UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

"ASYMPTOTIC EQUIPARTITION PROPERTY E APPLICAZIONI NELLA COMPRESSIONE DEI DATI"

Relatore: Prof. Giancarlo Calvagno

Laureando: Francesco Maberino

ANNO ACCADEMICO 2021 – 2022

Data di laurea 17/03/2022

INDICE

Sommario……...3

1 – Introduzione: cenni storici……4

2 – Entropia dell'informazione ……..7

2.a – Contenuto di informazione di Shannon ……...7

2.b – Definizione di Entropia ……7

2.c – Proprietà dell'entropia……..8

2.d – Molteplici variabili e quarta proprietà……..9

2.e – Esempio di possibile misura del contenuto di informazione di Shannon ……10

3 – Compressione dati……13

3.a – Introduzione……..13

3.b – Esempio di compressione con perdite……..14

3.c – Esempio di compressione senza perdite……15

3.d – Prefix code e decodificabilità univoca……...16

3.e – Limiti che non è possibile superare con la compressione……17

4 – Asymptotic equipartition property……..18

4.a – Legge debole dei grandi numeri……18

4.b – Definizione di asymptotic equipartition property……18

4.c – Definizione di typical set e sue proprietà……..19

4.d – Dimensioni del typical set rispetto ad altri insiemi……....21

4.e – Approfondimento sulla relazione tra typical set e il più piccolo insieme δ-sufficiente……..22

4.f – Relazione tra asymptotic equipartition property e primo teorema di Shannon….23

4.g – Conclusione ……25

Sommario

La teoria dell'informazione è la scienza alla base della compressione e della trasmissione dei dati. Questa tesi ha lo scopo di analizzare alcune delle nozioni principali della teoria dell'informazione, con lo scopo di studiare il legame che unisce la asymptotic equipartition property e la compressione dei dati. Vedremo inoltre degli esempi introduttivi per inquadrare meglio i temi principali di ogni argomento.

Inizieremo analizzando l'entropia dell'informazione e le sue proprietà. Vedremo che essa è misura della quantità media di informazione contenuta in un messaggio o in un esito. Studieremo inoltre alcune delle proprietà fondamentali che la caratterizzano.

Dopodiché introdurremo la compressione dei dati. Approfondiremo che tipi di compressione si possono effettuare ed esamineremo inoltre la relazione indissolubile che lega la compressione dei dati all'entropia dell'informazione.

Ci soffermeremo poi sulla Asymptotic equipartition property. Questa proprietà permette di dividere lo spazio di tutte le sequenze di lunghezza n in due insiemi: il typical set e il suo insieme complementare, il non-typical set. Analizzeremo conseguentemente alcune delle proprietà principali che definiscono la natura del typical set. Infine, mostreremo che è possibile provare il primo teorema di Shannon, ossia il teorema che stabilisce i limiti della compressione dati possibile e il significato operativo di entropia dell'informazione, proprio grazie alla asymptotic equipartition property.

1- Introduzione: cenni storici

Fra luglio e ottobre 1948, Claude Shannon pubblicò il suo capolavoro: "A Mathematical Theory of Communication". L'opera, considerata "la Magna Carta dell'età dell'Informazione", [1], segna la nascita della teoria dell'informazione.

Prima di Shannon, l'ingegneria della comunicazione era sempre stata fortemente legata alla specifica fonte e al mezzo fisico di comunicazione. Il problema della comunicazione era affrontato come un problema di ricostruzione del segnale. Si trattava di trasformare un segnale, distorto dal mezzo fisico di comunicazione, per poi ricostruire l'originale il più fedelmente possibile.

Shannon si chiese invece se fosse possibile avere una grande teoria generale della comunicazione. Con la sua opera rispose proprio a questa domanda e diede alla luce una teoria unificante, che affonda le sue radici nel mondo della probabilità, della statistica, della matematica e dell'informatica, e che contiene le leggi fondamentali della compressione e della trasmissione dei dati.

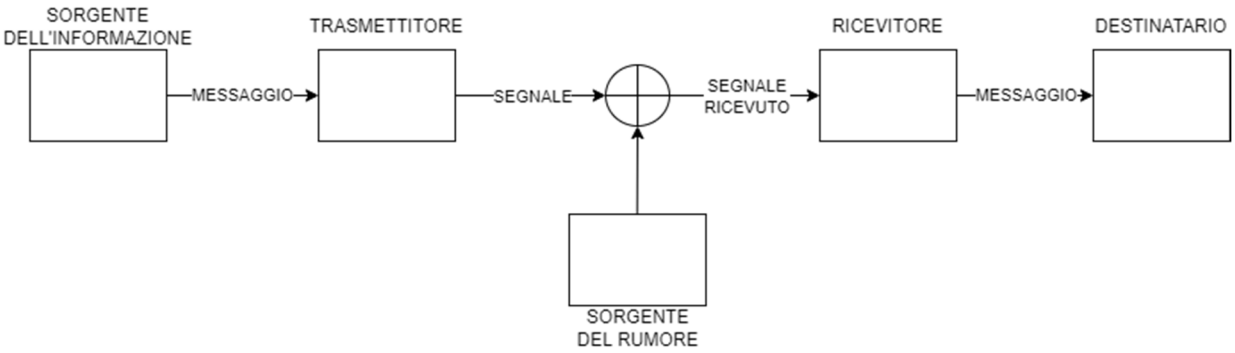


Fig. 1: Diagramma schematico del modello di comunicazione di Shannon

Secondo Shannon, "Gli aspetti semantici della comunicazione sono irrilevanti per il problema ingegneristico. L'aspetto significativo è che il messaggio che viene trasmesso è quello che viene selezionato da un set di possibili messaggi." [5]. Shannon comprese per primo che la chiave dell'informazione è l'incertezza. Così, cercò di reinterpretare la comunicazione attraverso una prospettiva probabilistica, portandola dal piano fisico a quello astratto, modellando l'incertezza usando la probabilità.

Tuttavia, era stato Hartley, vent'anni prima, a intuire che l'informazione è il risultato di una selezione tra un numero finito di possibilità: fu lui il primo a riconoscere il bisogno di introdurre

una misura per quantificare l'informazione. Usò la lettera H per denotare la quantità di informazione associata con n selezioni:

$$H = n \log(s)$$

dove s è il numero di simboli disponibile in ogni selezione. Ciò nonostante, Hartley non aveva preso in considerazione gli effetti del rumore, e nemmeno aveva considerato la fonte di informazione come un modello probabilistico.

Prima ancora che per la teoria dell'informazione, la modellazione probabilistica di fonti di informazione era stata usata e fu di grande importanza nel settore della crittografia. Rispettivamente, fin dal 1380 e dal 1658, tabelle di frequenze di lettere e coppie di lettere, vennero scritte per decrittare messaggi segreti.

Verso la fine della Seconda guerra mondiale, anche Shannon produsse un'opera sulla crittografia, nella quale incluse alcuni dei concetti che poi si ritrovano anche in "A Mathematical Theory of Communication". Nonostante Shannon a questo punto avesse già iniziato a lavorare alla sua grade opera, è innegabile che gli studi sulla crittografia ne abbiano influenzato il pensiero.

La grande intuizione di Shannon fu che la fonte dell'informazione deve essere modellata come un processo aleatorio: "Possiamo immaginare una fonte discreta generare il messaggio, simbolo per simbolo. I simboli successivi vengono scelti secondo alcune probabilità, che in generale dipendono sia da scelte precedenti, sia dai simboli in questione. Un sistema fisico, o un modello matematico di un sistema, che produce una tale sequenza di simboli, governato da un insieme di probabilità, è detto processo stocastico. Per converso, un processo stocastico che produce una sequenza discreta di simboli, scelti tra un insieme finito di elementi, può essere considerato una fonte discreta" [5].

Dato il contesto caratterizzato da incertezza e probabilità, Shannon si adoperò per determinare sistematicamente i limiti fondamentali della comunicazione. Egli articolò la risposta a questo quesito in tre parti, ideando la nozione di "bit" di informazione, come unità di misura dell'incertezza allo scopo di raggiungere tale obiettivo. Si servì, per la prima volta nella storia, del bit, quale cifra binaria che può assumere i valori 0 o 1.

Per prima cosa trovò un'espressione che determina il massimo numero di bit per secondo che possono essere comunicati in modo affidabile: la capacità di sistema, C.

Dopodiché stabilì il numero minimo di bit per secondo con cui un'informazione può essere comunicata. Chiamò questo concetto tasso di entropia, H. Esso rappresenta la quantità di incertezza che caratterizza un'informazione: più basso è il valore dell'entropia di un'informazione, più bassa è l'incertezza relativa a questa ed è quindi minore il numero di bit con cui si può rappresentare.

Infine, mostrò che è possibile raggiungere l'affidabilità della comunicazione se e solo se:

$$H < C$$

Quindi "l'informazione è come l'acqua, che scorre attraverso un tubo in modo sicuro, solo se la portata dell'acqua che vi scorre attraverso è inferiore alla capacità del tubo" [2].

Non possiamo comprendere appieno la Asymptotic Equipartition Property o la compressione dei dati, senza aver prima inteso il significato dell'entropia dell'informazione, così come pensata da Shannon.

2- Entropia dell'informazione

2.a – Contenuto di informazione di Shannon

Con l'entropia dell'informazione, Shannon fornisce un fondamento assiomatico per la misura dell'informazione, rifacendosi, per analogia, all'entropia di Boltzmann. Essa è uno dei concetti più importanti della teoria di Shannon, in quanto analizza quantitativamente la quantità di informazione di una distribuzione.

Prima di darne una definizione, è però importante comprendere il significato del contenuto di informazione di Shannon. Il contenuto di informazione di Shannon di un esito x è:

$$h(x) = \log \frac{1}{P(x)}$$

ed è misurato in bit. Esso rappresenta la quantità di sorpresa, o di informazione, contenuto da un particolare simbolo *x* di una distribuzione.

In generale, un esito è sorprendente quando il simbolo è raro, quindi quando $h(x)$ è grande e $p(x)$ è piccolo. Perciò, più grande è $h(x)$, più sorprendente è il risultato. D'altra parte, eventi più improbabili contengono meno informazione, mentre eventi certi non ne contengono proprio.

2.b – Definizione di Entropia

L'entropia di una variabile aleatoria X invece è definita come la media del contenuto di informazione di Shannon su uno spazio di probabilità:

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

Anche questa viene misurata in bit. L'entropia rappresenta quindi il valore atteso della "sorpresa/incertezza" contenuta da una distribuzione e misura quanto "sorprendente" è in media un esito. Intuitivamente, maggiore è l'entropia, e quindi più "sorprendente" è la distribuzione, più difficile è rappresentarla. Approfondiremo quest'ultimo argomento più nel dettaglio nei capitoli successivi.

2.c – Proprietà dell'entropia

L'entropia è caratterizzata da numerose proprietà, che ci aiutano a comprenderne meglio la natura. Le principali sono elencate di seguito. Vedremo le prime tre in questo paragrafo mentre la quarta in un paragrafo successivo.

Senza perdita di generalità, supponiamo di avere un alfabeto $A_x = \{1, 2, ..., m\}$.

- **Proprietà 1**: $H(X) \leq \log m$, con uguaglianza se e solo se $P(x) = 1/m$ per ogni x (lo spazio è uniforme)
- **Proprietà 2**: $H(X) \geq 0$, con uguaglianza se e solo se A_x è deterministico
- **Proprietà 3**: per una pmf Q, definita sullo stesso alfabeto di P, data una variabile aleatoria U,

  $$H_q(U) = \sum_{u \in U} p(u) log \frac{1}{q(u)}$$

  Quest'ultimo punto è di fatto il contenuto atteso dell'informazione di Shannon, ma invece che essere relativo alla pmf P, rappresenta la sorpresa associata alla v.a. U, che è distribuita secondo la pmf P, però incorrettamente assunta come se lo fosse secondo la pmf Q.

  La seguente disuguaglianza mostra invece che, se scegliamo la distribuzione sbagliata, saremo in media più sorpresi dagli esiti:

  $$H(U) \leq H_q(U)$$

  con uguaglianza se e solo se $P = Q$.

  Inoltre, questa proprietà può essere considerata equivalente alla disuguaglianza di Gibbs, che vedremo a breve.

Tutte queste proprietà possono essere dimostrate grazie alla **disuguaglianza di Jensen**, che afferma che, data una funzione convessa f, e una variabile aleatoria X, allora:

$$E[f(X)] \geq f(E[X])$$

con E che rappresenta il valore atteso. Se f è strettamente convessa e $E[f(X)] = f(E[X])$, allora la variabile aleatoria X è costante. La disuguaglianza può anche essere riscritta per funzioni concave, ma in tal caso la direzione della disuguaglianza deve essere invertita.

La divergenza Kullback-Leiber o entropia relativa tra due distribuzioni di probabilità *P(x)* e *Q(x)*, definite su uno stesso alfabeto A_x, è:

$$D_{KL}(P||Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

L'entropia relativa soddisfa la **Disuguaglianza di Gibbs**:

$$D_{KL}(P||Q) \geq 0$$

Con uguaglianza se e solo se $P = Q$. Può essere dimostrata usando il concetto di convessità e la disuguaglianza di Jensen.

Poiché l'entropia relativa risulta sempre maggiore di 0 grazie alla disuguaglianza di Gibbs, alcuni la considerano come una misura di distanza delle due distribuzioni. Tuttavia, generalmente l'entropia relativa non è simmetrica: se interscambiamo le distribuzioni *P* e *Q*, di solito ne risulta che $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Quindi, anche se l'entropia relativa viene a volte chiamata "distanza KL", essa non è esattamente una distanza.

Un'importantissima conseguenza di questa disuguaglianza è che l'entropia è massima quando un sistema è equiprobabile. Ovvero, l'esito di un esperimento casuale comporta informazione massima quando la distribuzione probabilistica degli esiti è uniforme.

2.d – Variabili multiple e quarta proprietà

Studiamo ora come si calcola l'entropia quando consideriamo più di una variabile aleatoria. Prendendo in esame due variabili aleatorie *X* e *Y*, possiamo definire l'entropia congiunta:

$$H(X,Y) = E\left[log\frac{1}{P(X,Y)}\right] = E\left[log\frac{1}{P(X)P(Y|X)}\right]$$

che quantifica la nostra sorpresa riguardo alla coppia *(X, Y)*. Si può invece definire l'entropia condizionale nel caso in cui conosciamo già l'esito di *Y* e abbiamo già conosciuto gli *H(Y)* bit di informazione che la variabile comporta. Si può quindi scrivere:

$$H(X|\,Y) = E\left[log\frac{1}{P(X|Y)}\right] = \sum_y P(Y=y)H(X|Y=y)$$

Dopo aver visto l'entropia congiunta, abbiamo tutti gli elementi per poter introdurre la quarta proprietà.

- **Proprietà 4**: L'entropia di variabili aleatorie indipendenti è uguale alla somma delle entropie delle singole variabili:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i)$$

Vale lo stesso per il contenuto di informazione di Shannon.

Tuttavia, l'entropia è caratterizzata dalla proprietà di subadditività. In generale quindi, vale che:

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

Con uguaglianza se e solo se X e Y sono appunto indipendenti. Ciò è dovuto al fatto che parte dell'informazione può essere contenuta nelle correlazioni delle due variabili.

La quantità di informazione contenuta nelle correlazioni delle due variabili, come emerge nella proprietà 4, è altresì misurabile. Questa misura, che quantifica quanta informazione una variabile aleatoria porta riguardo a un'altra variabile, si chiama **mutua informazione**, ed è definita come segue. Date due variabili X e Y, con pmf congiunta *P(x,y)*:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Questa definizione mostra che la mutua informazione rappresenta la riduzione in media della sorpresa, quando osserviamo variabili aleatorie correlate.

2.e – Esempio di possibile misura del contenuto di informazione di Shannon

Dopo aver inquadrato tutti questi concetti, può essere utile applicarli ad un esempio, per comprendere meglio la natura del contenuto di informazione di Shannon e dell'entropia, e come entrambi vengano misurati in bit. Prendiamo in esame il gioco della battaglia navale. In questo gioco si affrontano due avversari, che piazzano le loro navi in una posizione fissa nel mare, rappresentato da una griglia quadrata, e che a turno cercano di colpire e affondare le navi del nemico, sparando a uno specifico quadratino del mare nemico. L'esito di un attacco a uno specifico quadratino può essere "mancato", "colpito" o "colpito e affondato".

Noi però studieremo una versione semplificata del gioco. Nella nostra versione, ogni giocatore posiziona una sola nave su un singolo quadratino, in una griglia da 8x8 = 64 quadratini. Ogni attacco di un giocatore costituisce uno spazio di probabilità, con due possibili esiti $\{s, n\}$, che rispettivamente corrispondono il primo all'esito "colpito" e il secondo a "mancato". Inizialmente, $P(s) = 1/64$, mentre $P(n) = 63/64$. Invece al terzo tentativo, se entrambi i primi due sono andati a vuoto, si ha che $P(s) = 1/62$ e $P(n) = 61/62$.

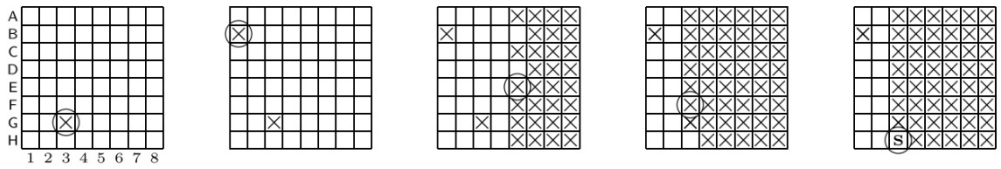
Il contenuto di informazione di Shannon di un esito *x* corrisponde a $h(x) = \log(1/P(x))$, che significa che, nell'improbabile caso in cui la nave avversaria venga colpita al primo colpo, otteniamo $h(x) = \log 64 = 6$ bit, dato che la nave poteva trovarsi in ognuno dei 64 quadratini. Se d'altra parte, l'esito del primo tentativo è "mancato", l'informazione di Shannon ottenuta è $h(x) = h_{(1)}(n) = \log\frac{64}{63} = 0.0227$ bit. Se falliamo anche il secondo tentativo, il contenuto di informazione di Shannon del secondo esito è $h_{(2)}(n) = \log\frac{63}{62} = 0.023$ bit. Se proseguiamo in questo modo, e manchiamo consecutivamente i primi 32 tentativi, l'informazione di Shannon raccolta è

$$\log\frac{64}{63} + \log\frac{63}{62} + \ldots + \log\frac{33}{32} = 0.0227 + 0.023 + \ldots + 0.043 = 1.0 \text{ bit}$$

Questo corrisponde alla domanda "x è uno dei 32 quadratini a cui abbiamo sparato?", e la risposta è no. Questa risposta esclude la prima metà delle ipotesi, e ci rende così un bit. Dopo 48 tentativi falliti, l'informazione ottenuta è 2 bit, e la posizione sconosciuta viene ridotta a un quarto dello spazio ipotizzato inizialmente.

Se colpiamo la nave con il 49 tentativo, quando sono rimasti 16 quadratini, il contenuto di informazione di Shannon dell'esito è $h_{(49)}(s) = \log 16 = 4.0$ bit. Il contenuto di informazione di Shannon di tutti gli esiti è $\log\frac{64}{63} + \log\frac{63}{62} + \ldots + \log\frac{17}{16} + \log\frac{16}{1} = 6.0$ bit.

Quindi, una volta scoperta la posizione della nave nemica, il contenuto dell'informazione di Shannon ottenuto è 6.0 bit. Questo risultato vale a prescindere da quando colpiamo la nave avversaria.



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Fig. 2: Battaglia navale semplificata, preso da [3]

Questo esempio mostra come il contenuto di informazione di Shannon sia una misura ragionevole del contenuto di informazione, e come questo venga effettivamente misurato in bit.

Di conseguenza, anche l'entropia dell'informazione, che è la media dei contenuti di informazione di Shannon per ogni elemento di un alfabeto, è allo stesso modo misurabile in bit.

Questo esempio dimostra inoltre che il contenuto di informazione di Shannon è strettamente connesso alla lunghezza di un codice binario che codifica gli esiti di un esperimento aleatorio, il che evidenzia l'esistenza di un legame tra l'entropia e la compressione dei dati.

Approfondiremo ulteriormente questa correlazione nel prossimo capitolo, sulla compressione dei dati.

3- Compressione dei dati

3.a - Introduzione

In questo capitolo approfondiremo la natura del legame che unisce entropia dell'informazione e compressione dei dati, fornendo altresì alcune nozioni generali su quest'ultima, utili alla comprensione di questo legame.

Per comprendere come funziona la compressione dati, immaginiamo di avere un generico messaggio (o file) da comprimere. Come intuì Shannon, possiamo pensare che questo sia composto da simboli che si susseguono quali esiti di una variabile aleatoria, generando il messaggio da comprimere, come in un processo stocastico.

La compressione non è che il processo che permette di rappresentare l'informazione in una forma compatta, al posto della forma originale o di una forma non compressa. Il generico file originale può venire codificato in forma compressa, usando codici costituiti da bit, in modo da risparmiare tempo di trasmissione e spazio in memoria. Quando la compressione di dati viene utilizzata per applicazioni come la trasmissione dati, lo scopo principale è quello di ridurre il più possibile il tempo necessario a trasferire dati da un luogo a un altro. In questo caso la velocità di trasmissione dipende da molteplici fattori, come il numero totale di bit da inviare, il tempo che il codificatore impiega a codificare il messaggio e il tempo necessario al decodificatore per rigenerare il messaggio originale. Se il nostro scopo è invece quello dell'archiviazione di dati, il nostro interesse principale, per quanto riguarda la compressione degli stessi, è il comprimere il più possibile il file originale, riducendone così al massimo lo spazio occupato.

La compressione dati può essere classificata in due categorie:

1) Compressione dati senza perdite

   Questo tipo di compressione opera mappando ogni elemento dell'alfabeto di una variabile aleatoria su codifiche diverse. Tuttavia, se le codifiche di alcuni elementi vengono accorciate, quelle di altri devono per forza essere più lunghe. Con questo tipo di algoritmi si può ridurre la dimensione di un generico messaggio originale, composto da sequenze di esiti, ed è inoltre possibile ricostruire il messaggio originale dal file compresso senza che avvenga alcuna perdita di informazioni. Gli algoritmi che operano in questo modo vengono anche chiamati reversibili, in quanto il messaggio originale può essere ricostruito nella fase di decodifica. Tecniche di compressione senza perdite

vengono utilizzate, tra le altre cose, per comprimere immagini mediche e testi e immagini che vengono conservati per motivi legali.

2) Compressione dati con perdite

   La compressione con perdite opera associando una codifica ad ogni elemento dell'alfabeto, ma talvolta alcuni elementi vengono associati alla stessa codifica. Queste tipologie di algoritmo ricostruiscono quindi il messaggio originale generando una perdita di informazione: non è possibile infatti ricostruire perfettamente l'originale durante la fase di decodifica. Per questo motivo questo tipo di compressione viene anche chiamata irreversibile. Il prodotto della decompressione in questo caso è solo una ricostruzione approssimativa. In generale, la compressione con perdite viene usata per applicazioni nelle quali il cervello umano non è in grado di riconoscere gli errori dovuti alle perdite avvenute durante la decompressione. Esempi tipici di questo utilizzo sono la compressione di immagini, video e audio.

In ogni caso, qualunque tipo di compressore costruiamo, dobbiamo sempre prendere in considerazione le probabilità dei diversi esiti.

3.b – Esempio di compressione con perdite

Vediamo brevemente dapprima un esempio di compressione con perdite, prima di affrontare il tema di quella senza perdite, che sarà approfondito maggiormente nei paragrafi successivi.

Prendiamo in considerazione il seguente alfabeto: A_X = {a, b, c, d, e, f, g, h}, caratterizzato da probabilità P_X = {1/4, 1/4, 1/4, 3/16, 1/64, /64, /64, /64}.

Introduciamo ora una nuova quantità, il raw bit content. Il raw bit content di una variabile aleatoria X è:

$$H_0(X) = \log|A_X|$$

Questa quantità consiste in un lower bound per il numero di domande binarie che sono sempre in grado di identificare un esito della variabile.

Il raw bit content dell'esempio che abbiamo preso in considerazione è 3 bit. Mediante questi 3 bit possiamo comporre gli 8 codici binari con cui è possibile mappare gli esiti. Possiamo però notare che $P(x \in \{a, b, c, d\}) = 15/16$.

Introduciamo la lettera δ come parametro per indicare il rischio che prendiamo quando usiamo una tecnica di compressione. In altre parole, il parametro δ rappresenta la probabilità che non ci sia il codice binario per un esito *x*.

Ritornando all'esempio, se siamo disposti a correre il rischio $\delta = 1/16$ di non avere un codice binario su cui mappare l'esito *x*, possiamo effettuare una compressione usando solo 4 codici binari al posto di 8, e solo 2 bit al posto di 3 per codificare gli esiti.

$\delta = 0$	
x	$c(x)$
a	000
b	001
c	010
d	011
e	100
f	101
g	110
h	111

$\delta = 1/16$	
x	$c(x)$
a	00
b	01
c	10
d	11
e	—
f	—
g	—
h	—

Fig. 3: Esempio di compressione con perdite, preso da [3]

La tecnica di compressione vista nell'esempio ci permette di risparmiare spazio in memoria e guadagnare velocità di trasmissione, dato che usiamo sia un numero minore di bit che di codici. Tuttavia, c'è un prezzo da pagare. Infatti, nell'esempio, mancano i codici binari per ben quattro degli otto esiti.

3.c – Esempio di compressione senza perdite

Vediamo adesso un esempio di compressione senza perdite. Questo ci sarà utile perché aprirà dei collegamenti molto interessanti con il concetto di entropia dell'informazione.

Prendiamo in esame una variabile aleatoria *X*, con alfabeto $A_X = \{a, b, c, d\}$ e probabilità $P_X = \{1/2, 1/4, 1/8, 1/8\}$.

Il raw bit content di questo esempio è 2 bit, con cui possiamo comporre i 4 codici binari, ai quali possiamo poi associare gli esiti. È quindi possibile codificare l'insieme in questione, assegnando a ogni elemento un codice binario di due bit. Può quindi risultare naturale pensare di assegnare i codici binari nel seguente modo: $\{a = 00, b = 01, c = 10, d = 11\}$. Questo è un primo esempio di codifica, ma possiamo anche renderla più efficiente. Il sistema in questione non è infatti caratterizzato da probabilità uniforme, ma, per esempio, l'esito "a" uscirà circa in metà degli esperimenti. Possiamo quindi accorciare il codice binario associatogli,

assegnandogli il codice "0" al posto di quello precedente. Questo ha però delle conseguenze significative. Per accorciare la codifica associata a uno degli esiti, è necessario allungare quella di altri esiti, come spiegheremo meglio più avanti. Possiamo quindi codificare gli esiti nel seguente modo: $\{a = 0,\ b = 10,\ c = 110,\ d = 111\}$.

Questo esempio mostra un primo utilizzo piuttosto rudimentale di prefix code, ossia un codice in cui nessun codice associato ad un elemento può essere il prefisso di un altro codice. Il vantaggio di questo è che è univocamente decodificabile (ovvero non esistono due sequenze che vengono codificate su uno stesso codice).

3.d – Prefix code e decodificabilità univoca

I prefix code sono molto utili in quanto non c'è bisogno di caratteri aggiuntivi, all'interno della codifica di una sequenza di esiti, che segnalino la fine del codice binario di un esito e l'inizio di quello successivo. Pertanto, risultano univocamente decodificabili.

Però è legittimo chiedersi quali siano le limitazioni imposte dalla condizione di decodificabilità univoca.

Abbiamo visto prima che se vogliamo accorciare il codice associato a un elemento dell'alfabeto, dobbiamo necessariamente allungare codici associati ad altri elementi. È come se esistesse una sorta di budget fisso, che possiamo spendere sulle codifiche. Questo è associato alle lunghezze in caratteri binari dei codici, con i codici più corti che sono più "costosi" di quelli più lunghi.

Per formalizzare questo concetto possiamo definire un budget di valore 1, da spendere in codice binario attraverso il quale codificare gli elementi in questione. D'altra parte, possiamo definire il costo di una codifica di lunghezza l come 2^{-l}. Per esempio, codifiche di lunghezza 3 costano 1/8 l'una. Possiamo spendere questo budget come vogliamo, ma se lo superiamo, sarà impossibile raggiungere la decodificabilità univoca. Questa condizione si chiama Disuguaglianza di Kraft, e afferma che per creare una tecnica di compressione univocamente decodificabile, le lunghezze delle codifiche degli esiti devono soddisfare la condizione:

$$\sum_{i=1}^{N} 2^{-l_i} \leq 1$$

con N che rappresenta la cardinalità dell'alfabeto considerato.

Quindi, possiamo creare una codifica senza perdite e univocamente decodificabile solo se non superiamo il nostro budget, rispettando la disuguaglianza di Kraft.

Comunque, anche rispettando la disuguaglianza possiamo chiederci fino a che punto è possibile comprimere, e se esiste un limite oltre cui non possiamo andare.

3.e – Limiti che non è possibile superare con la compressione

Introduciamo una nuova grandezza chiamata lunghezza attesa o lunghezza media di una codifica C, che, per la variabile aleatoria X è:

$$L(C,X) = \sum_{x \in A_x} P(x) l(x)$$

dove $l(x)$ rappresenta la lunghezza del codice binario a cui viene associato l'esito x.

L'estremo inferiore oltre cui non possiamo minimizzare questa grandezza è proprio il valore dell'entropia stessa $H(X)$, che rappresenta esattamente il lower bound della lunghezza attesa di una codifica univocamente decodificabile.

In ogni caso, questo risultato è raggiungibile solo se la disuguaglianza di Kraft è rispettata e se le lunghezze dei codici binari su cui vengono mappati sono ottimali. Tali lunghezze si dicono ottimali se le lunghezze dei codici binari associate agli esiti sono uguali ai contenuti di informazione di Shannon degli stessi: $l(x_i) = \log(1/p_i)$.

Non si può quindi comprimere al di sotto dell'entropia, ma quanto possiamo avvicinarci ad essa? Risponde a questa domanda il teorema della codifica di sorgente per i simboli di codice. Questo afferma che, per una variabile aleatoria X esiste un prefix code con lunghezza attesa che soddisfa la disuguaglianza seguente:

$$H(X) \leq L(C,X) \leq H(X) + 1$$

Abbiamo così chiarito quanto il concetto di entropia sia impattante per la compressione dati e come in un certo senso ne definisca la natura e i limiti. Ma questo è solo l'inizio, avremo modo di vedere anche nel prossimo capitolo quanto l'entropia sia fondamentale per la asymptotic equipartition property.

4- Asymptotic Equipartition Property

4.a – Legge debole dei grandi numeri

La asymptotic equipartition Property è una delle leggi cardine della teoria dell'informazione. Rappresenta per quest'ultima quello che la legge dei grandi numeri rappresenta per la teoria della probabilità. Le due leggi sono indissolubilmente legate, in quanto la asymptotic equipartition property è diretta conseguenza della legge debole dei grandi numeri.

Quest'ultima afferma che: data una sequenza di variabili aleatorie indipendenti e identicamente distribuite X_n, $\forall \varepsilon > 0, \forall \delta > 0$, esiste un N, tale che $\forall\, n > N$

$$P(|\hat{X}_{\mathrm{n}} - \mu| \geq \varepsilon) < \delta$$

ovvero che:

$$P(|\hat{X}_{\mathrm{n}} - \mu| < \varepsilon) > 1 - \delta$$

dove $\hat{X}_{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è la media campionaria, e μ che rappresenta il valore atteso di X_n, dato che si tratta di variabili aleatorie i.i.d.

Quindi, con altre parole, questa legge afferma che la media campionaria $\hat{X}_{\mathrm{n}}$ tende al valore atteso $E[X_n]$, per grandi valori di n.

4.b – Definizione di asymptotic equipartition property

La **Asymptotic Equipartition Property** afferma similarmente che, data una sequenza di n variabili aleatorie indipendenti e identicamente distribuite $(X_1, X_2, \ldots, X_n)$ con probabilità $p(x)$, per $n \to \infty$ si ha

$$-\frac{1}{n}\log[p(X_1, X_2, \ldots, X_n)] \to H(X)$$

La dimostrazione di questa proprietà è un'applicazione diretta della legge debole dei grandi numeri.

Dimostrazione: le funzioni di variabili aleatorie indipendenti e identicamente distribuite sono anch'esse variabili aleatorie indipendenti. Pertanto, dato che X_n sono i.i.d., anche $\log[p(X_n)]$ lo sono. Quindi per la legge debole dei grandi numeri

$$-\frac{1}{n}\log[p(X_1, X_2, \ldots, X_n)] = -\frac{1}{n}\sum_{i=1}^{n}\log[p(\mathrm{Xi})] \to -E\log[p(X)] = H(X)$$

Questo dimostra il teorema.

Uno dei vantaggi principali di questa proprietà è che permette di dividere l'insieme di tutte le possibili sequenze $(X_1, X_2, ..., X_n)$ di lunghezza n in due insiemi: il primo è il "typical set", o insieme delle sequenze tipiche, nel quale l'entropia campione è vicina a $H(X)$, mentre il secondo è il set non tipico, che contiene tutte le altre sequenze.

La nostra attenzione sarà tuttavia rivolta principalmente al typical set, per le sue peculiarità: qualunque proprietà verrà dimostrata per le sequenze tipiche, questa sarà vera con alta probabilità e determinerà il comportamento generale di sequenze lunghe.

4.c – Definizione di typical set e sue proprietà

Cerchiamo quindi di trovare una definizione per il typical set e per le sequenze che lo compongono. Una sequenza di esiti $\boldsymbol{x_n} = (X_1, X_2, ..., X_n)$ si dice ε-tipica per una *sorgente senza memoria X*, per ogni $\varepsilon > 0$, se

$$\left| -\frac{1}{n}\log[p(\boldsymbol{x_n})] - H(X) \right| \leq \varepsilon$$

o, equivalentemente

$$2^{-n[H(X)+\varepsilon]} \leq p(\boldsymbol{x_n}) \leq 2^{-n[H(X)-\varepsilon]}$$

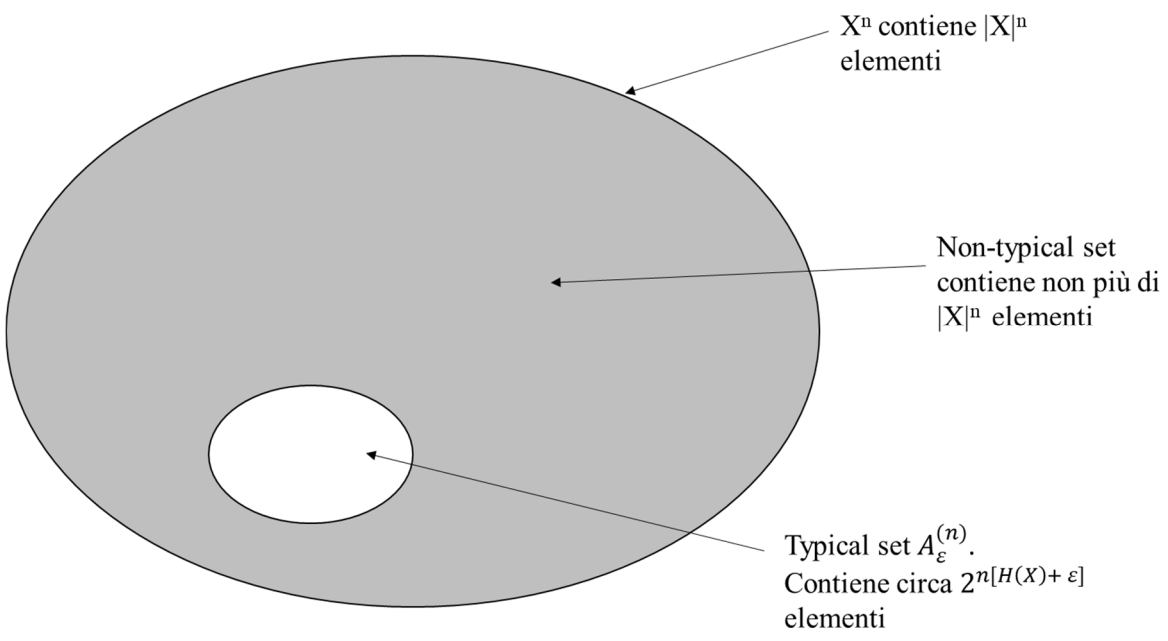


Fig. 4: Typical set e non typical set

Una sorgente priva di memoria è una sorgente tale per cui ogni messaggio è una variabile aleatoria indipendente e identicamente distribuita.

La costante ε invece ha la funzione di specificare quanto la probabilità di una sequenza di lunghezza n debba essere vicina a 2^{-nH}, affinché questa possa essere considerata tipica. L'insieme che contiene tutte tali sequenze, denotato con il simbolo $A_\varepsilon^{(n)}$ viene appunto chiamato typical set.

Tale insieme gode di alcune proprietà, che discendono direttamente dalla asymptotic equipartition property:

- **Proprietà 1**: Se una sequenza $\boldsymbol{x_n} \in A_\varepsilon^{(n)}$, allora:$H(X) - \varepsilon \leq \frac{1}{n}\log[p(\boldsymbol{x_n})] \leq H(X) + \varepsilon$
- **Proprietà 2**: $\mathrm{P}\left(A_\varepsilon^{(n)}\right) > 1 - \varepsilon$, per n sufficientemente grande
- **Proprietà 3**: $|A_\varepsilon^{(n)}| \leq 2^{-n[H(X)+\varepsilon]}$, dove il primo elemento della disuguaglianza denota il numero di elementi dell'insieme $A_\varepsilon^{(n)}$
- **Proprietà 4**: $|A_\varepsilon^{(n)}| \geq (1 - \varepsilon)\, 2^{-n[H(X)-\varepsilon]}$, per n sufficientemente grande

Si evince in tal modo che il typical set ha quasi probabilità 1, che tutti gli elementi che lo compongono sono circa equiprobabili, e che il numero di elementi che lo compone è circa 2^{nH}.

Dimostrazioni:

La dimostrazione della prima proprietà discende direttamente dalla definizione di typical set. Dalla definizione, discende infatti che una sequenza tipica è una sequenza che soddisfa la condizione

$$2^{-n[H(X)+\varepsilon]} \leq p(\boldsymbol{x_n}) \leq 2^{-n[H(X)-\varepsilon]}$$

Prendendo il logaritmo di questa espressione si ottiene la disuguaglianza che troviamo nella prima proprietà.

Quella della seconda deriva invece proprio dalla asymptotic equipartition property. Infatti, la probabilità di una sequenza $\boldsymbol{x_n}$, appartenente al typical set, tende a 1 con $n \to \infty$. Quindi, per ogni $\delta > 0$, esiste un N, tale che per ogni $n > N$,

$$\mathrm{P}\left(\left|-\frac{1}{n}\log[p(X_1, X_2, \dots, X_n)] - H(X)\right| \leq \varepsilon\right) > 1 - \delta$$

Otteniamo la seconda parte del teorema ponendo $\delta = \varepsilon$. È importante osservare che stiamo usando la lettera ε per due scopi differenti, anziché usare le due lettere ε e δ.

Per quanto riguarda la terza proprietà, possiamo scrivere:

$$1 = \sum_{x_n \in X} p(x_n)$$

Per quanto il typical set abbia probabilità molto alta, essa resta comunque inferiore o uguale a 1

$$1 \geq \sum_{x_n \in A_\varepsilon^{(n)}} p(x_n)$$

Sappiamo però dalla definizione di typical set vista precedentemente che una sequenza tipica ha probabilità maggiore o uguale a $2^{-n[H(X)+\varepsilon]}$, per cui

$$1 \geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n[H(X)+\varepsilon]} = 2^{-n[H(X)+\varepsilon]} |A_\varepsilon^{(n)}|$$

Possiamo quindi concludere che il lower bound della cardinalità per l'insieme in questione è

$$|A_\varepsilon^{(n)}| \leq 2^{n[H(X)+\varepsilon]}$$

Infine, per *n* sufficientemente grande, $P(A_\varepsilon^{(n)}) > 1 - \varepsilon$, per cui

$$1 - \varepsilon < P(A_\varepsilon^{(n)}) \leq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n[H(X)-\varepsilon]}$$

$$= 2^{-n[H(X)-\varepsilon]} |A_\varepsilon^{(n)}|$$

dove, anche in questo caso, la seconda disuguaglianza deriva dalla definizione di typical set. Possiamo così concludere le dimostrazioni delle proprietà di quest'ultimo, trovando che l'upper bound della cardinalità del typical set è:

$$|A_\varepsilon^{(n)}| \geq (1 - \varepsilon) 2^{n[H(X)-\varepsilon]}$$

4.d – Dimensioni del typical set rispetto ad altri insiemi

Abbiamo visto la definizione di typical set e abbiamo studiato alcune delle proprietà principali di cui gode. Sappiamo inoltre che è un insieme relativamente piccolo che contiene la maggior parte della probabilità. A questo punto però è legittimo chiedersi quanto piccolo sia il typical set rispetto allo spazio che contiene tutte le sequenze possibili.

Nonostante $|A_\varepsilon^{(n)}|$ cresca esponenzialmente con *n*, rimane piuttosto piccolo rispetto X^n, l'insieme di tutte le sequenze. Per qualche $\varepsilon > 0$, abbiamo che

$$\frac{|A_\varepsilon^{(n)}|}{|X^n|} \leq \frac{2^{n[H(X)+\varepsilon]}}{2^{nlog|X|}} = 2^{-n[\log|X|-H(X)-\varepsilon]} \to 0$$

quando $n \to \infty$, dato che $H(X) < \log|X|$ (con disuguaglianza stretta, se i simboli non sono equiprobabili). Quindi l'insieme delle sequenze tipiche resta comunque molto piccolo rispetto all'insieme di tutte le sequenze.

Allo stesso modo, possiamo chiederci se esistono insiemi con proprietà simili al typical set, ma più piccoli di esso. In risposta a questa domanda proveremo che il typical set ha essenzialmente lo stesso numero di elementi del set più piccolo possibile.

Per ogni $n = 1, 2, ..., \forall\delta > 0$, chiamiamo $B_\delta^{(n)} \subset X^n$ il più piccolo insieme δ-sufficiente. Questo è il più piccolo sottoinsieme dello spazio A_X considerato, con probabilità $P(B_\delta^{(n)}) \geq 1-\delta$.

Data la similarità nelle peculiarità dei due insiemi, possiamo ipotizzare che quest'ultimo insieme debba avere un'intersezione importante con il typical set e che quindi i due insiemi debbano avere un numero simile di elementi. Introduciamo a questo punto un altro importante teorema utile al nostro scopo.

Teorema: siano $X_1, X_2, ..., X_n$ indipendenti e identicamente distribuite con probabilità *p(x)*. Per $\delta < \frac{1}{2}$, e per ogni $\delta' > 0$, se $P(B_\delta^{(n)}) > 1 - \delta$, per *n* abbastanza grande, allora

$$\frac{1}{n} log|B_\delta^{(n)}| > H - \delta'$$

Perciò $B_\delta^{(n)}$ deve contenere almeno 2^{nH} elementi. Tuttavia, dato che $2^{n[H(X)-\varepsilon]} \leq |A_\varepsilon^{(n)}| \leq 2^{n[H(X)+\varepsilon]}$, possiamo concludere che il typical set ha circa la stessa cardinalità del più piccolo insieme di alta probabilità. Matematicamente si può scrivere: $\lim_{n\to\infty} log\frac{|A_\varepsilon^{(n)}|}{|B_\delta^{(n)}|} = 0$.

4.e – Approfondimento sulla relazione tra typical set e il più piccolo insieme δ-sufficiente

Per comprendere meglio la relazione tra i due insiemi, possiamo considerare un esempio. Prendiamo in esame una sequenza di variabili aleatorie bernoulliane $X_1, X_2, ..., X_n$, con parametro $p = 0.9$. Il teorema appena visto sottintende che sia $A_\varepsilon^{(n)}$ che $B_\delta^{(n)}$ devono comprendere le sequenze che sono composte per circa il 90% di 1, e che i due insiemi devono avere cardinalità molto simile. Le sequenze tipiche sono quindi, in questo caso, sequenze nelle quali all'incirca solo un numero su dieci è uno 0, mentre i restanti nove sono 1. Tuttavia, nel

typical set non è inclusa la sequenza più probabile, ossia quella in cui ogni numero è un 1, mentre $B_\delta^{(n)}$, che include tutte le sequenze più probabili, la include.

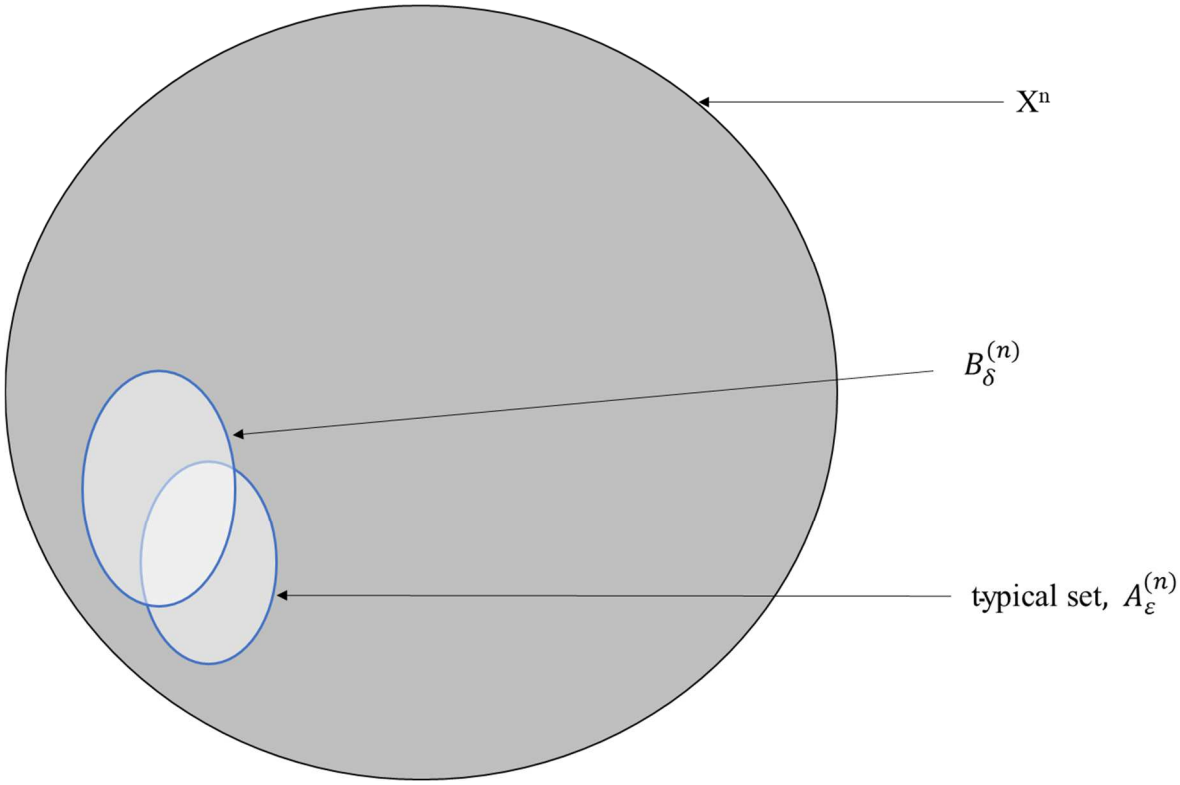


Fig. 5: Confronto visivo fra typical set, non typical set e $B_\delta^{(n)}$

4.f – Relazione tra asymptotic equipartition property e primo teorema di Shannon

Per arrivare a una conclusione, è importante chiarire che la asymptotic equipartition property è equivalente a uno dei teoremi cardine della compressione dei dati: il primo teorema di Shannon (o teorema di codifica della sorgente).

Il primo teorema di Shannon afferma in sintesi che, data una sequenza di N variabili aleatorie i.i.d., ognuna con entropia $H(X)$, questa può essere compressa in più di $NH(X)$ bit, con rischio trascurabile di perdita di informazione, con N che tende a infinito. Al contrario, se vengono compresse in meno di $NH(X)$ bit, siamo virtualmente certi che ci sarà perdita di informazione.

Vediamo invece ora il primo teorema di Shannon in maniera più formale. Sia X una variabile aleatoria con entropia $H(X) = H$ bit. Dati $\varepsilon > 0$ e $0 < \delta < 1$, esiste un N_0 tale che per ogni $N > N_0$ si ha

$$\left|\frac{1}{N} H_\delta(X^N) - H\right| < \varepsilon$$

con $H_\delta(X) = \log|B_\delta^{(n)}|$ che è l'essential bit content della variabile aleatoria *X*, mentre $B_\delta^{(n)}$ è il più piccolo insieme δ-sufficiente.

Questi due teoremi sono equivalenti in quanto possiamo definire un algoritmo di compressione in grado di dare una codifica diversa di lunghezza pari a circa $NH(X)$ bit a ognuna delle sequenze nel typical set.

Dimostrazione:

Possiamo reimmaginare il problema della compressione dati attraverso la prospettiva della Asymptotic equipartition property. Questa è fondamentale per la compressione dati perché, come abbiamo visto precedentemente, ci dà la possibilità di dividere lo spazio X^ndi tutte le sequenze di lunghezza n in due insiemi: il typical set e il non-typical set.

Vediamo ora un metodo molto semplice per comprimere le sequenze del typical set. Dato che in questo vi sono meno di $2^{n[H(X)+\varepsilon]}$ elementi, basteranno non più di $n(H+\varepsilon)+1$ bit a rappresentarle, con il +1 che è dovuto al fatto che potrebbe essere necessario un bit in più, in quanto $n(H+\varepsilon)$ potrebbe non essere un numero intero. A questo punto possiamo aggiungere un prefisso "0" a tutte le codifiche delle sequenze del typical set, generando così codifiche di lunghezza inferiore o uguale a $n(H+\varepsilon)+2$ bit per ogni sequenza tipica.
Allo stesso modo, è possibile aggiungere un "1" come prefisso a ogni codifica di sequenze non tipiche. Si riesce in questo modo a utilizzare meno di $n\log|A_X|+2$ bit per parola, dato l'alfabeto A_X. Il primo bit, in questo modo, ha la funzione di segnalare la lunghezza della codifica che lo segue.

Nello specifico, tuttavia, questa tecnica ha degli svantaggi. Infatti, in questo esempio prendiamo come dimensione del non-typical set quella di X^n. Nonostante il typical set sia molto più piccolo dello spazio di tutte le sequenze, la dimensione che abbiamo considerato è comunque più grande di quanto non sia effettivamente il non-typical set. Ciononostante, siamo comunque in grado di rappresentare in modo efficiente le sequenze degli insiemi considerati.

Il primo teorema di Shannon, come visto sopra, sostiene infatti "in sintesi" che è possibile rappresentare sequenze $\boldsymbol{x_n}$ usando in media $nH(X)$ bit, se n è sufficientemente grande.

Prendiamo ora in considerazione una sequenza di esiti $\boldsymbol{x_n}$. Sia $l(\boldsymbol{x_n})$ la lunghezza del codice binario su cui viene mappata la sequenza in questione. Siano $A_\varepsilon^{(n)}$ il typical set, e $A_\varepsilon^{(n)c}$ il suo complementare, ossia il non-typical set. Se n è grande a sufficienza, così che $P(A_\varepsilon^{(n)}) \geq 1-\varepsilon$ è rispettata, allora il valore atteso della codifica della sequenza è:

$$E[l(\boldsymbol{x_n})] = \sum_{x^n} \mathrm{p}(\boldsymbol{x_n}) l(\boldsymbol{x_n})$$

Possiamo espandere l'uguaglianza, considerando rispettivamente la probabilità che la sequenza data appartenga al typical set o al non-typical set. Da tale appartenenza ne derivano due differenti lunghezze della codifica, come vedremo nei passi successivi:

$$E[l(\boldsymbol{x_n})] = \sum_{\boldsymbol{x_n} \in A_\varepsilon^{(n)}} p(\boldsymbol{x_n}) l(\boldsymbol{x_n}) + \sum_{\boldsymbol{x_n} \in A_\varepsilon^{(n)c}} p(\boldsymbol{x_n}) l(\boldsymbol{x_n})$$

Ma a sua volta, per quanto riguarda $l(\boldsymbol{x_n})$, sappiamo che, con il nostro algoritmo, per comprimere una sequenza tipica, non servono più di $n(H+\varepsilon)+2$ bit, mentre per una non tipica non ne servono più di $n\log|A_X|+2$, includendo anche i prefissi. Pertanto:

$$E[l(\boldsymbol{x_n})] \leq \sum_{\boldsymbol{x_n} \in A_\varepsilon^{(n)}} p(\boldsymbol{x_n}) \, [n(H + \varepsilon) + 2] + \sum_{\boldsymbol{x_n} \in A_\varepsilon^{(n)c}} p(\boldsymbol{x_n}) [n log|A_X| + 2]$$

$$= P(A_\varepsilon^{(n)})[n(H + \varepsilon) + 2] + P(A_\varepsilon^{(n)c})[n log|A_X| + 2]$$

Abbiamo visto in precedenza che l'insieme delle sequenze tipiche contiene la maggior parte della probabilità. Infatti, per n che tende a infinito la probabilità del typical set tende a 1, mentre quella del suo insieme complementare a un ε piccolo a piacere, per cui

$$E[l(\boldsymbol{x_n})] \leq n(H + \varepsilon) + \varepsilon n(\log|A_X|) + 2$$

$$= n(H + \varepsilon')$$

dove $\varepsilon' = \varepsilon + \varepsilon \log|A_X| + 2/n$ può essere reso piccolo a piacere con una scelta appropriata di ε e di n. Se dividiamo il risultato trovato per il numero di elementi presenti in una sequenza, n, troviamo una quantità molto vicina all'entropia.

4.g - Conclusione

Abbiamo in tal modo dimostrato il primo teorema di Shannon grazie alla Asymptotic equipartition property, e abbiamo inoltre evidenziato come questa proprietà abbia implicazioni importantissime nella compressione dei dati, essendo la stessa la base per la dimostrazione di uno dei teoremi cardine di questa disciplina.

Nella teoria dell'informazione, il primo teorema di Shannon, e quindi anche la asymptotic equipartition property che gli è equivalente, stabilisce i limiti della compressione dei dati possibile e praticabile, nonché il significato operativo di entropia dell'informazione. Inoltre, stabilisce gli estremi inferiore e superiore della lunghezza attesa minima possibile per una codifica.

In conclusione, la teoria di Shannon ha gettato le fondamenta su cui, tra le altre cose, abbiamo costruito l'infrastruttura tecnologica alla base del mondo così come lo viviamo oggi. Viviamo in un'era data driven, in cui i dati, compressi, trasmessi e archiviati sono alla base di ogni piccola attività quotidiana che svolgiamo. Chi sa se Shannon, quando pubblicò il suo paper "A Mathematical Theory of Communication", avesse intuito la portata dei cambiamenti che sarebbero sopraggiunti.

Bibliografia

[1] S. Verdu, "Fifty years of Shannon theory," in IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2057-2078, Oct. 1998, doi: 10.1109/18.720531.

[2] David Tse, “How Claude Shannon Invented the Future”, Quanta Magazine, https://www.quantamagazine.org/how-claude-shannons-information-theory-invented-the-future-20201222/

[3] Thomas M. Cover, and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012. (Capitolo 3).

[4] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, UK, 2003 (Capitoli: Preface, I, II).

[5] C. E. Shannon, A mathematical theory of communication. The Bell System Technical Journal, 1948.

[6] Sharma Komal, and Kunal Gupta, "Lossless data compression techniques and their performance." 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017.