(F1X) Teoria dell'Informazione e della Trasmissione

10 – Proprietà di Equipartizione Asintotica

Docente: *Nicolò Cesa-Bianchi* versione 6 aprile 2016

Come nel caso della codifica sorgente, anche nel caso della codifica canale utilizziamo più simboli di ingresso per codificare un messaggio da trasmettere. Per far ciò, estendiamo il modello di canale in modo da considerare la trasmissione di sequenze di simboli.

La n-esima estensione di un canale $\langle \mathcal{X}, \mathcal{Y}, p(y \mid x) \rangle$ è il canale $\big\langle \mathcal{X}^n, \mathcal{Y}^n, p(y^n \mid x^n) \big\rangle$. La quantità $p(y^n \mid x^n)$ indica la probabilità di ottenere in uscita la sequenza $y^n \in \mathcal{Y}^n$ quando è stata trasmessa la sequenza $x^n \in \mathcal{X}^n$. Dato che il canale è senza memoria, vale

$$p(y^n \mid x^n) = p(y_n \mid y^{n-1}, x^n) \times p(y_{n-1} \mid y^{n-2}, x^n) \times \cdots \times p(y_1 \mid x^n) = \prod_{t=1}^{n} p(y_t \mid x_t) \ .$$

Un codice canale di tipo (M, n) per il canale $\langle \mathcal{X}, \mathcal{Y}, p(y \mid x) \rangle$ è definito da:

- Un insieme di M messaggi, indicati con $\{1, \ldots, M\}$.
- Una funzione di codifica $\boldsymbol{x}^n : \{1, \ldots, M\} \to \mathcal{X}^n$.
- Una funzione di decodifica $g : \mathcal{Y}^n \to \{1, \ldots, M\}$.

Dato un codice canale, la probabilità di errore di decodifica del messaggio i è definita come

$$\lambda_i = \mathbb{P}\big(g(Y^n) \neq i \mid X^n = \boldsymbol{x}^n(i)\big)$$

dove X^n, Y^n sono variabili casuali che indicano i simboli trasmessi e quelli ricevuti. Possiamo aggregare le probabilità di errore di singoli messaggi in due modi:

$$\lambda^{(n)} = \max_{i=1,\ldots,M} \lambda_i \qquad \text{probabilità massima di errore}$$

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \qquad \text{probabilità media di errore}$$

dove, ovviamente $P_e^{(n)} \leq \lambda^{(n)}$.

Il *tasso di trasmissione* di un codice di tipo (M, n) è dato dal rapporto (tutti i logaritmi in questa dispensa sono in base 2)

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

fra lunghezza in bit del messaggio e numero di simboli della sua codifica. Il tasso quindi rappresenta il numero di bit che il codice trasmette per uso del canale. Un tasso R è *raggiungibile* se esiste una sequenza di codici di tipo $\big(2^{\lceil nR \rceil}, n\big)$, per $n = 1, 2, \ldots$ tale che $\lim_{n \to \infty} \lambda^{(n)} = 0$.

Per capire i concetti di tasso e raggiungibilità, supponiamo che il canale sia binario ($\mathcal{X} = \mathcal{Y} = \{0, 1\}$) e non abbia rumore. Allora usando codifiche di n bit posso trasmettere $M = 2^n$ messaggi (tasso di trasmissione pari a uno) con probabilità di errore pari a zero. Se aggiungo rumore nel canale, allora per tenere la probabilità di errore sotto controllo devo ridurre il numero di messaggi, passando da 2^n a $2^{\lceil nR \rceil}$ per un qualche $R < 1$, che rappresenta appunto il tasso di trasmissione. Infine, come nel caso del primo teorema di Shannon, è ragionevole pensare che al crescere di n, e corrispondentemente del numero di messaggi $2^{\lceil nR \rceil}$, la probabilità di errore raggiunga asintoticamente il valore minimo a parità di tasso di trasmissione R.

Preliminarmente allo studio del secondo teorema di Shannon, introduciamo la proprietà di equipartizione asintotica (AEP). Cominciamo col rammentare la legge dei grandi numeri.

Teorema 1 *Per ogni sequenza $X_1, X_2, \ldots$ di variabili casuali indipendenti e identicamente distribuite con valore atteso $\mu < \infty$, per ogni $\varepsilon > 0$, vale*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \mu\right| > \varepsilon\right) = 0 \ .$$

In altre parole, se il campione è di taglia n sufficientemente grande rispetto ad ε, allora con alta probabilità la media campionaria è entro una distanza ε dal valore atteso.

La proprietà AEP può essere dimostrata come un corollario della legge dei grandi numeri.

Corollario 2 (AEP) *Per ogni sequenza $X_1, X_2, \ldots$ di variabili casuali indipendenti e identicamente distribuite con valore atteso finito e per ogni $\varepsilon > 0$, vale*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\log\frac{1}{p(X_1)\cdots p(X_n)} - H(X)\right| > \varepsilon\right) = 0 \ .$$

dove $H(X)$ è l'entropia della distribuzione di ciascuna delle X_t.

Dimostrazione. Introduciamo le variabili casuali $Z_1, Z_2, \ldots$ con $Z_t = -\log p(X_t)$. Dato che funzioni di variabili casuali indipendenti e identicamente distribuite sono variabili casuali ancora i.i.d., ne deduciamo che $Z_1, Z_2, \ldots$ sono anch'esse indipendenti e identicamente distribuite. Inoltre,

$$\frac{1}{n}\log\frac{1}{p(X_1)\cdots p(X_n)} = \frac{1}{n}\sum_{t=1}^{n}\big(-\log p(X_t)\big) = \frac{1}{n}\sum_{t=1}^{n} Z_t \ .$$

Dato che per ogni t vale $\mathbb{E}[Z_t] = H(X_t) = H(X)$, possiamo applicare la legge dei grandi numeri alle Z_t ottenendo la tesi. □

Si noti che nella definizione di AEP compaiono delle quantità $p(X_t)$. Queste sono variabili casuali, funzioni delle X_t, il cui valore è la probabilità calcolata in un punto casuale X_t. Un modo più intuitivo di interpretare AEP è attraverso la nozione di *insieme tipico*.

$$A_\varepsilon^{(n)} \equiv \left\{(x_1, \ldots, x_n) \in \mathcal{X}^n \ : \ 2^{-n\left(H(X)+\varepsilon\right)} \leq p(x_1)\cdots p(x_n) \leq 2^{-n\left(H(X)-\varepsilon\right)}\right\}$$

Dato che

$$\left\{(x_1,\ldots,x_n)\in\mathcal{X}^n\,:\,\left|\frac{1}{n}\log\frac{1}{p(x_1)\cdots p(x_n)}-H(X)\right|\le\varepsilon\right\}\equiv A_\varepsilon^{(n)}$$

vediamo subito che AEP implica

$$\lim_{n\to\infty}\mathbb{P}\big(A_\varepsilon^{(n)}\big)=1$$

e quindi per n abbastanza grande vale $\mathbb{P}\big(A_\varepsilon^{(n)}\big)>1-\varepsilon$. In altre parole, asintoticamente le sequenze nell'insieme tipico hanno tutte la stessa la probabilità pari a $2^{-nH(X)}$, mentre la probabilità delle altre sequenze va a zero.

Due altre proprietà notevoli dell'insieme tipico sono le seguenti.

Teorema 3 *Siano $X_1, X_2, \ldots$ variabili casuali indipendenti e identicamente distribuite e sia $A_\varepsilon^{(n)}$ l'insieme tipico corrispondente.*

1. *Per ogni n, $\big|A_\varepsilon^{(n)}\big|\le 2^{n\left(H(X)+\varepsilon\right)}$.*
2. *Per ogni n abbastanza grande vale $\big|A_\varepsilon^{(n)}\big|\ge(1-\varepsilon)2^{n\left(H(X)-\varepsilon\right)}$.*

Dimostrazione. Ricordando che $p(x^n)=p(x_1)\times\cdots\times p(x_n)$, cominciamo col dimostrare la prima proprietà. Abbiamo che

$$\begin{aligned}
1&=\sum_{x^n\in\mathcal{X}^n}p(x^n)\\
&\ge\sum_{x^n\in A_\varepsilon^{(n)}}p(x^n)\\
&\ge\sum_{x^n\in A_\varepsilon^{(n)}}2^{-n\left(H(X)+\varepsilon\right)}\qquad\text{per definizione di } A_\varepsilon^{(n)}\\
&=\big|A_\varepsilon^{(n)}\big|2^{-n\left(H(X)+\varepsilon\right)}\ .
\end{aligned}$$

Per dimostrare la seconda proprietà osserviamo che per ogni n abbastanza grande vale $\mathbb{P}\big(A_\varepsilon^{(n)}\big)>1-\varepsilon$, come avevamo già dimostrato. Quindi, per questi n abbiamo che

$$\begin{aligned}
1-\varepsilon&<\mathbb{P}\big(A_\varepsilon^{(n)}\big)\\
&=\sum_{x^n\in A_\varepsilon^{(n)}}p(x^n)\\
&\le\sum_{x^n\in A_\varepsilon^{(n)}}2^{-n\left(H(X)-\varepsilon\right)}\qquad\text{per definizione di } A_\varepsilon^{(n)}\\
&=\big|A_\varepsilon^{(n)}\big|2^{-n\left(H(X)-\varepsilon\right)}\ .
\end{aligned}$$

□

Per la dimostrazione del secondo teorema di Shannon useremo una nozione un po' più complessa di quella di insieme tipico, ovvero la nozione di insieme congiuntamente tipico. Però è possibile già sviluppare dell'intuizione sul legame fra capacità (o informazione mutua massimale) e tasso di trasmissione utilizzando gli insiemi tipici.

Fissiamo un canale $\langle \mathcal{X}, \mathcal{Y}, p(y \mid x) \rangle$ e una qualunque distribuzione $p(x)$ su $\mathcal{X}$. Abbiamo dimostrato che l'insieme tipico per $p(x)$ contiene all'incirca $2^{nH(X)}$ sequenze $x^n \in \mathcal{X}^n$. Anche l'insieme tipico per $p(y)$ contiene quindi circa $2^{nH(Y)}$ sequenze $y^n \in \mathcal{Y}^n$. Si può dimostrare che se trasmettiamo una sequenza x^n qualsiasi dall'insieme tipico, a causa del rumore osserveremo con probabilità pressoché uguale una fra circa $2^{nH(Y|X)}$ sequenze tipiche $y^n \in \mathcal{Y}^n$. Quindi, il numero massimo di x^n distinte che posso trasmettere in modo tale che gli insiemi corrispondenti in $\mathcal{Y}^n$, ciascuno di cardinalità circa $2^{nH(Y|X)}$, non si sovrappongano è

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n\left(H(Y)-H(Y|X)\right)} = 2^{nI(X,Y)} \ .$$

Posso utilizzare queste $2^{nI(X,Y)}$ sequenze per codificare altrettanti messaggi. Il tasso del codice che ottengo è

$$R = \frac{\log 2^{nI(X,Y)}}{n} = I(X,Y) \ .$$

Siano X, Y variabili casuali con valori in $\mathcal{X} \times \mathcal{Y}$ e distribuzione congiunta $p(X,Y)$. Siano $X^n = (X_1, \ldots, X_n)$ e $Y^n = (Y_1, \ldots, Y_n)$ le variabili casuali associate alle estrazioni indipendenti di n coppie $(X_1, Y_1), \ldots, (X_n, Y_n)$. Quindi,

$$\mathbb{P}\left(X^n = x^n, Y^n = y^n\right) = \prod_{i=1}^{n} p\left(x_i, y_i\right) = p(x^n, y^n) \ .$$

Per ogni $\varepsilon > 0$, l'insieme **congiuntamente tipico** $B_\varepsilon^{(n)}$ è definito come

$$B_\varepsilon^{(n)} = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ \begin{array}{l} \left| \frac{1}{n} \log \frac{1}{p(x^n)} - H(X) \right| \leq \varepsilon, \\ \left| \frac{1}{n} \log \frac{1}{p(y^n)} - H(Y) \right| \leq \varepsilon, \\ \left| \frac{1}{n} \log \frac{1}{p(x^n, y^n)} - H(X,Y) \right| \leq \varepsilon \end{array} \right\} .$$

Lemma 4 *Per ogni $\varepsilon > 0$ valgono le seguenti proprietà*

1. $\lim_{n \to \infty} \mathbb{P}\left(B_\varepsilon^{(n)}\right) = 1$.
2. *Se $\widetilde{X}^n$ ha distribuzione $\mathbb{P}(\widetilde{X}^n = x^n) = \prod_i p(x_i)$ dove $p(x)$ è la marginale di X, e $\widetilde{Y}^n$ ha distribuzione $\mathbb{P}(\widetilde{Y}^n = y^n) = \prod_i p(y_i)$ dove $p(y)$ è la marginale di Y allora, per ogni $n \geq 1$,*

$$\mathbb{P}\left(\left(\widetilde{X}^n, \widetilde{Y}^n\right) \in B_\varepsilon^{(n)}\right) \leq 2^{-n(I(X,Y)-3\varepsilon)} \ .$$

DIMOSTRAZIONE. Usando un'argomentazione del tutto simile a quella utilizzata per la dimostrazione dell'analogo enunciato relativo a $A_\varepsilon^{(n)}$, applichiamo tre volte la legge dei grandi numeri in modo da ottenere che, per ogni $\delta > 0$ fissato,

$$\exists n_1\ \forall n \geq n_1 \qquad \mathbb{P}\Bigg(\underbrace{\left|-\frac{1}{n}\log_2 p(X^n) - H(X)\right| > \varepsilon}_{[A]}\Bigg) \leq \frac{\delta}{3}$$

$$\exists n_2\ \forall n \geq n_2 \qquad \mathbb{P}\Bigg(\underbrace{\left|-\frac{1}{n}\log_2 p(Y^n) - H(Y)\right| > \varepsilon}_{[B]}\Bigg) \leq \frac{\delta}{3}$$

$$\exists n_3\ \forall n \geq n_3 \qquad \mathbb{P}\Bigg(\underbrace{\left|-\frac{1}{n}\log_2 p(X^n, Y^n) - H(X,Y)\right| > \varepsilon}_{[C]}\Bigg) \leq \frac{\delta}{3}\ .$$

Quindi, per ogni $n \geq \max\{n_1, n_2, n_3\}$, se indichiamo con $\overline{B}_\varepsilon^{(n)}$ l'insieme complemento di $B_\varepsilon^{(n)}$,

$$\mathbb{P}\big(\overline{B}_\varepsilon^{(n)}\big) = \mathbb{P}\big([A] \vee [B] \vee [C]\big) \leq \mathbb{P}([A]) + \mathbb{P}([B]) + \mathbb{P}([C]) \leq \delta\ .$$

Abbiamo appena dimostrato che

$$\lim_{n\to\infty} \mathbb{P}\big(\overline{B}_\varepsilon^{(n)}\big) = 0 \qquad \text{da cui otteniamo} \qquad \lim_{n\to\infty} \mathbb{P}\big(B_\varepsilon^{(n)}\big) = 1\ .$$

Per la seconda parte, osserviamo che

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

dove la disuguaglianza vale perché $(x^n, y^n) \in B_\varepsilon^{(n)}$ implica che $p(x^n) \leq 2^{-n\left(H(X)-\varepsilon\right)}$ e $p(y^n) \leq 2^{-n\left(H(Y)-\varepsilon\right)}$. Infine notiamo che

$$1 = \sum_{(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n} p(x^n, y^n) \geq \sum_{(x^n,y^n)\in B_\varepsilon^{(n)}} p(x^n, y^n) \geq \left|B_\varepsilon^{(n)}\right| 2^{-n\left(H(X,Y)+\varepsilon\right)}$$

in quanto $(x^n, y^n) \in B_\varepsilon^{(n)}$ implica $p(x^n, y^n) \geq 2^{-n\left(H(X,Y)+\varepsilon\right)}$. Quindi

$$\left|B_\varepsilon^{(n)}\right| \leq 2^{-n\left(-H(X,Y)-\varepsilon\right)}$$

che sostituito in (1) conclude la dimostrazione. □