6}$$

Se si massimizza su $p(a)$, $\mathcal{C}$ dipende solo dal canale stesso e *rappresenta il massimo flusso informativo che può essere sopportato dal canale.* Se, inoltre, la sorgente genera simboli con una frequenza f ($[simboli/s]$) la capacità di canale per unità di tempo è data da

$$\mathcal{C}_t \coloneqq f\mathcal{C}, \qquad [bit/s] \tag{4.1.7}$$

La capacità di canale per unità di tempo $\mathcal{C}_t$ rappresenta la *massima velocità di trasferimento dell'informazione permessa dal canale.* Si hanno alcune proprietà per la capacitȍdi canale:

- $\mathcal{C} \geq 0$
- $\mathcal{C} \leq \log|A|$ (dato che $I(A;B) \leq H(A) \leq \log|A|$)
- $\mathcal{C} \leq \log|B|$ (dato che $I(A;B) \leq H(B) \leq \log|B|$)
- $I(A;B)$ è una funzione continua di $p(a)$
- $I(A;B)$ è una funzione concava di $p(a)$, per cui si ha coincidenza tra massimi locali e massimi globali.

Per determinare la capacità si deve quindi effettuare una massimizzazione. Si possono quindi usare tecniche di ottimizzazione, anche se in generale non è semplice. Ci sono però dei casi particolari in cui la capacità può essere calcolata senza troppi sforzi, vediamone alcuni.

4.1.3 Canale senza rumore

Un canale si definisce **senza rumore** (noiseless) se è caratterizzato da una matrice di canale $\mathcal{P}$ con 1 solo elemento non nullo in ogni colonna. In questo tipo di canale quindi il destinatario sa sempre che simbolo è stato inviato mentre il mittente non è in grado di sapere cosa il destinatario abbia ricevuto. Ricordando che $\mathcal{P}$ è data da

$$\mathcal{P} = \begin{bmatrix} p(b_1|a_1) & p(b_2|a_1) & \dots & p(b_s|a_1) \\ p(b_1|a_2) & p(b_2|a_2) & \dots & p(b_s|a_2) \\ \vdots & \vdots & & \vdots \\ p(b_1|a_r) & p(b_2|a_r) & \dots & p(b_s|a_r) \end{bmatrix} \tag{4.1.8}$$

se un solo elemento per colonna j è diverso da 0 quel simbolo b_j può essere ottenuto solo con un possibile a_i. Se il canale è noiseless allora $H(A|B) = 0$, ovvero l'equivocazione di canale è nulla, e l'osservazione dell'uscita ci restituisce esattamente l'ingresso.

Esempio 2 : Canale senza rumore

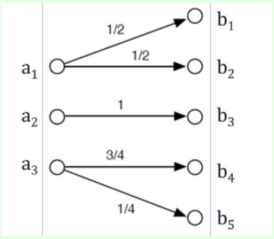


Figura 17: Grafo di un canale senza rumore.

$$\mathcal{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 3/4 & 1/4 \end{bmatrix}$$

Infatti, conoscendo b si determina univocamente a e la probabilità a posteriori è data da

$$p(a|b) = \begin{cases} 1, & \text{se è stato trasmesso } a \\ 0, & \text{altrimenti.} \end{cases}$$

da cui

$$H(A|B) = \sum_{a\in A}\sum_{b\in B} p(a,b)\log\frac{1}{p(a|b)} = 0$$

Si ha quindi che

$$I(A;B) = H(A) - H(A|B) = H(A) = \sum_{a\in A} p(a)\log\frac{1}{p(a)}$$

e che, ricordando che l'entropia di una sorgente è massima quando i simboli sono tutti equiprobabili[11] vale

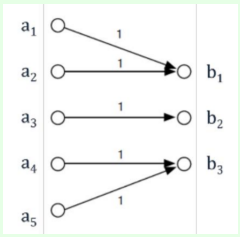
$$\mathcal{C} = \max_{p(a)} I(A;B) = \max_{p(a)} \sum_{a\in A} p(a)\log\frac{1}{p(a)} = \log|A| \tag{4.1.9}$$

4.1.4 Canale deterministico

Un canale si definisce **deterministico** se è caratterizzato da una matrice di canale $\mathcal{P}$ con un solo elemento unitario in ogni riga. *Il mittente sa sempre che simbolo viene ricevuto mentre il destinatario non sa ciò che è stato inviato.*

[11] Cioè $p(a) = 1/|A| = 1/r$.

Esempio 3 : Canale deterministico



$$\mathcal{P} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Figura 18: Grafo di un canale deterministico.

Se un solo elemento per riga i è uguale a 1 allora quel valore a_i può essere ottenuto solo con un possibile b_j. Nel canale deterministico infatti conoscere l'ingresso vuol dire conoscere l'uscita, il che implica $H(B|A) = 0$. Poichè quindi l'incertezza che rimane su B conoscendo A è nulla si ha

$$I(A;B) = I(B;A) = H(B) - H(B|A) = H(B) = \sum_{b \in B} p(b) \log \log \frac{1}{p(b)}$$

da cui, similmente al caso precedente

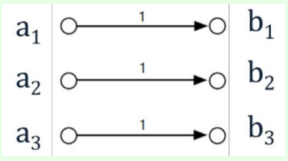
$$\mathcal{C} = \max_{p(a)} I(A;B) = \max_{p(a)} \sum_{b \in B} p(b) \log \frac{1}{p(b)} = \log |B| \tag{4.1.10}$$

In questo caso, infine, l'equivocazione di canale è data da $H(A|B) = H(A) - H(B)$.

4.1.5 Canale completamente ceterministico

Un canale si dice **completamente deterministico** (corrispondenza uno ad uno) se è sia noiseless che deterministico. In questo caso quindi sia il mittente che il destinatario sanno cosa è stato ricevuto/inviato. Alfabeto di ingresso e di uscita hanno quindi la stessa dimensione, dovendo essere il canale biettivo.

Esempio 4 : Canale completamente deterministico



$$\mathcal{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figura 19: Grafo di un canale completamente deterministico.

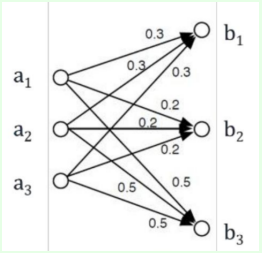
Si ha quindi $I(A;B) = H(A) = H(B)$ da cui

$$\mathcal{C} = \log r = \log s \tag{4.1.11}$$

4.1.6 Canale inutile

Un canale di definisce **inutile** se le uscite sono indipendenti dagli ingressi: $p(b|a) = p(b)$. Il destinatario non può derivare alcuna informazione dalla comunicazione, che è stata appunto inutile. Le colonne sono composte da elementi uguali, per cui tutte le righe sono identiche. L'uscita non dipende quindi dall'ingresso.

Esempio 5 : Canale inutile



$$\mathcal{P} = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.3 & 0.2 & 0.5 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$$

Figura 20: Grafo di un canale inutile.

Si ha quindi che

$$H(B|A) = \sum_{b\in B}\sum_{a\in A} p(a,b)\log\frac{1}{p(b|a)} = \sum_{b\in B}\sum_{a\in A} p(b|a)p(a)\log\frac{1}{p(b)} =$$
$$= \underbrace{\sum_{a\in A} p(a)}_{=1}\sum_{b\in B}\underbrace{p(b|a)}_{=p(b)}\log\frac{1}{p(b)} = \sum_{b\in B} p(b)\log\frac{1}{p(b)} = H(B)$$

da cui

$$I(A;B) = I(B;A) = H(B) - H(B|A) = H(B) - H(B) = 0$$

e quindi

$$\mathcal{C} = \max_{p(a)} I(A;B) = 0 \tag{4.1.12}$$

4.1.7 Canale simmetrico e Canale simmetrico binario

Un canale di definisce **simmetrico** se la sua matrice di canale è caratterizzata da righe e colonne che sono permutazioni degli stessi numeri. Ad esempio:

$$\mathcal{P} = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.2 & 0.5 & 0.3 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}$$

In questo caso si ha che $H(B|A)$ è indipendente dalla sorgente ma dipende unicamente dalla matrice $\mathcal{P}$. Dal momento che ogni riga contiene gli stessi valori, indicando con k una costante, si ha che vale

$$\sum_{j=1}^{s} p(b_j|a_i)\log\frac{1}{p(b_j|a_i)} = k \tag{4.1.13}$$

rendendo questa quantità invariante rispetto alla riga (valendo $\forall i = 1, \dots, r$). Si ha quindi che

$$H(B|A) = \sum_{b\in B}\sum_{a\in A} p(a,b)\log\frac{1}{p(b|a)} = \sum_{b\in B}\sum_{a\in A} p(b|a)p(a)\log\frac{1}{p(b|a)} =$$
$$= \sum_{a\in A} p(a)\underbrace{\sum_{b\in B} p(b|a)\log\frac{1}{p(b|a)}}_{=k} = k\sum_{a\in A} p(a) = k$$

da cui

$$I(A;B) = I(B;A) = H(B) - H(B|A) = H(B) - k$$

ovvero che l'informazione mutua dipende da sia da $H(B)$ che dalla matrice di canale e l'unico modo per avere la massima $I(A;B)$ è massimizzare $H(B)$, il che vuol dire avere simboli equiprobabili. *In un canale simmetrico però si hanno simboli in uscita equiprobabili se i simboli in ingresso sono equiprobabili.* Infatti se $p(a) = 1/r$ vale

$$p(b) = \sum_{a\in A} p(a,b) = \sum_{a\in A} p(b|a)p(a) = \frac{1}{r}\sum_{a\in A} p(b|a) = \frac{col}{r} = \frac{1}{s} \tag{4.1.14}$$

dove *col* rappresenta la somma degli elementi di una colonna della matrice (tutte le colonne sommano a *col* per definizione).

Prendiamo ad esempio un canale debolmente simmetrico (un canale si dice debolmente simmetrico quando ogni riga è una permutazione di ogni altra riga e le colonne sommano allo stesso valore *col*) di questo tipo

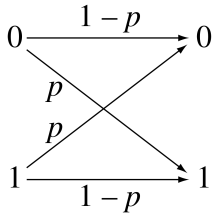
$$\mathcal{P} = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

Se imponiamo $\forall a \in A$, $p(a) = 1/2$ si ha $p(b) = \frac{1}{2}\times\frac{2}{3} = \frac{1}{3}, \forall b \in B$.

In generale si ha quindi che la capacità di un canale simmetrico è

$$\mathcal{C} = \log(s) - k \tag{4.1.15}$$

Il **canale simmetrico binario** (BSC) è un particolare canale simmetrico composto da due simboli in ingresso e da due simboli in uscita $A = B = \{0,1\}$ in cui si ha che p rappresenta la probabilità che ci sia un errore di trasmissione. Si ha quindi che $Pr\{B = 1|A = 0\} = p = Pr\{B = 0|A = 1\}$ e $Pr\{B = 1|A = 1\} = 1 - p = Pr\{B = 0|A = 0\}$:



$$\mathcal{P} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Figura 21: Grafo di un canale simmetrico binario.

Verifichiamo che una probabilità uniforme in ingresso implica una probabilità uniforme in uscita:

$$\begin{aligned} Pr\{B=0\} &= Pr\{B=0|A=0\}Pr\{A=0\} + Pr\{B=0|A=1\}Pr\{A=1\} = \\ &= (1-p)Pr\{A=0\} + pPr\{A=1\} \\ Pr\{B=1\} &= Pr\{B=1|A=0\}Pr\{A=0\} + Pr\{B=1|A=1\}Pr\{A=1\} = \\ &= pPr\{A=0\} + (1-p)Pr\{A=1\} \end{aligned}$$

$$\text{Quindi } Pr\{A=0\} = Pr\{A=1\} = \frac{1}{2} \implies Pr\{B=0\} = Pr\{B=1\} = \frac{1}{2}$$

Il BSC è il modello più semplice di canale con errori, tuttavia riesce a rappresentare bene la complessità del problema.

$$\begin{aligned} I(A;B) = H(B) - H(B|A) = H(B) - \sum_{a\in A} p(a)H(B|a) = \\ \stackrel{\rho}{=} H(B) - \sum_{a\in A} p(a)H(p) = H(B) - H(p) \leq 1 - H(p) \end{aligned}$$

Dove l'ultima disuguaglianza vale perchè $H(B) \leq 1$ e $H(p)$ è l'entropia di una sorgente binaria con probabilità p:

$$H(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

mentre l'uguaglianza ρ vale perchè, chiamando

- $p(b_0) \coloneqq Pr\{B=0\}$
- $p(b_0|a_0) \coloneqq Pr\{B=0|A=0\} = 1-p$
- $p(a_1) \coloneqq Pr\{A=1\}$
- ...

si ha

$$\begin{aligned} &H(B|A) = \sum_{a\in A}\sum_{b\in B} p(b|a)p(a)\log\frac{1}{p(b|a)} = p(b_0|a_0)p(a_0)\log\frac{1}{p(b_0|a_0)} + \\ &+ p(b_0|a_1)p(a_1)\log\frac{1}{p(b_0|a_1)} + p(b_1|a_0)p(a_0)\log\frac{1}{p(b_1|a_0)} + p(b_1|a_1)p(a_1)\log\frac{1}{p(b_1|a_1)} = \\ &= p(a_0)\Big[p(b_0|a_0)\log\frac{1}{p(b_0|a_0)} + p(b_1|a_0)\log\frac{1}{p(b_1|a_0)}\Big] + p(a_1)\Big[p(b_0|a_1)\log\frac{1}{p(b_0|a_1)} + p(b_1|a_1)\log\frac{1}{p(b_1|a_1)}\Big] = \\ &= p(a_0)\Big[(1-p)\log\frac{1}{1-p} + p\log\frac{1}{p}\Big] + p(a_1)\Big[p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}\Big] = \\ &\Big[p(a_0) + p(a_1)\Big]\Big[(1-p)\log\frac{1}{1-p} + p\log\frac{1}{p}\Big] = H(p) \end{aligned}$$

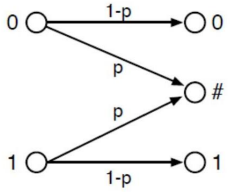
Ricordando (vedi Fig. 3) che l'entropia di una sorgente binaria è massima quando la probabilità è uniforme si ha

$$\mathcal{C} = \max_{p(a)} I(A;B) = 1 - H(p) \tag{4.1.16}$$

L'entropia di B è massima quando $p(b)$ è uniforme, ma ciò vuol dire che la $p(a)$ deve essere uniforme. La capacità di canale di un canale binario simmetrico è quindi massima quando la probabilità a priori $p(a)$ è uniforme.

4.1.8 Binary Erasure Channel

Nel **binary erasure channel** il mittente può inviare due diversi simboli al destinatario. Un simbolo può essere ricevuto correttamente oppure può essere **perso** (in questo caso il destinatario riceve uno speciale simbolo #). Il canale è caratterizzato dalla probabilità p che il simbolo venga perso:



$$\mathcal{P} = \overbrace{\begin{bmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{bmatrix}}^{\begin{matrix} 0 & \# & 1 \end{matrix}}$$

Figura 22: Grafo di un BEC.

Se è stato trasmesso 0 non si può ricevere 1 e viceversa. Questo modello assume quindi che gli 0 e 1 ricevuti vengano ricevuti sempre correttamente e, viceversa, se si è ricevuto # si ha un'ambiguità sul simbolo trasmesso e questa ricezione viene scartata (da quì il nome di BEC). Chiamando $Pr\{A=0\} \coloneqq w, Pr\{A=1\} = 1-w$ e avendo che

- $Pr\{B=0|A=0\} = Pr\{B=1|A=1\} = 1-p$
- $Pr\{B=\#|A=0\} = Pr\{B=\#|A=1\} = p$
- $Pr\{B=1|A=0\} = Pr\{B=0|A=1\} = 0$

calcoliamo $H(B)$ e $H(B|A)$. Si ha

$$\begin{aligned} H(B) &= (1-p)w\log\frac{1}{(1-p)w} + (1-p)(1-w)\log\frac{1}{(1-p)(1-w)} + p\log\frac{1}{p} = \\ &= (1-p)\Big[w\log\frac{1}{(1-p)w} + (1-w)\log\frac{1}{(1-p)(1-w)}\Big] + p\log\frac{1}{p} = \\ &= (1-p)\Big[w\log\frac{1}{1-p} + w\log\frac{1}{w} + (1-w)\log\frac{1}{1-p} + (1-w)\log\frac{1}{1-w}\Big] + p\log\frac{1}{p} = \\ &= (1-p)\Big[\log\frac{1}{1-p} + H(w)\Big] + p\log\frac{1}{p} = H(p) + (1-p)H(w) \end{aligned}$$

.

$$\begin{aligned} H(B|A) &= \sum_{a\in A}\sum_{b\in B} p(b|a)p(a)\log\frac{1}{p(b|a)} = \\ &= w\Big[(1-p)\log\frac{1}{1-p} + p\log\frac{1}{p} + 0\Big] + (1-w)\Big[0 + p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}\Big] = \\ &= wH(p) + (1-w)H(p) = H(p) \end{aligned}$$

Quindi si ha

$$I(A;B) = H(B) - H(B|A) = H(p) + (1-p)H(w) - H(p) = (1-p)H(w)$$

da cui

$$\mathcal{C} = \max_{p(a)} I(A;B) = \max_{w}(1-p)H(w) = 1-p \tag{4.1.17}$$

4.2 Canali in cascata

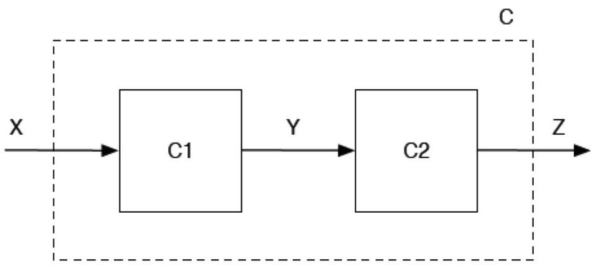


Figura 23: Due canali in cascata C_1 e C_2.

Due canali C_1 e C_2 possono essere posti in *cascata*, ovvero uno di seguito all'altro. La matrice di canale che si ottiene ponendo in cascata due canali C_1, C_2 è il prodotto delle matrici di canale dei due canali, ovvero

$$\mathcal{P}_c = \mathcal{P}_1 \mathcal{P}_2 \tag{4.2.1}$$

Si può verificare che per i canali in cascata l'informazione mutua, man mano che si attraversano i canali, non può aumentare: ad ogni passaggio si può **solo perdere** informazione e mai guadagnarne:

$$\begin{cases} H(X|Z) \geq H(X|Y) \implies I(X;Z) \leq I(X;Y) \\ H(X|Z) \geq H(Y|Z) \implies I(X;Z) \leq I(Y;Z) \end{cases}$$

Esempio 6: BSC in cascata

Siano C_1 e C_2 due BSC in cascata con matrici di canale

$$\mathcal{P}_1 = \begin{bmatrix} 1-p_1 & p_1 \\ p_1 & 1-p_1 \end{bmatrix}, \quad \mathcal{P}_2 = \begin{bmatrix} 1-p_2 & p_2 \\ p_2 & 1-p_2 \end{bmatrix}$$

Allora si ha che la matrice di canale risultante è data da

$$\mathcal{P}_c = \mathcal{P}_1 \mathcal{P}_2 = \begin{bmatrix} (1-p_1)(1-p_2) + p_1 p_2 & (1-p_1)p_2 + (1-p_2)p_1 \\ (1-p_1)p_2 + (1-p_2)p_1 & (1-p_1)(1-p_2) + p_1 p_2 \end{bmatrix}$$

da cui, se $p_1 = p_2 = p$ e $(1-p_1) = (1-p_2) = \bar{p}$ si ha

$$\mathcal{P}_c = \begin{bmatrix} p^2 + \bar{p}^2 & 2p\bar{p} \\ 2p\bar{p} & p^2 + \bar{p}^2 \end{bmatrix}$$

Le due capacità dei singoli tratti valgono quindi

$$\mathcal{C}_1 = \mathcal{C}_2 = 1 - H(p)$$

mentre la capacità del canale in cascata è

$$\mathcal{C}_c = 1 - H(2p\bar{p}) \leq \mathcal{C}_1 = \mathcal{C}_2$$

4.3 Probabilità di Errore e Regola di Decisione

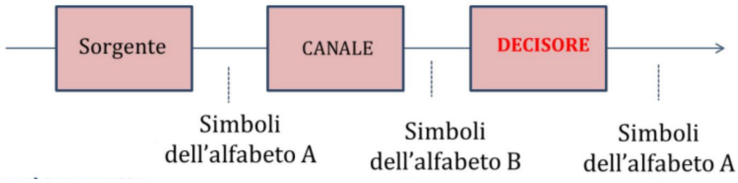
Se il canale introduce rumore, $H(A|B) \neq 0$, ciò che arriva:

- Può essere uguale a quello trasmesso con una certa probabilità.
- Può essere diverso da quello trasmesso con un'altra probabilità.

quindi il sistema è caratterizzato ovviamente da una certa probabilità di errore.

Il secondo teorema di Shannon vuole determinare la quantità di informazione che, con una probabiltà di errore piccola a piacere, può attraversare il canale ed essere ricevuta correttamente.

C'è quindi innanzitutto bisogno di una regola di decisione che consenta di passare dai simboli ricevuti a quelli inviati. Questa è di fondamentale importanza dal momento che, in generale, si ha a che fare con canali "continui" e non discreti, in cui l'alfabeto di uscita può assumere infiniti valori.



Definizione 22. *Regola di decisione:* Dato un alfabeto in ingresso $A = \{a_1, \dots, a_r\}$ e uno di uscita $B = \{b_1, \dots, b_s\}$ una regola di decisione è una funzione $d : B \to A$ tale che

$$\forall b \in B \quad \exists!\ \hat{a} \in A : d(b) = \hat{a} \tag{4.3.1}$$

Esempio 7 : Regola di decisione per BSC

Se si ha un BSC con matrice di canale

$$\mathcal{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

si può decidere in due modi

$$\begin{cases} d(0) = 0 \wedge d(1) = 1 \\ d(0) = 1 \wedge d(1) = 0 \end{cases}$$

La prima regola di decisione non assicura l'assenza di errori ma, mediamente, sbaglierà solo nel 10% dei casi. L'altra regola invece porta a una media del 90% di errore.

In generale si ha che, per un canale con r ingressi ed s uscite, si possono avere r^s regole di decisione. Ad ogni regola di decisione è associata un'incertezza ed il nostro obiettivo è determinare la migliore regola di decisione. La probabilità di errore media P_e è quindi associata ad una certa regola di decisione, in particolare, data una certa regola di decisione d, essa può essere scritta come

$$P_e \coloneqq \sum_{b\in B} \sum_{\substack{a\in A \\ a\neq\hat{a}}} p(a,b) = \sum_{b\in B} \sum_{\substack{a\in A \\ a\neq\hat{a}}} p(b|a)p(a) = 1 - \underbrace{\sum_{b\in B} p(b|\hat{a})p(\hat{a})}_{=P_c} \tag{4.3.2}$$

Dove P_c indica la probabilità di corretta decisione. Come si minimizza questa probabilità? La regola di decisione ottima è data dal criterio di **massima verosimiglianza**:

$$\min_{d(b)} P_e = \min_{d(b)} 1 - P_c = \max_{d(b)} P_c = \max_{d(b)} \sum_{b \in B} p(b|\hat{a})p(\hat{a})$$
$$\text{ovvero } \min_{d(b)} p(b|\hat{a})p(\hat{a}), \quad \forall b \in B$$

Il simbolo $\hat{a}$ viene quindi scelto come quello che massimizza la probabilità di aver ricevuto il simbolo b condizionata alla trasmissione di a:

$$d(b) = \hat{a} \iff p(b|\hat{a})p(\hat{a}) \geq p(b|a)p(a), \quad \forall a \in A, b \in B \tag{4.3.3}$$

che, nel caso di simboli equiprobabili[12], diventa

$$d(b) = \hat{a} \iff p(b|\hat{a}) \geq p(b|a), \quad \forall a \in A, b \in B \tag{4.3.4}$$

4.4 Disuguaglianza di Fano

Supponiamo, osservando una variabile aleatoria B, di voler stimare il valore assunto da un'altra variabile aleatoria A correlata a B. La disuguaglianza di Fano connette la probabilità di errore di indovinare il valore di A con l'entropia condizionata $H(A|B)$. L'idea è che ci aspettiamo di poter stimare A con una bassa probabilità di errore solo se la probabilità condizionata $H(A|B)$ è *piccola*.

Indicando con P_e la probabilità di errore[13] si ha la che

$$H(A|B) \leq H(P_e) + P_e \log(r-1) \tag{4.4.1}$$

Il membro a destra è dato dalla somma di due contributi. Una volta osservato B infatti:

- Non si può determinare se il simbolo ricevuto è affetto da errore o no, e questo è il contributo dato da $H(P_e)$, quest'ultima infatti è la quantità di informazione di una sorgente binaria con simboli $\{errore, non\ errore\}$, quindi la quantità di informazione necessaria all'osservatore per dire se c'è stato un errore.
- Se c'è errore, con probabilità P_e, non si può determinare quale dei restanti $r-1$ simboli si è sbagliato a stimare. Essendo una probabilità condizionata si ha che il valore massimo che può assumere è quello di una sorgente senza condizionamento con $r-1$ simboli tutti equiprobabili, la cui incertezza è $\log(r-1)$ pesata per P_e.

Ricordando che $r = |A|$ la disuguaglianza di Fano ci dice che l'entropia condizionata è massima per $P_e = 1 - \frac{1}{r}$ e vale $H(A|B) = \log r$. Questo avviene quindi quando la probabilità di corretta decisione è uniforme sull'insieme A, ovvero quando $P_c = \frac{1}{r}$.

[12] Se non si conosce la statistica della sorgente si applica il criterio ML considerando simboli equiprobabili. In questo caso si ottiene una regola di decisione sub-ottima.

[13] In generale P_e puè essere dato da qualunque stimatore $g(B) = \hat{A}$: $P_e = Pr\{\hat{A} \neq A\}$.

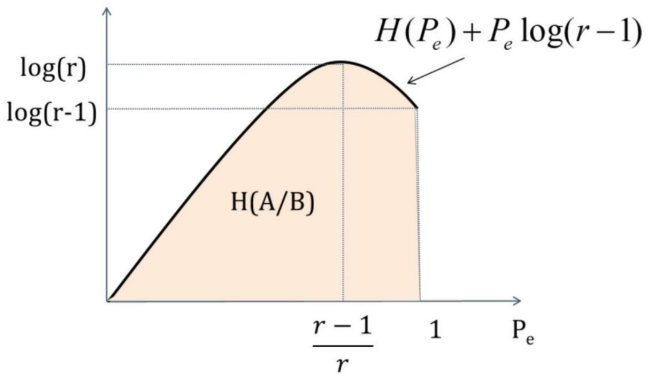


Figura 24: Rappresentazione grafica della disuguaglianza di Fano.

Si ha infatti che la funzione $H(P_e)+\log(r-1)$ è concava nell'intervallo $P_e \in [0,1]$, quindi il massimo si ottiene in

$$\begin{aligned}\frac{\partial}{\partial P_e}\Big\{H(P_e)+P_e\log(r-1)\Big\} &= \frac{\partial}{\partial P_e}\Big\{P_e\log\frac{1}{P_e}+(1-P_e)\log\frac{1}{1-P_e}+P_e\log(r-1)\Big\} = \\ &= \log\frac{1}{P_e}-\log\frac{1}{1-P_e}+\log(r-1) = \\ &= \log\frac{(1-P_e)(r-1)}{P_e}=0 \implies \frac{(1-P_e)(r-1)}{P_e}=1 \\ &\implies P_e = 1-\frac{1}{r}=\frac{r-1}{r}\end{aligned}$$

Vediamo ora la dimostrazione della disuguaglianza di Fano data nella (4.4.1). Questa non è l'unica forma in cui si può presentare ma è la forma che più rapidamente ci permette di mostrare come questa fornisca un limite inferiore alla probabilità d'errore.

Dim: Proviamo che $H(A|B)-H(P_e)-P_e\log(r-1)\leq 0$. Scriviamo i due termini di entropia in funzione di P_e:

$$\begin{aligned}H(P_e)+P_e\log(r-1) &= P_e\log\frac{1}{P_e}+P_c\log\frac{1}{P_c}+P_e\log(r-1) = P_e\log\frac{r-1}{P_e}+P_c\log\frac{1}{P_c} = \\ &= \sum_{b\in B}\sum_{\substack{a\in A\\ a\neq\hat{a}}} p(a,b)\log\frac{r-1}{P_e}+\sum_{b\in B}p(\hat{a},b)\log\frac{1}{P_c}\end{aligned}$$

$$H(A|B)=\sum_{a\in A}\sum_{b\in B}p(a,b)\log\frac{1}{p(a|b)}=\sum_{\substack{a\in A\\ a\neq\hat{a}}}\sum_{b\in B}p(a,b)\log\frac{1}{p(a|b)}+\sum_{b\in B}p(\hat{a},b)\log\frac{1}{p(\hat{a}|b)}$$

da cui, facendo la differenza, si ha

$$
\begin{aligned}
&H(A|B) - H(P_e) + P_e \log(r-1) = \\
&= \sum_{\substack{a\in A \\ a\neq \hat{a}}} \sum_{b\in B} p(a,b) \log \frac{1}{p(a|b)} + \sum_{b\in B} p(\hat{a},b) \log \frac{1}{p(\hat{a}|b)} - \sum_{b\in B} \sum_{\substack{a\in A \\ a\neq \hat{a}}} p(a,b) \log \frac{r-1}{P_e} - \sum_{b\in B} p(\hat{a},b) \log \frac{1}{P_c} \\
&= \sum_{\substack{a\in A \\ a\neq \hat{a}}} \sum_{b\in B} p(a,b) \log \frac{P_e}{p(a|b)(r-1)} + \sum_{b\in B} p(\hat{a},b) \log \frac{P_c}{p(\hat{a}|b)} = \\
&= \log e \left[\sum_{\substack{a\in A \\ a\neq \hat{a}}} \sum_{b\in B} p(a,b) \ln \frac{P_e}{p(a|b)(r-1)} + \sum_{b\in B} p(\hat{a},b) \ln \frac{P_c}{p(\hat{a}|b)} \right] \leq \\
&\leq \log e \left[\sum_{\substack{a\in A \\ a\neq \hat{a}}} \sum_{b\in B} p(a,b) \Big(\frac{P_e}{p(a|b)(r-1)} - 1 \Big) + \sum_{b\in B} p(\hat{a},b) \Big(\frac{P_c}{p(\hat{a}|b)} - 1 \Big) \right] = \\
&= \log e \left[\sum_{\substack{a\in A \\ a\neq \hat{a}}} \sum_{b\in B} \Big(\frac{P_e p(b)}{r-1} - p(a,b) \Big) + \sum_{b\in B} P_c p(b) - p(\hat{a},b) \Big) \right] = \\
&= \log e \left[\sum_{\substack{a\in A \\ a\neq \hat{a}}} \frac{P_e}{r-1} - p(a) + P_c - p(\hat{a}) \right] = \\
&= \log e \left[\frac{P_e(r-1)}{r-1} - \big(1 - p(\hat{a})\big) + P_c - p(\hat{a}) \right] = \\
&= \log e \Big[P_e + P_c - 1 \Big] = 0
\end{aligned}
$$

□

Si ha inoltre che la disuguaglianza vale come uguaglianza quando, $\forall b \in B$

$$
\begin{cases} \frac{P_e}{p(a|b)(r-1)} = 1, \quad \forall a \in A \backslash \{\hat{a}\} \\ \frac{P_c}{p(\hat{a}|b)} = 1 \end{cases} \implies \begin{cases} P_e = p(a|b)(r-1), \quad \forall a \in A \backslash \{\hat{a}\} \\ P_c = p(\hat{a}|b) \end{cases}
$$

ovvero quando, se si trasmette un simbolo a, si ha la stessa probabilità di sbagliare con uno qualsiasi degli altri simboli.

Se prendiamo la disuguaglianza di Fano e la relazione che lega informazione mutua ed equivocazione di canale si ha:

$$
H(A|B) \leq H(P_e) + P_e \log(r-1) \wedge H(A|B) = H(A) - I(A;B) \geq H(A) - \mathcal{C}
$$

da cui

$$
H(A) \leq H(A|B) \leq H(P_e) + P_e \log(r-1) + \mathcal{C} \tag{4.4.2}
$$

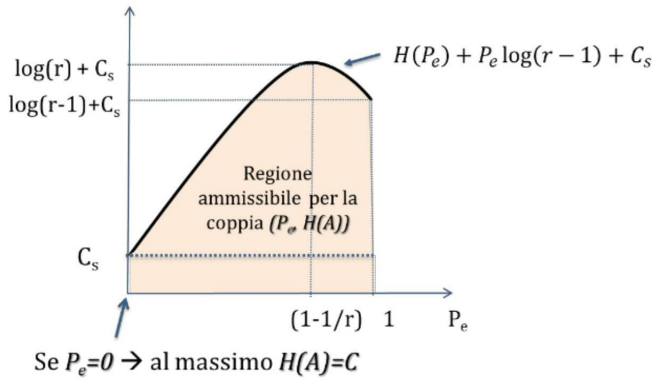


Figura 25: Regione di di ammissibilità per la coppia $\big(P_e, H(A)\big)$.

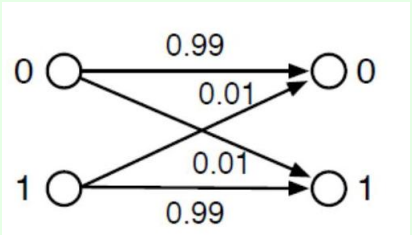
Si ha quindi che *se l'entropia della sorgente (il rate) supera la capacità di canale* ***è impossibile trasmettere con probabilità di errore piccola a piacere***. Se $H(A) - \mathcal{C} = \tau > 0$ si ha

$$H(P_e) + P_e \log(r-1) \geq \tau > 0$$

e quindi un limite inferiore per la P_e. Dal momento che esiste questo limite, dato un certo canale con capacità $\mathcal{C}$, vogliamo capire se c'è un modo per rendere la comunicazione più affidabile. **L'obiettivo è quindi quello di trovare un limite per la quantità di informazione che può essere trasmessa in modo affidabile su un canale rumoroso**.
L'affidabilità di una comunicazione può essere migliorata con la codifica di canale, il cui obiettivo consiste proprio nell'aumentare la resistenza dell'informazione al rumore presente sul canale. La codifica di canale *trasforma* la sequenza di dati in ingresso al canale in una nuova sequenza intrinsecamente più robusta agli effetti del rumore. *L'approccio adottato consiste solitamente nell'introdurre ridondanza*. Sfruttando tale ridondanza il decodificatore può ricostruire il messaggio originale anche in presenza di bit errati.

Esempio 6 : Codice a ripetizione

Consideriamo un canale simmetrico binario con $p = 0.01$ con simboli in ingresso equiprobabili.



Si ha che

$$\mathcal{P} = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}$$

in cui la decisione ottima d è data da

$$\begin{cases} d(0) = 0 \\ d(1) = 1 \end{cases} \implies P_e = \frac{0.01 + 0.01}{2} = 0.01$$

Proviamo a migliorare l'affidabilità con una semplice codifica di canale: il *codice a ripetizione*, inviando il *bit*, ad esempio, 3 volte invece di una sola. In sostanza quindi si considera l'estensione di un BSC in cui i simboli che possono essere inviati sul canale sono solamente due $000, 1111$, mentre possono essere ricevute tutte le combinazioni di 3 bit (a seguito degli errori di trasmissione).

In ricezione la regola di decisione è abbastanza semplice: si sceglie 0 se sono stati ricevuti più 0 che 1, mentre si sceglie 1 nell'altro caso (avendo preso una ripetizione dispari non si ha mai ambiguità). Poiché la probabilità che 1 bit venga trasmesso in maniera non corretta e uguale a $p = 0.01$ per tutti i simboli, si ha dalla distribuzione binomiale

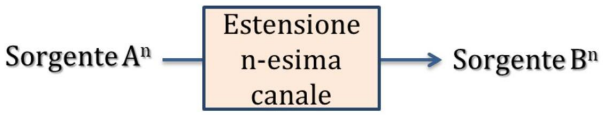
$$P_e = p^2(1-p)\begin{pmatrix} 3 \\ 2 \end{pmatrix} + p^3(1-p)^0\begin{pmatrix} 3 \\ 3 \end{pmatrix} =$$
$$= 3p^2(1-p) + p^3 = 3 \times 0.01^2 \times 0.99 + 0.01^3 \approx 3 \cdot 10^{-4}$$

L'affidabilità è aumentata a costo di aumentare il numero di bit di ben 3 volte. Possiamo quindi dire che la velocità di trasmissione è ridotta di 1/3. Se aumentiamo ancora il numero di bit vedremo sempre lo stesso andamento:

n	P_e	R
1	10^{-2}	1
3	$3 \cdot 10^{-4}$	1/3
5	$3 \cdot 10^{-5}$	1/5
7	$4 \cdot 10^{-7}$	1/7
9	10^{-8}	1/8
$\vdots$	$\vdots$	$\vdots$

Quindi se nel tempo di k simboli si trasmette un solo simbolo informativo la velocità di trasmissione si riduce di k volte.

4.5 Canale esteso



Nell'estensione del canale andiamo a considerare sorgenti estese A^n, B^n. Si ha che vale

$$I(A^n; B^n) \leq n\mathcal{C}$$

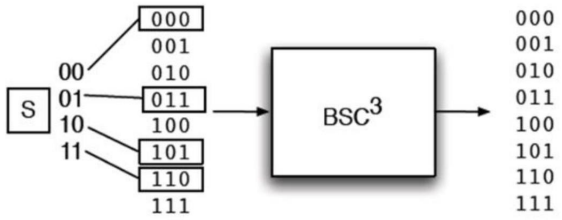
Dim:

$$I(A^n; B^n) = H(B^n) - H(B^n|A^n) = H(B^n) - \sum_{i=1}^{n} H(B_i|B_1, \ldots, B_{i-1}, A^n) =$$
$$\stackrel{\eta}{=} H(B^n) - \sum_{i=1}^{n} H(B_i|A_i) \leq \sum_{i=1}^{n} H(B_i) - \sum_{i=1}^{n} H(B_i|A_i) = \sum_{i=1}^{n} I(A_i; B_i) \leq$$
$$\leq n\mathcal{C} \qquad \square$$

Dove l'uguaglianza η vale perche il canale è discreto e senza memoria, da cui B_i dipdende unicamente da A_i ed è quindi condizionalmente indipendente da tutto il resto. La prima disuguaglianza vale invece dalla (1.2.13) e la seconda dalla definizione di capactità.
Si può estendere la disuguaglianza di Fano all'estensione n-esima del canale come

$$H(A^n|B^n) \leq H(P_e^n) + P_e^n \log(r^n - 1) \tag{4.5.1}$$

dove P_e^n è la probabilità di errore associata a una sequenza di n simboli e non la potenza n-esima della P_e per un simbolo. Abbiamo visto che aumentando il numero di bit usati per codificare un simbolo della sorgente si riduce la P_e ma anche la velocità con cui si trasmettono i dati. Consideriamo allora un altro esempio: dato lo stesso canale BSC visto sopra, cambiamo codifica, invece di inviare un bit per volta, ne inviamo due e codifichiamo il messaggio fatto da due bit in una parola di codice fatta da tre bit:



La decisione ottima d è data da

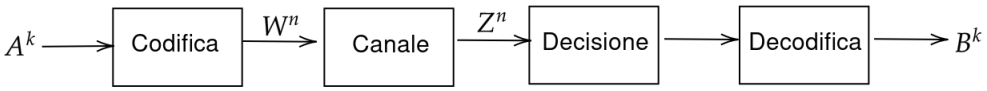
$$\begin{cases} d(000) = d(001) = 00 \\ d(010) = d(011) = 01 \\ d(100) = d(101) = 10 \\ d(110) = d(111) = 11 \end{cases} \implies P_e \approx 2 \cdot 10^{-2}$$

Con un rate $R = \frac{2}{3}$ dal momento che vengono inviati 3 bit per rappresentarne 2.

La codifica di canale associa a blocchi di k bit in ingresso blocchi di n bit, con $n \geq k$ il che comporta che il **coding rate** R_c segua la legge

$$R_c := \frac{k}{n} \tag{4.5.2}$$

dal momento che, se per ogni parola della sorgente lunga k bit se ne inviano n la velocità di trasmissione si ridurrà: se invio 1 *bit* ogni T unità di tempo (ad esempio secondi): senza codifica sono necessari kT unità di tempo per terminare la comunicazione, con la codifica sono necessari nT unità di tempo, con $n \geq k$. Abbiamo quindi una struttura di questo tipo:



In cui, seguendo l'esempio con $k = 2, n = 3$, potremmo avere $A^2 = \{00, 01, 10, 11\}, W^3 = \{000, 011, 101, 110\}$ e $Z^3 = \{000, 001, 010, 011, 100, 101, 110, 100\}$. In questo sistema si ha che

- $H(A^k|B^k) = H(A^k) - I(A^k; B^k)$
- Con i canali in cascata: $I(A^k; B^k) \leq I(W^n; Z^n)$
- Con un canale esteso: $I(W^n; Z^n) \leq n\mathcal{C}$

Questo comporta

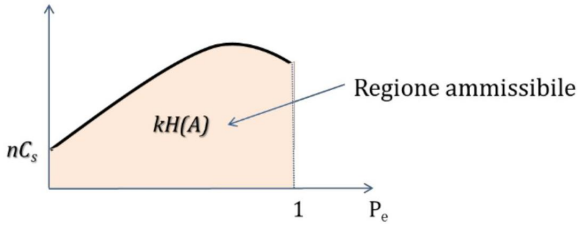
$$\begin{cases} H(A^k|B^k) \geq H(A^k) - n\mathcal{C} = kH(A) - n\mathcal{C} \\ H(A^k|B^k) \leq H(P_e^n) + P_e^n \log(r^k - 1) \end{cases} \implies kH(A) \leq H(P_e^n) + P_e^n \log(r^k - 1) + n\mathcal{C}$$

Nel limite in cui $P_e \to 0$ si ha che

$$kH(A) \leq n\mathcal{C} \implies H(A) \leq \frac{n}{k}\mathcal{C} = \frac{\mathcal{C}}{R_c}$$

quindi deve essere che

$$R = R_c H(A) \leq \mathcal{C} \qquad (4.5.3)$$



ovvero che, per poter lavorare nella regione $P_e \to 0$ si deve:

- **Ridurre il rate di trasmissione**:
  1. *Riducendo il rate* (entropia) *della sorgente* A, quindi comprimendo di più (non è detto che sia possibile).
  2. Ridcuendo R_c, ovvero *aumentando la ridondanza della codifica di canale*: fissata $H(A)$ se la sorgente emette un bit ogni T_s unità di tempo, con modulazione binaria la banda necessaria B sarebbe, se $R_c = 1$
  $$B \approx \frac{1}{T_s}$$
  se invece il canale deve far passare n bit in kT_s unità di tempo la banda necessaria sarebbe
  $$B \approx \frac{n}{kT_s} = \frac{R_c}{T_s}$$
  il che implica che *serve una banda maggiore.*
- **Aumentare la capacità di canale** $\mathcal{C}$ aumentando il *Rapporto Segnale Rumore* sul canale: la P_e si ridurrà dal momento che il canale introduce una minore equivocazione.

4.6 Secondo Teorema di Shannon

Il secondo teorema di Shannon da un'altra possibilità, ci dice che fissato il rate di sorgente e la capacità di canale, si può mantenere costante il rapporto $R_c = \frac{n}{k}$ (ovvero non aumentando la banda) **incrementando la lunghezza del blocco in ingresso al codificatore**. In sostanza dimostra che *è possibile ridurre arbitrariamente la probabilità di errore, con il vincolo che il rate di trasmissione sia inferiore alla capacita di canale.*

Supponiamo di avere una sorgente discreta senza memoria con alfabeto A, entropia $H(A)$ e un canale discreto senza memoria con capacità per simbolo pari a $\mathcal{C}$ $[bit/simbolo]$. Il **Secondo Teorema di Shannon** afferma che

Esiste un sistema di codifica con $R_c = \frac{k}{n}$ *e* $R = R_c H(A) < \mathcal{C}$ *che permette la trasmissione dell'informazione emessa dalla sorgente sul canale con una probabilità di errore arbitrariamente piccola.*

Questo può essere espresso come

$$\forall \epsilon > 0, \exists n_0 \in \mathbb{N} : \forall n > n_0 \text{ si ha } P_e < \epsilon \tag{4.6.1}$$

ed in particolare si può scrivere, detta $E(\cdot)$ una funzione convessa decrescente e positiva in $[0, \mathcal{C}]$

$$P_e < e^{-nE(R_s)} \text{ con } R_s \coloneqq R_c H(A) \tag{4.6.2}$$

Il Secondo Teorema di Shannon *non fornisce dettagli sul sistema di codifica necessario* per ottenere probabilità d'errore arbitrariamente piccola, ma afferma solo che, *se* $R < \mathcal{C}$, esso ***esiste***. Nella pratica, si può verificare che per ridurre la probabilità d'errore si deve incrementare il numero di *bit* in ingresso al codificatore. Per $n \to \infty$ però il tempo necessario per la trasmissione del messaggio codificato tende all'infinito e la complessità di codifica e decodifica cresce.

4.7 Canale Gaussiano

Fino ad ora abbiamo considerato canali discreti. Nella realtà le informazioni vengono inviate in canali analogici e si devono quindi considerare sorgenti analogiche. Studiamo in particolare il canale con rumore bianco gaussiano ($AWGN$)[14], che è caratterizzato da un modello a rumore additivo tra ingresso X ed uscita Y:

$$Y = X + Z \tag{4.7.1}$$

in cui il rumore $Z = \mathcal{N}(0; \sigma_z^2)$ con $X \perp Z$ (X e Z incorrelate). *Data la bianchezza del rumore il canale è quindi senza memoria e stazionario.*
Anche nel caso di segnali analogici poi vale quanto abbiamo visto, e la capacità è definita come il massimo dell'informazione mutua:

$$\begin{aligned}
\mathcal{C} &= \max_X I(X;Y) = \max_X \big[h(Y) - h(Y|X)\big] = \\
&= \max_X \big[h(Y) - h(X+Z|X)\big] = \max_X \big[h(Y) - h(Z|X)\big] = \\
&= \max_X \big[h(Y) - h(Z)\big] = \max_X \big[h(Y) - \frac{1}{2}\log(2\pi\sigma_z^2 e)\big]
\end{aligned}$$

che si ottiene quando Y è gaussiana

$$\mathcal{C} = \frac{1}{2}\log(2\pi\sigma_y^2 e) - \frac{1}{2}\log(2\pi\sigma_z^2 e) = \frac{1}{2}\log\frac{\sigma_y^2}{\sigma_z^2}$$

ed essendo $X \perp Z$ si ha $\sigma_y^2 = \sigma_x^2 + \sigma_z^2$ da cui

$$\mathcal{C} = \frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right) \tag{4.7.2}$$

Stiamo però considerando un canale con banda infinita: nella pratica tutti i canali hanno banda limitata $[-B, B]$ e, se non vogliamo avere aliasing, il segnale deve essere campionato[15] con una frequenza almeno pari a $2B$ campioni/s. Ogni campione di segnale ha potenza $P_X = \sigma_x^2$ mentre la

[14] Questo canale è un buon modello per molte comunicazioni, come quella telefonica o quella satellitare.
[15] Per il Teorema del Campionamento di Nyquist-Shannon.

potenza del rumore, assumento che la densità spettrale di potenza del segnale sia $\sigma_x^2 = \frac{N_0}{2}\frac{[W]}{[Hz]}$, è data da $P_Z = \sigma_z^2 = N_0 B$ da cui

$$\mathcal{C} = \frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{N_0 B}\right) \quad \frac{bit}{\text{trasmissione}}$$

e se teniamo in considerazione la frequenza di campionamento $f = 2B$ possiamo usare il canale $2B$ volte al secondo, quindi

$$\mathcal{C} = f\frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{N_0 B}\right) = B\log\left(1 + \frac{\sigma_x^2}{N_0 B}\right) = B\log\left(1 + SNR\right) \ \frac{bit}{s} \tag{4.7.3}$$

che è famosa formula di Shannon della capacità di un canale *AWGN*.
Dalla formula si vede che la capacità dipende dalla banda e dal *SNR*, quindi dalla potenza del segnale in ingresso P_x. Aumentando la potenza di trasmissione aumenta il *SNR* e ovviamente aumenta la capacità, ma la crescita è logaritmica. Aumentando la banda B aumenta il termine a moltiplicare ma anche la quantità di rumore che entra nel segnale, quindi si ha un asintoto

$$\begin{aligned}\lim_{B\to\infty}\mathcal{C} &= \lim_{B\to\infty} B\log\left(1 + \frac{\sigma_x^2}{BN_0}\right) = \\ &= \lim_{B\to\infty}(\log e)B\ln\left(1 + \frac{\sigma_x^2}{BN_0}\right) \approx \lim_{B\to\infty}(\log e)B\frac{\sigma_x^2}{BN_0} = (\log e)\frac{\sigma_x^2}{N_0}\end{aligned}$$

dove si è usato l'approssimazione al primo ordine $\ln(1+x) \approx x$ con $x \ll 1$. È chiaro quindi che, anche aumentando la banda B non si può scendere sotto il limite di $\approx 1.44\sigma_x^2/N_0$.

4.8 Curva di Shannon

Introduciamo adesso la curva di Shannon, che mostra l'esistenza di un *trade-off* tra potenza del segnale in ingresso P_x e banda B in ogni sistema di comunicazione. Dal momento che deve essere $R < \mathcal{C}$ si deve avere

$$R < B\log\left(1 + \frac{P_x}{N_0 B}\right) \tag{4.8.1}$$

dividendo entrambi i membri per B si ottiene l'*efficienza spettrale* $r \coloneqq R/B$

$$r < \log\left(1 + \frac{P_x}{N_0 B}\right) \tag{4.8.2}$$

ovvero il numero di *bits* per secondo che possono essere trasmessi in un'unità di banda (in 1 Hz). Osservando che $P_x = E_b R$, dove E_b è l'energia per bit trasmesso, otteniamo

$$r < \log\left(1 + r\frac{E_b}{N_0}\right) \tag{4.8.3}$$

Questa relazione definisce le regioni di ammissibilità di efficienza spettrale al variare del rapporto E_b/N_0. Il luogo dei punti in cui $r = \log(1 + rE_b/N_0)$ prende il nome di **curva di Shannon** (vedi Figura 26). Quest'ultima divide il piano $(E_b/N_0, r)$ in due parti: la regione sotto la curva (in blu) rappresenta l'insieme dei punti per cui è possibile ottenere una comunicazione affidabile (quella

sopra in cui non è possibile). Per studiare il comportamento del ENR (*Energy-Noise Ratio*) al variare di r eleviamo alla potenza di 2 entrambi i membri della 4.8.3, da cui

$$\frac{2^r - 1}{r} < \frac{E_b}{N_0} \tag{4.8.4}$$

Possiamo quindi studiare il comportamento asintotico:

$$\begin{cases} r \to \infty \implies \frac{E_b}{N_0} \to \infty \\ r \to 0 \implies \frac{E_b}{N_0} \to \ln 2 \end{cases}$$

ovvero che la curva di Shannon ha un asintoto verticale in $E_b/N_0 = \ln 2 \approx 1.6\ dB$ al di sotto del quale non è possibile avere una trasmissione affidabile, per ogni valore di r.

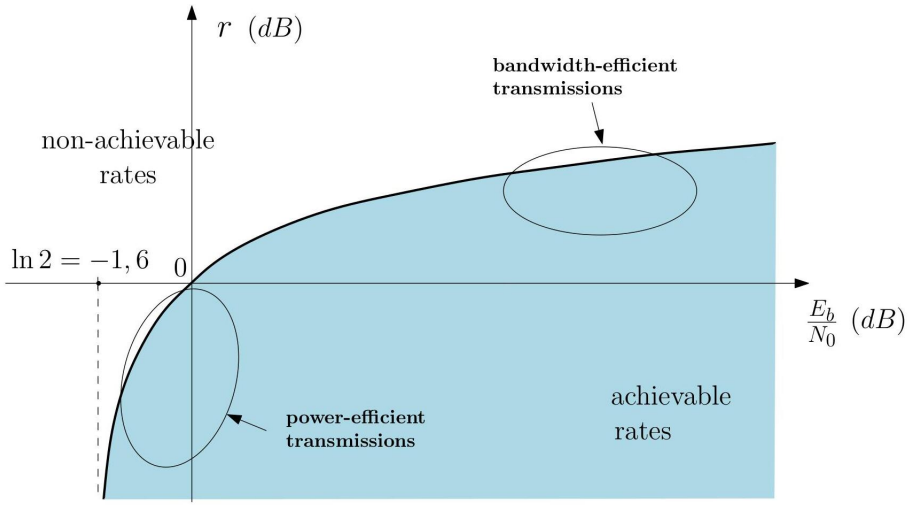


Figura 26: Curva di capacità di Shannon.

La curva che si ottiene è teorica, ma tali risultati sono molto utili nella progettazione di un sistema reale, in quanto definiscono il ***limite superiore*** *delle prestazioni ottenibili da un sistema* modellabile con un canale corrotto da rumore *AWGN*. Nella regione dei punti operativi (raggiungibili) ovviamente più il punto di lavoro reale che si riesce ad ottenere si avvicina alla curva limite ($R = C$) più il sistema è efficiente. Per le comunicazioni il cui principale limite è la potenza di trasmissione ($r \ll 1$) si hanno trasmissioni efficienti in potenza, mentre per i sistemi dove è la banda ad essere il limite principale ($r \gg 1$) si hanno trasmissioni efficienti in banda.

Appendices

A Richiami (TODO)

A.1 Richiami di probabilità

A.2 Richiami sui processi stocastici

A.3 Notazione

Alcuni apici che useremo nel corso di questi appunti, che verranno inseriti ogni qual volta si debba esplicitare il motivo del passaggio matematico in cui sono stati invocati:

$$\alpha : \sum_x p(x) = 1$$
$$\beta : \ln x \leq x - 1 \text{ con } \ln x = x - 1 \iff x = 1$$
$$\gamma : p(a, b) = p(a|b)p(b)$$
$$\delta : \sum_y p(x, y) = p(x)$$