UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI INGEGNERIA

Corso di Laurea in

INGEGNERIA DELL'INFORMAZIONE

Proprietà di equipartizione asintotica e sue applicazioni nella teoria dell'informazione

Relatore

Prof. Giancarlo Calvagno

Candidato

Matteo Pagin

Anno Accademico 2011/2012

Prefazione

In questo elaborato vedremo uno dei risultati fondamentali nella teoria dell'informazione: la proprietà di equipartizione asintotica. Essa deriva direttamente dalla legge debole dei grandi numeri ed esprime come gli esiti di una sequenza di variabili aleatorie indipendenti e identicamente distribuite possano essere divisi sostanzialmente in due insiemi, di cui il più piccolo contiene la quasi totalità della probabilità, e per questo è detto tipico. Ciò significa che siamo abbastanza certi che sotto opportune condizioni, un esito qualsiasi appartenga a questo insieme tipico. Grazie a questo risultato vedremo come realizzare un sistema di compressione dati senza perdita di informazione.

Estenderemo poi l'argomento a due sequenze di variabili aleatorie descritte da una densità di probabilità congiunta, il cui caso tipico è quello di una sequenza casuale in ingresso a un canale di comunicazione e la sua rispettiva uscita. Definiremo anche per queste coppie di sequenze un insieme tipico e grazie alle sue proprietà è possibile dimostrare il teorema sulla codifica di canale, che permette di identificare esattamente quanta informazione è possibile trasferire attraverso un particolare canale affetto da rumore senza avere perdita.

Indice

Capitolo 1

Definizioni Preliminari

Prima di entrare nel vivo dell'argomento riporto alcune definizioni e teoremi di base che serviranno per dimostrare i risultati che vedremo più avanti, così da introdurre anche la notazione che sarà utilizzata nel resto del documento.

Come convenzione indicherò le variabili aleatorie (v.a.) con lettere maiuscole, e a volte le distribuzioni saranno abbreviate scrivendo $p\left(x\right)$ al posto di $p_{\mathbf{X}}(x)$, quando questo non sarà soggetto a fraintendimenti. Diamo ora la definizione di entropia di una v.a., che rappresenta una misura dell'incertezza di quest'ultima.

Definizione 1.1. *(Entropia di una v.a.). Data una v.a. X discreta su un alfabeto $\mathcal{X}$ con densità di probabilità $p_{\boldsymbol{X}}(x)$ definiamo la sua entropia $H(X)$ come:*

$$H(X) = -\sum_{x\in\mathcal{X}} p_{\boldsymbol{X}}(x) log(p_{\boldsymbol{X}}(x)) = E\left(\log\frac{1}{p_{\boldsymbol{X}}(x)}\right) \tag{1.1}$$

dove useremo 2 come base del logaritmo, a meno che non venga espressamente indicato diversamente: in questo modo l'unità di misura dell'entropia è il bit.

Pensiamo al numero di domande binarie (risposta si o no) che serve per descrivere una certa v.a.: si può dimostrare che l'aspettazione del minimo numero di domande è compresa tra $H(X)$ e $H(X)+1$. Questo fa vedere come l'entropia sia strettamente legata al contenuto di informazione di una v.a.

Indichiamo inoltre con

$$H\left(X|Y\right), H\left(X,Y\right), I\left(X,Y\right)$$

rispettivamente l'entropia condizionata, l'entropia congiunta e l'informazione mutua di una coppia di v.a. Per una descrizione esaustiva di queste grandezze e le loro proprietà rimando ad altri testi, ad esempio [1].

Definizione 1.2. *Si che dice che le v.a. X,Y,Z formano una catena di Markov se la probabilità di Z dipende solo da Y e data questa è indipendente da X ovvero*

$$p(x, z|y) = p(x|y)\, p(z|y) \tag{1.2}$$

e si indica con $X \to Y \to Z$

Riporto, senza dimostrazione, due teoremi che serviranno poi per la dimostrazione delle proprietà che vedremo.

Teorema 1.1. *(Disuguaglianza sull'elaborazione dati). Se X,Y,Z formano una catena di Markov, $X \to Y \to Z$ allora*

$$I(X, Y) \geq I(X, Z) \tag{1.3}$$

Teorema 1.2. *(Disuguaglianza di Fano). Per una catena di Markov $X \to Y \to \tilde{X}$, definita la probabilità di errore $P_e = P\left(X \neq \tilde{X}\right)$ si ha:*

$$1 + P_e \log |\mathcal{X}| \geq H\left(X|\tilde{X}\right) \geq H(X|Y) \tag{1.4}$$

dove $|\mathcal{X}|$ è la cardinalità dell'alfabeto di X.

Riporto la definizione di convergenza in probabiltà di una sequenza di variabili aleatorie.

Definizione 1.3. *Data una sequenza di v.a. $\{X_n\}$=X_1,X_2,... si dice che questa converge in probabilità alla v.a. X se $\forall\, \varepsilon > 0$*

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \tag{1.5}$$

e si indica con $X_n \xrightarrow{\mathcal{P}} X$.

Capitolo 2

Proprietà di Equipartizione Asintotica

Nella teoria della probabilità si è vista la legge debole dei grandi numeri, nella teoria dell'informazione esiste un suo analogo rappresentato dalla proprietà di equipartizione asintotica (asymptotic equipartition property, AEP da qui in avanti).

2.1 AEP e l'insieme tipico

Teorema 2.1. *Proprietà di equipartizione asintotica. Si consideri una sequenza di v.a.* $(X_1, X_2, ..., X_n)$ *con* X_i *i.i.d. con densità di probabilità* $p_{\boldsymbol{X}}(x)$ $(X_i \sim \mathrm{p}_{\boldsymbol{X}}(x))$. *Allora:*

$$-\frac{1}{n}\log p\left(X_1, X_2, ..., X_n\right) \xrightarrow{\mathcal{P}} H(X) \tag{2.1}$$

dove $p\left(X_1, X_2, .., X_n\right)$ *rappresenta la probabilità di un certo esito* $(X_1, X_2, ..., X_n)$ *e* $H(X)$ *l'entropia di una qualsiasi v.a.* X_i *appartenente alla sequenza.*

Dimostrazione: Sfruttiamo l'indipendenza delle X_i e le proprietà dei logaritmi:

$$-\frac{1}{n}\log p\left(X_1, X_2, ..., X_n\right) = -\frac{1}{n}\sum_{i=1}^{n}\log p\left(X_i\right). \tag{2.2}$$

Dal fatto che funzioni di v.a. indipendenti sono ancora v.a. indipendenti si osserva che la parte a destra dell'equazione è la media campionaria di $-\log p(X)$ con $X \sim \mathrm{p}_{\mathbf{X}}(x)$, quindi per la legge debole dei grandi numeri:

$$-\frac{1}{n}\sum_{i=1}^{n}\log p\left(X_i\right) \xrightarrow{\mathcal{P}} -E\left(\log p(X)\right). \tag{2.3}$$

Infine è immediato vedere che la seconda parte dell'equazione corrisponde alla definizione di entropia per una v.a.

$$-E\left(\log p(x)\right) = -\sum_{x\in X} p(x)\log p(x) = H(X). \tag{2.4}$$

□

Vediamo ora che questo semplice risultato, derivante direttamente dalla legge dei grandi numeri, ci permette di definire un insieme di possibili esiti della sequenza $(X_1, X_2, ..., X_n)$ con interessanti proprietà. Ad esempio questo insieme, detto insieme tipico, conterrà la quasi totalità della probabilità dell'insieme di tutti i possibili esiti della sequenza $\{X_n\}$. Ovvero prendendo una sequenza casuale tra tutte quelle possibili, con alta probabilità questa sarà all'interno dell'insieme tipico.

Definizione 2.1. *Insieme tipico. Si definisce l'insieme tipico $A_\varepsilon^{(n)}$ rispetto alla densità $p_{\boldsymbol{X}}(x)$ come l'insieme di tutti gli esiti $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$ tali che*

$$2^{-n(H(x)+\varepsilon)} \leq p(x_1, x_2, ..., x_n) \leq 2^{-n(H(x)-\varepsilon)} \tag{2.5}$$

Il prossimo teorema mostra le proprietà di cui gode questo insieme.

Teorema 2.2.

1. $\forall (x_1, x_2, ..., x_n) \in A_\varepsilon^{(n)}$ *vale* $H(x) - \varepsilon \leq -\frac{1}{n}\log p(x_1, x_2, ..., x_n) \leq H(X) + \varepsilon$
2. $P\left(A_\varepsilon^{(n)}\right) \geq 1 - \varepsilon$ *Per n abbastanza grande*
3. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$
4. $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$ *Per n abbastanza grande*

Dimostrazione:
1. Segue direttamente da (2.5) passando ai logaritmi:

$$-n(H(X)+\varepsilon) \leq \log p(x_1, x_2, ..., x_n) \leq -n(H(x)-\varepsilon) \tag{2.6}$$

moltiplicando per $-\frac{1}{n}$ si ha la proprietà 1.

Per la seconda proprietà usiamo (2.1) e, passando all'evento complementare nella definizione di convergenza in probabilità, si ha

$$\lim_{n\to\infty} P\left(\left|-\frac{1}{n}\log p(X_1, X_2, ..., X_n) - H(X)\right| \leq \varepsilon\right) = 1. \tag{2.7}$$

Dalla definizione di limite $\forall\varepsilon > 0$, $\forall\delta > 0$, $\exists N$ tale che $\forall n \geq N$

$$1 - \delta \leq P\left(\left|-\frac{1}{n}\log p(X_1, X_2, ..., X_n) - H(X)\right| \leq \varepsilon\right) \leq 1 + \delta. \tag{2.8}$$

Dove la seconda disequazione in (2.8) è sempre vera poichè $P(A) \leq 1$ per qualsiasi evento A, inoltre dalla prima proprietà segue che $\left|-\frac{1}{n}\log p(X_1, X_2, ..., X_n) - H(X)\right|$ siano tutti e soli gli elementi di $A_\varepsilon^{(n)}$, essendo questa diretta conseguenza di (2.5). Per quanto detto e scegliendo $\varepsilon = \delta$ otteniamo la proprietà 2 dalla prima delle due disequazioni.

Per dimostrare 3 utilizziamo la proprietà di normalizzazione di $p(x)$

$$\begin{aligned}
1 &= \sum_{x\in\mathcal{X}} p(x) \\
&= \sum_{x\in A_\varepsilon^{(n)}} p(x) + \sum_{x\notin A_\varepsilon^{(n)}} p(x) \\
&\geq \sum_{x\in A_\varepsilon^{(n)}} p(x) \\
&\geq \sum_{x\in A_\varepsilon^{(n)}} 2^{-n(H(x)+\varepsilon)} = 2^{-n(H(x)+\varepsilon)} \left|A_\varepsilon^{(n)}\right|
\end{aligned}$$

da cui si ottiene la proprietà 3.

Per l'ultima parte del teorema procediamo in maniera simile al punto precendente, sfruttando la seconda proprietà del teorema

$$\begin{aligned}
1 - \varepsilon &\leq \sum_{x\in A_\varepsilon^{(n)}} p(x) \\
&\leq \sum_{x\in A_\varepsilon^{(n)}} 2^{-n(H(x)-\varepsilon)} = 2^{-n(H(x)-\varepsilon)} \left|A_\varepsilon^{(n)}\right|
\end{aligned}$$

e in modo diretto ricaviamo la quarta e ultima proprietà. □

Dalla seconda proprietà del precedente teorema ricaviamo quindi che gran parte della probabilità di $\mathcal{X}^n$ è concentrata in $A_\varepsilon^{(n)}$ per valori sufficentemente grandi di n. Questo significa che una sequenza di $\mathcal{X}^n$ avrà un'elevata probabilità di appartenere all'insieme tipico. Questo avviene nonostante la dimensione dell'insieme tipico è praticamente trascurabile rispetto al numero totale di sequenze in $\mathcal{X}^n$. Confrontiamo il numero di elementi per entrambi gli insiemi:

$$|A_\varepsilon^{(n)}| \cong 2^{n(H(x))}$$

$$|\mathcal{X}^n| = 2^{n\log(|\mathcal{X}|)}$$

e si osserva che

$$\begin{aligned}\frac{|A_\varepsilon^{(n)}|}{|\mathcal{X}^n|} &\cong \frac{2^{n(H(x))}}{2^{n\log(|\mathcal{X}|)}} \\ &= 2^{n(H(X)-\log(|\mathcal{X}|))}\end{aligned}$$

dove l'ultimo termine tende a 0 per $n \to \infty$ e $H(X) < \log(|\mathcal{X}|)$. Il Teorema 2.2 ci dice che una sequenza di $\{X_n\}$ appartiene con probabilità elevata all'insieme tipico. Quindi nonostante l'insieme tipico abbia una dimensione trascurabile rispetto al numero totale di sequenze questo contiene la maggior parte della probabilità.

Per capire come questo sia possibile facciamo un esempio:

Esempio 1. *(Tratto da esercizio 3.7, [1], pag 66)*

Immaginiamo una sorgente di simboli binari che può essere modellizzata con una variabili di Bernoulli $X \sim b(p)$ con $p = 0.003$. Prendiamo sequenze di 200 bit, $(X_1, ..., X_{200})$ dove $X_i \sim b(p)$ e calcoliamo la probabilità P che una sequenza casuale abbia al più 3 bit pari a 1. Chiamiamo $Y = \sum_{i=1}^{200} X_i$

$$\begin{aligned}P &= P(Y \leq 3) = P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) \\ &= \binom{200}{0}0.997^{200} + \binom{200}{1}0.997^{199}0.003 + \binom{200}{2}0.997^{198}0.003^2 + \\ &+ \binom{200}{3}0.997^{197}0.003^3 \\ &= 0.5483 + 0.33 + 0.0988 + 0.0196 = 0.9967.\end{aligned}$$

Equivalentemente, la probabilità che una sequenza non sia nell'insieme di sequenze con almeno 3 bit pari a 1 è $1 - P = 0.0033$.

Sebbene nell'esempio non abbiamo trovato l'insieme tipico, abbiamo mostrato come sia possibile che un sottoinsieme abbia la maggior parte della probabilità dell'insieme totale seppure con un esiguo numero di elementi. Infatti se indichiamo con N il numero di sequenze con al più 3 bit uguali a 1 si ha

$$N = \binom{200}{0} + \binom{200}{1} + \binom{200}{2} + \binom{200}{3} = 1333501 \tag{2.9}$$

non paragonabile al numero totale di sequenze $2^{200} = 1.61 \times 10^{60}$.

2.2 Compressione dati

In [1] troviamo un'interessante applicazione per effettuare compressione dati utilizzando la AEP: vediamo qui di seguito il procedimento.

Vogliamo codificare le possibili sequenze $(X_1, X_2, ..., X_n)$ di n v.a. i.i.d. con distribuzione $p(x)$. Se diciamo $\mathcal{X}$ l'alfabeto delle v.a, la codifica delle sequenze corrisponde a trovare una codifica per $\mathcal{X}^n$.

Dividiamo $\mathcal{X}^n$ in $A_\varepsilon^{(n)}$ e $A_\varepsilon^{(n)^c}$. È possibile ordinare tutte le sequenze appartenenti agli insiemi secondo un qualche ordine, ad esempio ordine lessicografico, dopo l'ordinamento, possiamo semplicemente descrivere una particolare sequenza, dando il suo indice per l'ordinamento scelto e dicendo a quale dei due insiemi appartiene. Con questo metodo per descrivere una sequenza in $A_\varepsilon^{(n)}$, essendo $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$ sono sufficienti $n(H(X)+\varepsilon)+1$ bit, ovvero $\lceil n(H(X)+\varepsilon) \rceil$. A questi aggiungiamo all'inizio un bit per dire se si tratta di una sequenza appartenente a $A_\varepsilon^{(n)}$ o $A_\varepsilon^{(n)^c}$, per esempio rispettivamente per i due casi 0 e 1.

Per descrivere invece gli elementi di $A_\varepsilon^{(n)^c}$ possiamo dire che ci bastano al più $n \log |\mathcal{X}|+1$ bit, essendo questo il numero di bit sufficente a descrivere l'intero $\mathcal{X}^n$. Anche qui è presente un bit aggiuntivo per indicare l'insieme, 1 nel nostro caso.

Sia quindi $l(x^n)$ la lunghezza della parola di codice usata per codificare la sequenza $x^n = (x_1, x_2, ..., x_n)$. Abbiamo che

$$l(x^n) \leq n(H(X)+\varepsilon) + 2 \qquad \text{se } x^n \in A_\varepsilon^{(n)} \tag{2.10}$$

$$l(x^n) \leq n \log |\mathcal{X}| + 2 \qquad \text{se } x^n \in A_\varepsilon^{(n)^c}. \tag{2.11}$$

Calcoliamo l'aspettazione di $l(x^n)$ per vedere se questo metodo di codifica è efficiente.

$$\begin{aligned} E(l(x^n)) &= \sum_{x\in\mathcal{X}} l(x^n)p(x^n) \\ &\leq \sum_{x\in A_\varepsilon^{(n)}} (n(H(X)+\varepsilon)+2)p(x^n) + \sum_{x\in A_\varepsilon^{(n)^c}} (n\log|\mathcal{X}|+2)p(x^n) \\ &= P(A_\varepsilon^{(n)})(n(H(X)+\varepsilon)+2) + (1-P(A_\varepsilon^{(n)}))n\log|\mathcal{X}|+2 \end{aligned}$$

Per la proprietà 2 del Teorema 2.2 e per valori abbastanza grandi abbiamo che

$$E(l(x^n)) \leq n(H(X)+\varepsilon) + \varepsilon n\log|\mathcal{X}|+2 \tag{2.12}$$

Ponendo poi $\varepsilon' = \varepsilon + \log|\mathcal{X}| + \frac{2}{n}$ possiamo opportunamente scegliere ε e n per minimizzare a piacere il valore di ε'.

Abbiamo dimostrato così che esiste un codice per rappresentare $\mathcal{X}^n$, sicuramente decodificabile poichè la funzione che mappa le sequenze è biunivoca, e con lunghezza media delle parole di codice circa $nH(X)$. Formalizziamo ora il precedente risultato:

Teorema 2.3. *Siano X^n sequenza di v.a. i.i.d. $\sim \mathrm{p}_{\boldsymbol{X}}(x)$ e $\varepsilon > 0$. Allora esiste un codice che mappa le sequenze $(x_1, x_2, ..., x_n)$ in stringhe di bit, tale che la mappa sia invertibile e vale*

$$E\left(\frac{1}{n}l(X^n)\right) \leq H(X)+\varepsilon \tag{2.13}$$

per n abbastanza grande.

Dimostrazione: Fatta sopra. □

È interessante notare inoltre che per la codifica di $A_\varepsilon^{(n)^c}$ non abbiamo preso particolari precauzioni e abbiamo usato tanti bit quanti ne bastano per descrivere l'intero insieme $\mathcal{X}^n$, nonostante questo abbiamo ottenuto un codice che in media necessita di meno bit, infatti ricordiamo che $H(X) \leq \log|\mathcal{X}|$. Chiaramente il caso in cui $H(X) = \log|\mathcal{X}|$ non è di nostro interesse qui, essendo in tal caso tutte le sequenze equiprobabili.

Capitolo 3

Il teorema della codifica di Canale

In questo capitolo dimostreremo il teorema della codifica di canale, introdotto per la prima volta da Claude E. Shannon nel 1948, nel suo famoso articolo che aprì la strada all'analisi matematica dei sistemi di comunicazione [2]. Per la dimostrazione useremo sequenze congiuntamente tipiche: vedremo come l'insieme di queste sequenze presenti proprietà molto simili a quelle dell'insieme tipico, visto nel capitolo precedente, e non stupirà infatti vedere come anche queste derivino dalla legge debole dei grandi numeri.

3.1 Definizioni

L'obiettivo di questa tesi è di analizzare le applicazioni della proprietà di equipartizione asintotica nella teoria dell'informazione, non intendo quindi trattare per esteso le definizioni di capacità e di canale (comprese le differenti tipologie di canale). Tuttavia in questa sezione darò queste definizioni per coerenza di notazione visto che saranno poi usate nella dimostrazione del teorema della codifica di canale.

Il sistema preso in esame consiste di un insieme di parole $W \in \{1, 2, ..., M\}$, in pratica W rappresenta un indice, ad esempio $W = i$ per indicare la parola i-esima delle possibili parole generate da una certa sorgente. Codificheremo W in un segnale (una sequenza di simboli, x_i nel nostro caso) $X^n(W)$. Il segnale viene quindi inviato lungo il canale e raggiunge il ricevitore secondo una distribuzione $Y^n \sim p(y^n|x^n)$. Il ricevitore con una funzione di decodifica $g(Y^n)$ cercherà di ottenere una stima $\tilde{W}$ di W.

Definiamo formalmente un canale:

Definizione 3.1. *Un canale discreto è descritto come la terna* $(\mathcal{X}, p(y|x), \mathcal{Y})$*, dove* $\mathcal{X}, \mathcal{Y}$ *sono due insiemi finiti rappresentanti rispettivamente i simboli in ingresso e in uscita al canale.* $p(y|x)$ *è definita* $\forall x \in \mathcal{X}$ *ed è la p.m.f. della variabile* $Y|X$*.*

Possiamo estendere la precedente definizione nel caso pratico in cui il canale sia usato più volte consecutivamente, trasmettendo una sequenza di simboli.

Definizione 3.2. *Si dice estensione n-esima senza memoria del canale* $(\mathcal{X}, p(y|x), \mathcal{Y})$ *il canale* $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$ *per cui vale*

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k) \qquad k = 1, 2, ..., n. \tag{3.1}$$

(3.1) assicura che il canale sia senza memoria e che ogni simbolo in uscita dipenda esclusivamente dal simbolo in ingresso corrispondente e non dall'intera sequenza di input (simboli trasmessi prima o che verranno trasmessi) o dall'output prodotto dai simboli precedenti. Questo comporta inoltre:

$$p(y^k|x^k) = \prod_{i=1}^{k} p(y_i|x_i). \tag{3.2}$$

In questo capitolo, se non espressamente indicato, ci riferiamo sempre a un canale senza memoria, per cui vale la proprietà vista sopra. Ricordiamo la definizione di capacità di un canale

Definizione 3.3. *(Capacità di canale). Si definisce capacità di un canale discreto e senza memoria*

$$C = \max_{p(x)} I(X, Y) \tag{3.3}$$

Definiamo il "rate "di un codice (M, n) con funzione di codifica $X^n : \{1, 2, ...M\} \to \mathcal{X}^n$ e decodifica $g : \mathcal{Y}^n \to \{1, 2, ..., M\}$ la quantità

$$R = \frac{\log M}{n} \tag{3.4}$$

3.2 Proprietà di equipartizione asintotica per sequenze congiunte

Sempre avendo in mente la dimostrazione del teorema della codifica di canale sviluppiamo in questa sezione un metodo per decodificare le sequenze al ricevitore, $\mathcal{Y}^n$, in maniera da

minimizzare la probabilità di errore $P(W \neq \tilde{W})$. Useremo in seguito il procedimento per dimostrare, appunto, il teorema. Chiaramente esistono molti modi per fare questo, noi ci concentreremo su un metodo, basato sulla proprietà di equipartizione asintotica, che risulterà facilmente analizzabile nella dimostrazione del teorema della codifica di canale.

Cominciamo col definire l'insieme tipico, analogo a quanto visto nel precedente capitolo, ora considerando due diverse sequenze di v.a. Dalle sue proprietà vedremo che quando due sequenze sono dipendenti esse saranno contenute, molto probabilmente, nell'insieme tipico, questo è il caso in cui una sequenza di ingresso è la causa fisica della sequenza in uscita. Viceversa due sequenze indipendenti avranno bassa probabilità di appartenere all'insieme tipico.

Definizione 3.4. *Insieme delle sequenze congiuntamente tipiche, $A_\varepsilon^{(n)}$. È l'insieme delle sequenze $\{(x^n, y^n)\}$ rispettivamente alla distribuzione $p(x, y)$ così definito:*

$$\begin{aligned} A_\varepsilon^{(n)} = \{&(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ &\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \varepsilon; && (3.5) \\ &\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \varepsilon; && (3.6) \\ &\left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \varepsilon \Bigg\} && (3.7) \end{aligned}$$

con

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \tag{3.8}$$

dove (3.8) è conseguenza del fatto che i simboli di input sono indipendenti tra di loro. Le coppie di sequenze di $A_\varepsilon^{(n)}$ sono dunque quelle per cui la loro entropia empirica, quella che possiamo misurare, è vicina a quella reale della v.a. che le genera. Facciamo notare che questa era proprietà anche delle sequenze appartenenti all'insieme tipico per una sola variabile.

Osserviamo che (3.5) è la proprietà 1 del teorema 2.2, vorremo allora che anche questo nuovo insieme tipico definito per due sequenze avesse proprietà simili a $A_\varepsilon^{(n)}$ della definizione 2.1.

Facciamo vedere che ne possiede alcune che sono l'analogo del Teorema 2.2.

Prendiamo la sequenza X^n per questa vale (2.1), allora prendiamo nella definizione del limite per la convergenza in probabilità (1.5) $\delta = \frac{\varepsilon}{3}$ e $n = n_1$.

$$P\left(\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3} \qquad \forall n > n_1 \tag{3.9}$$

Va notato che l'evento di cui indichiamo la probabilità è il complementare di (3.5).

Grazie all'assenza di memoria del canale possiamo scrivere (2.1) anche per Y e poi per X e Y congiuntamente. Prendendo sempre $\frac{\varepsilon}{3}$ e rispettivamente $n = n_2$ $n = n_3$ nei due casi abbiamo:

$$P\left(\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3} \qquad \forall n > n_2 \tag{3.10}$$

$$P\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3} \qquad \forall n > n_3 \tag{3.11}$$

Possiamo ora scegliere $n > \max\{n_1, n_2, n_3\}$ in modo che valgano contemporaneamente le (3.9), (3.10), (3.11). Da queste relazioni e dalla proprietà che la probabilità di unione di eventi è minore uguale alla somma delle probabilità degli eventi, abbiamo che la probabilità dell'unione dei tre eventi definiti nelle tre relazioni è minore di ε. Osservato che i tre eventi corrispondono ai complementari delle tre proprietà con cui abbiamo definito $A_\varepsilon^{(n)}$, è facile vedere che la loro unione è ${A_\varepsilon^{(n)}}^c$. Allora $\forall n > \max\{n_1, n_2, n_3\}$, $\forall \varepsilon > 0$

$$P\left({A_\varepsilon^{(n)}}^c\right) < \varepsilon \tag{3.12}$$

$$\lim_{n\to\infty} P\left({A_\varepsilon^{(n)}}^c\right) = 0 \tag{3.13}$$

Infine otteniamo la prima importante proprietà di $A_\varepsilon^{(n)}$

$$\lim_{n\to\infty} P\left(A_\varepsilon^{(n)}\right) = 1 \tag{3.14}$$

Possiamo sfruttare la proprietà appena trovata per trovare una relazione sul numero di elementi di $A_\varepsilon^{(n)}$:

$$\begin{aligned} 1 &= \sum_{(x^n, y^n)} p(x^n, y^n) \\ &\geq \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n, y^n) \end{aligned}$$

da (3.7) si ricava facilmente che $p(x^n, y^n) \geq 2^{-n(H(X,Y)+\varepsilon)}$ e quindi

$$
\begin{aligned}
1 &\geq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} 2^{-n(H(X,Y)+\varepsilon)} \\
&\geq \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X,Y)+\varepsilon)}
\end{aligned}
$$

Riassiumiamo le due proprietà in

Teorema 3.1. *Siano (X^n, Y^n) sequenze di lunghezza n, distribuite i.i.d. con $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Allora*

1. $P\left(A_\varepsilon^{(n)}\right) \to 1 \qquad$ *per* $n \to \infty$
2. $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\varepsilon)}$
3. $\left|A_\varepsilon^{(n)}\right| \geq (1-\varepsilon)\, 2^{n(H(X,Y)-\varepsilon)}$

Dimostrazione: Le prime due proprietà sono state ricavate nei passaggi precedenti.

La terza proprietà è facilmente dimostrabile. Partendo dalla definizione del limite nella prima proprietà del teorema otteniamo:

$$
P\left(A_\varepsilon^{(n)}\right) \geq 1-\varepsilon \tag{3.15}
$$

Da questo, analogamente a come fatto nel precedente capitolo per il teorema 2.2, e ricavando da (3.7) che $p(x^n, y^n) \leq 2^{-n(H(X,Y)-\varepsilon)}$

$$
\begin{aligned}
1-\varepsilon &\leq P\left(A_\varepsilon^{(n)}\right) \\
&= \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n, y^n) \\
&\leq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} 2^{-n(H(X,Y)-\varepsilon)} = \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X,Y)-\varepsilon)}
\end{aligned}
$$

Da cui segue l'ultima proprietà. □

Il teorema precedente indica le stesse proprietà viste nel capitolo 2 per le sequenze tipiche, con la differenza che i suoi elementi non sono piu' singole sequenze, ma coppie di sequenze descritte da una probabilità congiunta, come già ripetuto questo è il caso di una sequenza in ingresso ed una in uscita da un canale di comunicazione. Valgono anche le

osservazioni fatte precedentemente, essendo arrivati ai risultati seguendo gli stessi procedimenti.

Grazie alle proprietà viste possiamo già dare un'idea su quello che sarà il nostro metodo di decodifica. Dopo aver trasmesso una sequenza tra gli input possibili, $x_n \in \mathcal{X}_n$, e aver ricevuto la sequenza y_n decidiamo che è stato trasmesso l'indice $\tilde{W}$ se $\left(X^n\left(\tilde{W}\right), y^n\right) \in A_\varepsilon^{(n)}$. In altre parole se l'input generato dall'indice stimato e il segnale ricevuto sono congiuntamente tipici allora scegliamo quell'indice. Da questo traiamo una conclusione piuttosto qualitativa: di sequenze Y^n tipiche c'è ne sono circa $2^{n(H(Y))}$ e per ognuna di queste $\approx 2^{n(H(X|Y))}$ sono le coppie (X^n, y^n) di sequenze congiuntamente tipiche. La probabilità che una sequenza x^n sia congiuntamente tipica con y^n è circa

$$\frac{2^{n(H(X|Y))}}{2^{n(H(X))}} = 2^{-n(I(X,Y))} \tag{3.16}$$

Ecco allora che possiamo scegliere $2^{n(I(X,Y))}$ sequenze x^n prima di trovarne una che sia congiuntamente tipica con un altro segnale d'uscita.

In alternativa possiamo pensare di dividere l'insieme di sequenze Y^n tipiche in circa $2^{n(H(Y|X))}$ sottoinsiemi, ovvero per ogni sottoinsieme abbiamo le sequenze tipiche congiunte con un determinato ingresso. Dobbiamo anche imporre che questi sottoinsiemi siano una partizione dell'insieme dei segnali d'uscita per poter decodificare poi efficacemente. Allora abbiamo a disposizione $2^{n(I(X,Y))}$ di questi insiemi, e ad ognuno di questi faremo corrispondere una parola di codice. Il che porta alla stessa conclusione del primo ragionamento.

Infine se possiamo scegliere circa $2^{n(I(X,Y))}$ parole di codice per avere una decodifica senza sovrapposizioni, è chiaro che il numero di bit che possiamo trasmettere è $nI(X,Y)$. Questo giustifica il fatto di aver chiamato informazione la quantità $I(X,Y)$ che per quanto detto esprime proprio il numero di bit che un canale può inviare, per una certa distribuzione $p(x)$, quindi quanta informazione può trasferire.

Come detto all'inizio del capitolo vogliamo arrivare a dimostrare il teorema sulla codifica di canale sfruttando le proprietà delle sequenze congiuntamente tipiche, questo fu infatti l'approccio usato da Shannon quando lo dimostrò per la prima volta. La prossima proprietà sarà cruciale per poter calcolare poi la probabilità di errore nella dimostrazione del teorema.

Teorema 3.2. *Siano* (X^n, Y^n) *come nel Teorema 3.1 e* $\left(\tilde{X}^n, \tilde{Y}^n\right)$ *due sequenze indipendenti, distribuite secondo* $p(x^n)$ *e* $p(y^n)$ *rispettivamente:* $\left(\tilde{X}^n, \tilde{Y}^n\right) \sim p(x^n)p(y^n)$. *Allora:*

$$P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) \leq 2^{-n(I(X,Y)-3\varepsilon)} \tag{3.17}$$

e per n sufficientemente grande:

$$P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) \geq (1-\varepsilon)\, 2^{-n(I(X,Y)+3\varepsilon)} \tag{3.18}$$

Dimostrazione: Per la prima parte

$$P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) = \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n)p(y^n) \tag{3.19}$$

avendo usato il fatto che le marginali di $\tilde{X}^n, \tilde{Y}^n$ sono quelle di X^n e Y^n. Da (3.5) e (3.6) si ha che $p(x^n) \leq 2^{-n(H(X))-\varepsilon)}$ e $p(y^n) \leq 2^{-n(H(Y))-\varepsilon)}$. Usando poi la *2* del Teorema 3.1:

$$\begin{aligned} P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) &\leq 2^{n(H(X,Y)+\varepsilon)} 2^{-n(H(X))-\varepsilon)} 2^{-n(H(Y))-\varepsilon)} \\ &= 2^{-n(I(X,Y)-3\varepsilon)} \end{aligned}$$

Che dimostra la prima parte del teorema. Per la seconda parte

$$\begin{aligned} P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) &= \sum_{(x^n,y^n)} p(x^n)p(y^n) \\ &\geq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} 2^{-n(H(X))+\varepsilon)} 2^{-n(H(Y))+\varepsilon)} \\ &\geq \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X))+\varepsilon)} 2^{-n(H(Y))+\varepsilon)}. \end{aligned}$$

Per n sufficientemente grande vale la proprietà *3* del Teorema 3.1 e quindi

$$\begin{aligned} P\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\varepsilon^{(n)}\right) &\geq (1-\varepsilon)\, 2^{n(H(X,Y)-\varepsilon)} 2^{-n(H(X))+\varepsilon)} 2^{-n(H(Y))+\varepsilon)} \\ &= (1-\varepsilon)\, 2^{-n(I(X,Y)+3\varepsilon)} \end{aligned}$$

□

3.3 Il teorema della codifica di canale

Fino ad ora non abbiamo ancora parlato di probabilità di errore, infatti può accadere che non siamo in grado di decodificare il messaggio ricevuto, oppure semplicemente la decodifica

è sbagliata perché ritorna un messaggio differente da quello effettivamente trasmesso. I motivi di errori nel canale possono essere dei più svariati, quello su cui ci concentreremo noi è ammettere una probabilità di errore, eventualmente da far diventare piccola a piacere, senza indagare sulle sue cause fisiche.

Introduciamo alcune definizioni per la probabilità di errore

Definizione 3.5. *Probabilità di errore per la i-esima parola di codice*

$$\lambda_i = P\left(\tilde{W} \neq i | W = i\right) = \sum_{y^n \in Y^n : g(y^n) \neq i} P\left(y^n | X^n(i)\right) \tag{3.20}$$

dove $g(\cdot)$ è la funzione di decodifica del codice, λ_i rappresenta la probabilità condizionata di errore dato che il messaggio in ingresso è $X^n(i)$, o equivalentemente che il messaggio scelto è l'i-esimo.

Definizione 3.6. *Probabilità massima di errore per un codice* (M, n)

$$\lambda_{max} = \max_i \lambda_i \tag{3.21}$$

Definizione 3.7. *Probabilità di errore di un codice* (M, n)

$$P_e = P(\tilde{W} \neq W) = \sum_{w \in 1,2,\ldots,M} P(W = w)\lambda_w \tag{3.22}$$

Se gli indici w sono distribuiti uniformemente, allora (3.22) diventa la probabilità di errore media:

$$P_e = \frac{1}{M} \sum_w \lambda_w \tag{3.23}$$

Definizione 3.8. *Un rate R si dice realizzabile se esiste una sequenza di un codici* (M, n) *per cui*

$$\lim_{n \to \infty} \lambda_{max} = 0 \tag{3.24}$$

Il teorema della codifica di canale mostrerà come solo i codici con rate $R < C$ siano realizzabili, fornendo così una definizione operativa di quella che è la capacità di un canale di comunicazione.

Abbiamo ora tutti gli strumenti necessari per dimostrare il teorema sulla codifica di canale, o secondo teorema di Shannon. Cominciamo con l'enunciato.

Teorema 3.3. *(Codifica di canale). Per un canale senza memoria, discreto, per ogni rate R tale che $R < C$, dove C è la capacità del canale, esiste una sequenza di codici (M, n) tale che*

$$\lim_{n\to\infty} \lambda_{max} = 0 \tag{3.25}$$

Al contrario ogni sequenza di codici (M, n) con $\lambda_{max} \to 0$ per $n \to \infty$ deve avere rate $R \leq C$.

Il teorema è fondamentale nella teoria delle comunicazioni, infatti esso ci dice che possiamo avere una trasmissione di dati affidabile nonostante il canale introduca degli errori. Chiaramente non sarebbe possibile pretendere una probabilità di errore nulla, ma ammetterla piccola a piacere al crescere di n è possibile, e questo permette di avere una trasmissione dati affidabile. La condizione importante espressa dal teorema è: per trasmettere in maniera sicura il rate del codice deve essere inferiore alla capacità del canale. La nozione di capacità acquista un ruolo importante, definisce il limite per cui possiamo inviare informazione in maniera affidabile, attraverso un canale che introduce rumore. Dimostriamo ora il Teorema 3.3

Dimostrazione: Il punto fondamentale è calcolare la probabilità di errore non per un solo codice ma per una scelta di codici differenti, un codice viene generato casualmente e poi reso noto a ricevitore e trasmettitore. Per generare il codice: per ogni indice $i \in \{1, 2, ..., M\}$ generiamo gli n simboli di $X^n(i)$ casualmente e in maniera indipendente secondo una distribuzione $p(x)$. Ora a entrambe le estremità del canale sono note le parole di codice generate per ogni indice. Gli indici sono estratti con una densità di probabilità uniforme, $P(W = w) = \frac{1}{M}$, $\forall w \in \{1, 2, ..., M\}$. Un indice w viene scelto casualmente e inviata attraverso il canale la parola di codice generata precedentemente secondo la distribuzione $p(x)$. Definito Err l'evento di un errore:

$$P(Err) = \sum_{w=1}^{M} P\left(\tilde{W} \neq w | W = w\right) P\left(W = w\right) \tag{3.26}$$

Questa, avendo scelto una distribuzione uniforme per gli indici è equivalente a (3.23). Inoltre poiché abbiamo scelto casualmente i simboli per la parola di codice, λ_w non dipende da w, e senza perdita di generalità prendiamo $w = 1$

$$P(Err) = P_e = \lambda_1 = P\left(\tilde{W} \neq 1 | W = 1\right) \tag{3.27}$$

Per calcolare esattamente la probabilità di errore dobbiamo descrivere nel dettaglio la procedura per la decodifica. Premettiamo che la decodifica ottima si ottiene con una stima di massima verosimiglianza fatta a posteriori sul segnale ricevuto (una descrizione del metodo si può trovare in [3]), tuttavia useremo un approccio più semplice da analizzare basato sulle proprietà delle sequenze tipiche, approccio, come detto utilizzato da Shannon nel suo lavoro originale.

Supponiamo che venga scelto l'indice w e trasmessa quindi la sequenza corrispondente in ingresso al canale, diciamo questa x^n. All'altra estremità del canale si riceve una sequenza y^n secondo la probabilità di transizione del canale, $p(y^n|x^n)$. Diremo poi che è stato inviato l'indice $\tilde{w}$ se $(X^n(\tilde{w}), y^n) \in A_\varepsilon^{(n)}$, dove $X^n(\tilde{w})$ rappresenta il segnale d'ingresso corrispondente all'indice $\tilde{w}$. Cerchiamo quindi una sequenza di ingresso che sia congiuntamente tipica con il segnale che effettivamente viene ricevuto, in questo modo potremo sfruttare le proprietà viste in precedenza per il calcolo di P_e. Imponiamo anche che l'indice $\tilde{w}$ sia unico, percui se l'indice non esiste oppure ne esistono più di uno la funzione di decodifica segnala un errore. Osserviamo come questo ultimo evento descritto non concorre ad incrementare P_e, infatti l'errore è segnalato e possiamo sapere quando avviene.

Ritorniamo al calcolo di P_e, definito l'evento percui la sequenza ricevuta è congiuntamente tipica con $X^n(i)$

$$E_i = \{(X^n(i), y^n)\} \tag{3.28}$$

Perché si verifichi un errore di decodifica o $X^n(1)$ non è congiuntamente tipica con y^n oppure una parola di codice diversa è congiuntamente tipica con y^n, in termini di eventi E_i:

$$\begin{aligned} P(Err) &= P\left({E_1}^c \cup E_2 \cup ... \cup E_M | W = 1\right) \\ &\leq P\left({E_1}^c\right) + \sum_{i=2}^{M} P\left(E_i\right) \end{aligned} \tag{3.29}$$

per la proprietà della probabilità dell'unione di eventi.

Per il primo termine condizioniamo per $W = 1$, i simboli di $X^n(1)$ sono generati indipendentemente l'uno dall'altro, e poiché y^n è il risultato dell'invio di $X^n(1)$ abbiamo che $(X^n(1), y^n) \sim p(y^n|X^n(1))p(W = 1) = p(X^n(1), y^n)$. Considerando infine che il canale è senza memoria per valori sufficientemente grandi di n, possiamo far valere la proprietà 1 del Teorema 3.1

$$P\left({E_1}^c\right) \to 0 \tag{3.30}$$

Condizionando rispetto a w con $2 \leq w \leq M$ valgono ancora le considerazioni fatte per il caso precedente tranne per il fatto che questa volta i segnali $X^n(w)$ non sono la causa di y^n, quindi i segnali hanno stessa distribuzione di X^n e Y^n, ma sono indipendenti. Per il Teorema 3.2 vale:

$$P\left(E_w\right) \leq 2^{-n(I(X,Y)-3\varepsilon)} \qquad 2 \leq w \leq M \tag{3.31}$$

L'equazione (3.29) per n abbastanza grande diventa

$$P(Err) \leq \varepsilon + \sum_{2}^{M} 2^{-n(I(X,Y)-3\varepsilon)} \tag{3.32}$$

Dalla definizione di rate di un codice riscriviamo $M = 2^{nR}$ e otteniamo

$$\begin{aligned} P(Err) &\leq \varepsilon + \left(2^{nR} - 1\right) 2^{-n(I(X,Y)-3\varepsilon)} \\ &= \varepsilon + 2^{nR} 2^{-n(I(X,Y)-3\varepsilon)} - 2^{-n(I(X,Y)-3\varepsilon)} \\ &\leq \varepsilon + 2^{-n(I(X,Y)-R-3\varepsilon)} \end{aligned}$$

Se è soddisfatta la condizione $R < I(X,Y) - 3\varepsilon$ il secondo termine della somma nell'ultima disuguaglianza tende a 0 per n grande. Possiamo quindi minimizzarlo e in particolare renderlo ε. Concludiamo con la relazione per la probabilità di errore:

$$P(Err) \leq 2\varepsilon \qquad \text{Per } n \text{ abbastanza grande} \tag{3.33}$$

Vediamo che la probabilità di errore trovata è la media di P_e su ogni codice generato. Un particolare codice $\mathcal{C}$ viene generato con probabilità $P(\mathcal{C})$. Condizionando su $\mathcal{C}$

$$P(Err) = P\left(\tilde{W} \neq W\right) = \sum_{\mathcal{C}} P(\mathcal{C}) P(Err|\mathcal{C}) \tag{3.34}$$

La relazione è vera poiché ogni parola di codice è generata indipendentemente, e quindi anche i codici. Questo permette di concentrarci sulla probabilità di errore assumendo che sia stato scelto un indice W, come fatto in (3.27).

Per quanto detto allora esiste un codice $\mathcal{C}^*$ per cui

$$P(Err|\mathcal{C}^*) \leq P(Err) \leq 2\varepsilon \tag{3.35}$$

Per dimostrare la prima parte del teorema sulla codifica di canale dobbiamo trovare una relazione su λ_{max} e non sulla probabilità di errore in media. Per fare questo, scegliamo $\mathcal{C}^*$ e per (3.27), (3.23), $P(Err)$ è la media aritmetica delle probabilità di errore quindi almeno

metà dei λ_w sono minori di 4ε. Scartiamo dal codice metà degli indici, in modo da tenere solo quella metà con $\lambda_w \leq 4\varepsilon$. Questo nuovo codice ha quindi $\lambda_{max} \leq 4\varepsilon$ e

$$R' = \frac{\log \frac{M}{2}}{n} = R - \frac{1}{n} \tag{3.36}$$

La variazione del tasso del codice per valori elevati di n è trascurabile dimezzando le parole di codice.

Per concludere prendiamo $p^*(x)$ quella distribuzione per cui $I(X,Y) = C$ e la condizione $R < I(X,Y) - 3\varepsilon$ diventa $R < C$. □

Fino a qui abbiamo dimostrato la prima parte del teorema, ovvero che dato un codice con rate $R < C$ allora questo deve essere un codice realizzabile. Facciamo alcune considerazioni su quanto visto.

Si è visto come un codice $\mathcal{C}$ con $P(Err) \leq 4\varepsilon$ esista, la dimostrazione non indica tuttavia come costruirlo. È possibile per trasmettitore e ricevitore decidere di utilizzare il miglior codice, generare poi con la procedura descritta tutti i codici e tramite una ricerca esaustiva trovare il codice ottimo e utilizzarlo. Questa procedura è difficile, se non impossibile, da realizzare anche per valori piccoli di n a causa dell'elevato numero di codici che bisogna poi ispezionare. Precisamente: un codice $\mathcal{C}$ (M, n) ha M parole di codice di n simboli

$$\mathcal{C} = \begin{bmatrix} x_1(1) & \cdots & x_n(1) \\ \vdots & \ddots & \vdots \\ x_1(M) & \cdots & x_n(M) \end{bmatrix} \tag{3.37}$$

Detto $\mathcal{X}$ l'alfabeto dei simboli per le parole di codice, abbiamo per ogni simbolo $|\mathcal{X}|$ possibilità, il numero di codici è quindi $|\mathcal{X}|^{nM}$. Per rate prossimi alla capacità il numero è circa $|\mathcal{X}|^{n2^{nC}}$, il numero di codici da ispezionare per trovare il migliore cresce con doppio esponenziale all'aumentare di n. Trovare il codice in questa maniera non è fattibile, infatti in ambito applicativo si usano altre tecniche di codifica.

3.4 L'inverso del teorema

Dedichiamo questa parte del capitolo alla dimostrazione dell'inverso del teorema sulla codifica di canale.

Dimostriamo che dato un codice realizzabile, questo ha rate $R \leq C$. Per dimostrarlo useremo la disuguaglianza di Fano (1.4).

Dimostrazione: Se un codice è realizzabile allora $\lambda_{max} \to 0$ per $n \to \infty$, e deve essere anche

$$\lim_{n\to\infty} P_e = 0 \tag{3.38}$$

infatti ricordiamo che P_e rappresenta la media delle probabilità di errore.

Osserviamo che l'intero processo di invio di un segnale è suddiviso in 4 punti fondamentali:

- Scelta di un indice W nell'insieme $\mathcal{W} = \{1, 2, ..., M\}$.
- Codifica nel segnale d'ingresso $X^n(W)$.
- Passaggio attraverso il canale di $X^n(W)$ e conseguente ricezione di Y^n.
- Stima del messaggio spedito: $\tilde{W} = g(Y^n)$.

Poiché per effettuare uno dei passaggi ci serve essere a conoscenza solo del risultato dell'operazione precedente, questi formano una catena di Markov, che sintetizziamo con:

$$W \to X^n \to Y^n \to \tilde{W}$$

Calcoliamo ora l'entropia per W

$$\begin{aligned} H(W) &= -E\left(\log p(w)\right) \\ &= -\sum_{w=1}^{M} p(w) \log p(w) \\ &= -\frac{1}{M} \sum_{w=1}^{M} \log \frac{1}{M} \end{aligned}$$

Usando la distribuzione di probabilità uniforme di W, $p(w) = \frac{1}{M}$. Ricordando infine la definizione di rate di un codice (3.4)

$$H(W) = \log M = nR$$

Dalle proprietà dell'entropia risulta inoltre:

$$H(W) = H(W|\tilde{W}) + I(W, \tilde{W}). \tag{3.39}$$

Per il primo termine della parte a destra dell'equazione possiamo usare (1.4)

$$1 + P_e \log |\mathcal{W}| \geq H(W|\tilde{W})$$

e ottenere:

$$H(W|\tilde{W}) \leq 1 + P_e nR.$$

Utilizzando questa ultima relazione (3.39) diventa

$$H(W) \leq 1 + P_e nR + I(W, \tilde{W})$$

inoltre per la disugualianza sull'informazione delle catene di Markov (1.3)

$$H(W) \leq 1 + P_e nR + I(X^n, Y^n). \tag{3.40}$$

Per l'informazione mutua a secondo membro di quest'ultima relazione, usiamo le proprietà dell'informazione mutua e otteniamo:

$$I(X^n, Y^n) = H(Y^n) - H(Y^n|X^n).$$

Sfruttando al secondo membro dell'ultima uguaglianza, le proprietà dell'entropia per sequenze di v.a. e la regola della catena per l'entropia:

$$I(X^n, Y^n) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|Y_1 ... Y_{i-1}, X^n)$$

Infine per l'assenza di memoria del canale:

$$\begin{aligned} I(X^n, Y^n) &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i, Y_i) \end{aligned}$$

e poiché $I(X_i, Y_i) \leq C$

$$I\left(X^n, Y^n\right) \leq nC \tag{3.41}$$

Quindi combinando (3.41) in (3.40)

$$\begin{aligned} H(W) &\leq 1 + P_e nR + nC \\ nR &\leq 1 + P_e nR + nC \\ R &\leq \frac{1}{n} + P_e R + C \end{aligned}$$

Per $n \to \infty$ otteniamo $R \leq C$. Dimostrando così anche la seconda parte dell'enunciato del teorema sulla codifica di canale. □

Il Teorema 3.3 nella sua interezza, ci fornisce una definizione operativa per la capacità, questa rappresenta la massima informazione trasferibile attraverso il canale. Infatti tentando di inviare più informazione attraverso il canale, infatti usando un rate $R > C$ andiamo incontro a probabilità d'errore non più tendenti a zero e quindi perdita di informazione.

3.5 Separazione tra codifica di sorgente e di canale

Abbiamo fino ad ora parlato di inviare informazione attraverso un canale, consideriamo adesso il problema partendo dalla sorgente che genera l'informazione e analizziamo in questa sezione se utilizzando un sistema di codifica unico per sorgente e canale otteniamo risultati diversi da quanto visto fin'ora. Infatti va notato che in questo capitolo non abbiamo considerato la codifica di sorgente, anche nella dimostrazione del Teorema 3.3 ci siamo limitati a guardare degli indici, che potevano rappresentare sequenze di una sorgente direttamente, oppure sequenze dopo una codifica. Quindi, se volessimo inviare l'informazione generata da una sorgente attraverso un canale con questo metodo dovremmo effettuare una codifica di sorgente, per rimuovere la ridondanza di questa, poi procedere ad una opportuna codifica di canale per trasmettere in maniera affidabile attraverso di esso.

Con una codifica unica di sorgente e canale invece codifichiamo le informazioni dalla sorgente e inviamo direttamente quanto ottenuto attraverso il canale. Pensiamo che questo ultimo approccio al problema porti a risultati migliori, poiché per realizzare un'unica codifica per entrambi, sorgente e canale, bisogna conoscere nello specifico la struttura della sorgente e del canale, e quindi avere una soluzione specifica per una particolare coppia sorgente canale che porti a prestazioni migliori rispetto ad una soluzione più generica per sorgente e canale separatamente.

Il prossimo teorema è analogo a quanto visto per la codifica di canale (Teorema 3.3) ma nella dimostrazione partiremo da una sorgente di informazione piuttosto che da indici di parole, in questo modo faremo anche una codifica di sorgente prima di inviare il messaggio. È interessante vedere come nel nostro caso, canale non condiviso da più utenti, i metodi diano risultati equivalenti. Questo ci permette di studiare separatamente i processi di codifica, focalizzandoci solo sulle caratteristiche della sorgente e del canale separatamente, riducendo la complessità della progettazione.

Il sistema che analizzeremo è molto simile a quello che abbiamo analizzato nella sezione precedente. Abbiamo una sorgente V e vogliamo trasmettere sequenze di n-simboli dell'alfabeto $\mathcal{V}$. Codifichiamo le sequenze V^n in segnali $X^n(V^n)$ da inviare lungo il canale,

riceviamo la sequenza Y^n all'altra estremità e tramite un'opportuna codifica stimiamo la sequenza originale, $\tilde{V^n}$.

Teorema 3.4. *(Codifica di sorgente e canale). Dato un processo aleatorio V_n con alfabeto $\mathcal{V}$ finito e che soddisfi il Teorema 2.1 e $H(V) < C$, con C capacità di un canale dato, esiste una codifica di sorgente e canale, per cui è possibile trasmettere in maniera affidabile i simboli di V_n attraverso il canale, $P(Err) \to 0$ per valori grandi di n. Vale anche l'inverso, se è possibile trasmettere con $P(Err) \to 0$ un processo aleatorio stazionario, allora deve valere $H(V) < C$.*

Dimostrazione: Osserviamo che da (3.26)

$$\begin{aligned} P(Err) &= \sum_{v^n} P\left(V^n \neq \tilde{V^n} | V^n = v^n\right) P\left(V^n = v^n\right) \\ &= \sum_{v^n} \sum_{y^n : g(y^n) \neq v^n} P\left(y^n | V^n = v^n\right) P\left(V^n = v^n\right) \end{aligned}$$

Come nella dimostrazione del teorema sulla codifica di canale calcoliamo la probabilità di errore per il sistema. Poichè V soddisfa la AEP possiamo codificare solo le parole che appartengono all'insieme tipico $A_\varepsilon^{(n)}$, infatti per il teorema 2.2, per n grandi

$$P(V^n \notin A_\varepsilon^{(n)}) \to 0 \tag{3.42}$$

ovvero il contributo delle parole che non stanno in $A_\varepsilon^{(n)}$ è trascurabile.

Ricordando che $|A_\varepsilon^{(n)}| \leq 2^{n(H(V)+\varepsilon)}$, dobbiamo codificare al più $2^{n(H(V)+\varepsilon)}$ parole di codice, ovvero dobbiamo trasmettere $n(H(V)+\varepsilon)$ bit e il rate del codice è

$$\frac{\log\left(2^{n(H(V)+\varepsilon)}\right)}{n} = H(V) + \varepsilon = R \tag{3.43}$$

Siamo nelle condizioni in cui vale il teorema della codifica di canale, con $R < C$ possiamo quindi inviare la nostra sequenza di bit con probabilità di errore tendente a zero per n abbastanza grandi. $P(Err)$ è il risultato o di un errore nelle codifica di sorgente o di trasmissione, sfruttando le proprietà per la funzione di probabilità sull'unione di eventi:

$$P(Err) \leq P\left(V^n \notin A_\varepsilon^{(n)}\right) + P\left(\tilde{V^n} \neq V^n\right) \leq 2\varepsilon \tag{3.44}$$

Dove l'ultimo passaggio seque utilizzando il Teorema 3.3 e per le considerazioni fatte prima. Questo conclude la dimostrazione per la prima parte del teorema.

Dimostriamo che se possiamo inviare una sequenza con $P(Err) \to 0$ allora l'entropia della sorgente è $H(V) \leq C$. Anche qui la dimostrazione segue quella fatta per l'inverso del teorema 3.3.

Ipotizziamo che esista una sequenza di codici come quella descritta sopra con $P(Err) \to 0$.

Dalla definizione di entropia per un processo stazionario

$$H(V) \leq \frac{H(V^n)}{n} = \frac{1}{n} H\left(V^n | \tilde{V^n}\right) + \frac{1}{n} I\left(V^n, \tilde{V^n}\right) \tag{3.45}$$

Utilizzando (1.4)

$$H(V) \leq \frac{1}{n}\left(1 + P(Err) \log |\mathcal{V}|^n\right) + \frac{1}{n} I\left(V^n, \tilde{V^n}\right) \tag{3.46}$$

Visto che $V \to X^n \to Y^n \to \tilde{V^n}$ formano una catena di markov possiamo usare (1.3) e ottenere

$$\begin{aligned} H(V) &\leq \frac{1}{n}\left(1 + P(Err)\, n \log |\mathcal{V}|\right) + \frac{1}{n} I\left(X^n, Y^n\right) \\ &\leq \frac{1}{n}\left(1 + P(Err)\, n \log |\mathcal{V}|\right) + C \end{aligned}$$

Dove l'ultimo passaggio segue da (3.41). Per $n \to \infty$ nell'ultima equazione otteniamo $H(V) \leq C$, che conclude la dimostrazione del teorema. □

Abbiamo così dimostrato anche questo importante risultato: come detto all'inizio la dimostrazione fa uso di un'unica codifica per sorgente e canale. Nonostante questo, il risultato non cambia dal caso in cui si considerino i problemi di codifica separatamente, permettendo appunto una più semplice progettazione. Il risultato può essere controintuitivo sotto un certo punto di vista perchè si può pensare che un'unica codifica possa ottenere prestazioni migliori, essendo specificamente studiata per una sola combinazione di sorgente e canale. Si possono anche fare esempi che portano a pensare che ciò sia vero, come la voce umana che anche nel caso di un alto rumore ambientale riusciamo lo stesso a comprendere. In questi casi evidentemente la ridondanza introdotta dalla sorgente è appropriata al canale, nel senso che ne corregge efficacemente gli errori.

Capitolo 4

Conclusioni

Con la proprietà di equipartizione asintotica è stato possibile dividere l'insieme degli esiti di una sequenza di variabili aleatorie i.i.d. in due insiemi: insieme tipico e il suo complementare. Le proprietà dell'insieme tipico ci dicono che preso un esito casuale della sequenza di v.a. esso ha un'elevata probabilità di appartenere all'insieme tipico, abbiamo sfruttato questo per effettuare una codifica di sorgente senza perdita di informazione.
Abbiamo visto anche, come la proprietà di equipartizione asintotica si riveli critica per dimostrare alcuni risultati fondamentali nella teoria dell'informazione. Grazie ad essa abbiamo dimostrato il Teorema della codifica di canale, e sebbene non ci sia un modo diretto per costruire codici efficienti è di fondamentale importanza sapere che una comunicazione affidabile è possibile nonostante il rumore. Senza questo risultato non avremmo le basi teoriche per ammettere che un messaggio possa essere ricevuto correttamente.
Di grande importanza è anche il Teorema della codifica di sorgente e canale, infatti sarebbe molto più oneroso progettare sistemi di comunicazione dovendo pensare contemporaneamente alle due codifiche. Poter affrontare il problema diviso in due parti risulta molto più pratico.

Bibliografia

[1] Thomas M. Cover, Joy A. Thomas, *Elements of information theory*, Wiley

[2] C.E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., 27:379-423,623-656,1948

[3] N. Benvenuto, M. Zorzi, *Principles of Communications Networks and Systems*, 257-264, Wiley, 2010

[4] Robert Gallager, course materials for 6.450 Principles of Digital Communications I, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [05/09/2012].