(F1X) Teoria dell'Informazione e della Trasmissione

Secondo Teorema di Shannon

Docente: *Nicolò Cesa-Bianchi* versione 12 aprile 2016

Siamo finalmente pronti ad enunciare e dimostrare il teorema fondamentale della codifica di canale.

Teorema 1 (Secondo teorema di Shannon - parte I) *Sia $\langle \mathcal{X}, \mathcal{Y}, p(y \mid x) \rangle$ un canale di capacità C. Per ogni $R < C$ esiste una sequenza $K_1, K_2, \ldots$ di codici, dove il codice K_n è di tipo $\big(\lceil 2^{nR_n} \rceil, n\big)$, tale che*

$$\lim_{n\to\infty} R_n = R \qquad \text{e} \qquad \lim_{n\to\infty} \lambda^{(n)}(K_n) = 0 \ .$$

Possiamo interpretare il teorema nel modo seguente. Per codificare M messaggi (senza assumere nulla circa la loro distribuzione) ci servono $n = \lceil \log_2 M \rceil$ bit. Quindi, se il canale non avesse rumore, trasmetteremmo al tasso massimo di $R = \frac{1}{n} \log_2 M = 1$ bit per uso di canale. In altri termini, senza rumore riusciamo a trasmettere senza errori fino a 2^n messaggi diversi usando n volte il canale per ogni messaggio. Se c'è rumore, il teorema ci dice che, per ogni $\delta > 0$ e per ogni n abbastanza grande, usando n volte il canale riusciamo a trasmettere uno qualunque fra $2^{n(C-\delta)}$ messaggi con probabilità di errore che tende a zero al crescere di n.

Dimostrazione. Scegliamo un $\delta > 0$ arbitrario; fissiamo n e $M_n = \lceil 2^{nR} \rceil$ dove $R = C - \delta$. Rappresentiamo una funzione di codifica $\boldsymbol{x}^n : \{1, \ldots, M_n\} \to \mathcal{X}^n$ per un codice di tipo $\big(M_n, n\big)$ con una matrice K di M_n righe e n colonne. La i-esima riga di K è la codifica $\boldsymbol{x}^n(i) = \big(x_1^n(i), \ldots, x_n^n(i)\big)$ dell'i-esimo messaggio. Indichiamo con $K_{i,j} = x_j^n(i)$ l'elemento di riga i e colonna j della matrice K. Sia $\mathcal{K}_n$ l'insieme di tutte le matrici $M_n \times n$ a elementi in $\mathcal{X}$. Assegniamo una distribuzione di probabilità P su $\mathcal{K}_n$ nel modo seguente: sia $p(x)$ una distribuzione arbitraria su $\mathcal{X}$; allora

$$P(K) = \prod_{i=1}^{M_n} \prod_{j=1}^{n} p\left(K_{i,j}\right) \ .$$

Ovvero, per estrarre una matrice K dalla distribuzione P estraiamo in modo indipendente e identicamente distribuito gli elementi $K_{i,j}$ dalla distribuzione $p(x)$.

Consideriamo le seguenti procedure di codifica e decodifica. Mittente e ricevente condividono i seguenti parametri: il canale $\langle \mathcal{X}, \mathcal{Y}, p(y \mid x) \rangle$, il tipo (M_n, n) del codice, un reale $0 < \varepsilon < \frac{\delta}{3}$ e la distribuzione $p(x)$ utilizzata per definire la P da cui estraiamo il codice. Definiamo l'insieme $B_\varepsilon^{(n)}$ rispetto alla distribuzione congiunta $p(x^n, y^n) = \prod_i p(x_i, y_i) = \prod_i p(x_i) p(y_i \mid x_i)$.

1. Estraggo un codice $K \in \mathcal{K}_n$ con probabilità $P(K)$; sia $\boldsymbol{x}^n$ la funzione di codifica rappresentata dal codice K. Il codice estratto è noto sia al mittente che al ricevente.
2. Ottengo un messaggio $w \in \{1, \ldots, M_n\}$ da trasmettere.
3. Codifico w con $\boldsymbol{x}^n(w) = x^n$ e lo invio nel canale.

4. Il ricevente riceve y^n e lo decodifica con $\widehat{w} \in \{1, \ldots, M_n\}$ dove $\widehat{w} = i$ se e solo se

$$(\boldsymbol{x}^n(i), y^n) \in B_\varepsilon^{(n)} \qquad \text{e} \qquad (\boldsymbol{x}^n(j), y^n) \notin B_\varepsilon^{(n)} \qquad \text{per ogni } j \neq i;$$

se non esiste nessun i con queste proprietà allora $\widehat{w}$ è scelto arbitrariamente in $\{1, \ldots, M_n\}$.

Si noti che questa procedura usa un codice casuale per codificare un messaggio. Inizialmente, dimostreremo il teorema in media rispetto all'estrazione del codice e rispetto alla probabilità media di errore $P_e^{(n)}$. In seguito, dimostreremo che il teorema vale per un codice specifico e per la probabilità massima di errore $\lambda^{(n)}$.

Denotiamo con W e $\widehat{W}$ le variabili casuali associate a w e $\widehat{w}$, dove $\mathbb{P}(W = i) = 1/M_n$ in quanto consideriamo la probabilità media di errore. Siano inoltre $X^n(i)$ la variabile casuale che rappresenta la codifica di i con un codice casuale K e $Y^n(i)$ la variabile casuale che rappresenta la sequenza di simboli d'uscita ottenuta inviando $X^n(i)$ nel canale.

Ora, per come è stata costruito il codice casuale, la distribuzione di $X^n(i)$ non dipende da i; infatti, ogni riga della matrice K è estratta in modo indipendente. Inoltre, per definizione di canale, la distribuzione dell'uscita $Y^n(i)$ è anch'essa indipendente dal messaggio i. Da ciò deduciamo che le coppie $\big(X^n(i), Y^n(i)\big)$ sono distribuite come $p(x^n, y^n) = \prod_i p(x_i, y_i)$ che è la stessa distribuzione congiunta usata per definire $B_\varepsilon^{(n)}$. Studiamo ora la probabilità media d'errore della procedura descritta sopra. Un errore si verifica quando $\widehat{W} \neq W$. Scriviamo

$$P_e^{(n)} = \frac{1}{M_n} \sum_{i=1}^{M_n} \mathbb{P}\left(\widehat{W} \neq W \mid W = i\right)$$

e maggioriamo separatamente ciascun $\mathbb{P}(\widehat{W} \neq W \mid W = i)$.

Denotiamo con E_j l'evento $(X^n(j), Y^n(i)) \in B_\varepsilon^{(n)}$. Per definizione della procedura di decodifica, notiamo che avviene un errore di decodifica sul messaggio i quando $X^n(i)$ non è congiuntamente tipico con $Y^n(i)$, oppure quando esistono dei $X^n(j)$ che sono congiuntamente tipici con $Y^n(i)$. Formalmente, dato $W = i$,

$$\widehat{W} \neq i \quad \text{implica} \quad \overline{E_i} \quad \text{oppure} \quad \bigvee_{j \neq i} E_j \ .$$

Allora,

$$\mathbb{P}(\widehat{W} \neq i \mid W = i) \leq \mathbb{P}(\overline{E}_i) + \sum_{j \neq i} \mathbb{P}(E_j) = 1 - \mathbb{P}(E_i) + \sum_{j \neq i} \mathbb{P}(E_j) \ .$$

Ricordando che la distribuzione congiunta di $\big(X^n(i), Y^n(i)\big)$ è la stessa $p\big(x^n, y^n\big)$ usata per definire $B_\varepsilon^{(n)}$, possiamo scrivere

$$\mathbb{P}\Big(\big(X^n(i), Y^n(i)\big) \in B_\varepsilon^{(n)}\Big) = \sum_{(x^n, y^n) \in B_\varepsilon^{(n)}} \mathbb{P}\big(X^n(i) = x^n, Y^n(i) = y^n\big) = \sum_{(x^n, y^n) \in B_\varepsilon^{(n)}} p(x^n, y^n) = \mathbb{P}\big(B_\varepsilon^{(n)}\big) \ .$$

Quindi vale

$$\lim_{n \to \infty} \mathbb{P}\Big(\big(X^n(i), Y^n(i)\big) \in B_\varepsilon^{(n)}\Big) = \lim_{n \to \infty} \mathbb{P}\big(B_\varepsilon^{(n)}\big) = 1 \ .$$

Ovvero, per ogni $\varepsilon > 0$ esiste un n_ε tale che per ogni $n \geq n_\varepsilon$ vale

$$\mathbb{P}(E_i) = \mathbb{P}\Big(\big(X^n(i), Y^n(i)\big) \in B_\varepsilon^{(n)}\Big) \geq 1 - \varepsilon \ .$$

Quindi, per ogni n sufficientemente grande rispetto a ε,

$$\mathbb{P}(\widehat{W} \neq i \mid W = i) \leq \varepsilon + \sum_{j \neq i} \mathbb{P}(E_j) \ .$$

Ora studiamo E_j per $j \neq i$. Per costruzione del codice casuale, $X^n(i)$ e $X^n(j)$ sono indipendenti e identicamente distribuiti in quanto corrispondono a righe diverse della matrice casuale K. Inoltre, dato che il rumore nel canale non dipende dal messaggio w codificato, anche $X^n(j)$ e $Y^n(i)$ sono indipendenti. Ma allora, $(X^n(j), Y^n(i))$ è una coppia di variabili casuali indipendenti tali che $\mathbb{P}\left(X^n(i) = x^n\right) = p(x^n)$ e $\mathbb{P}\left(Y^n(j) = y^n\right) = p(y^n)$. Quindi, per la proprietà (2) degli insiemi congiuntamente tipici,

$$\mathbb{P}(E_j) \leq 2^{-n(I(X,Y)-3\varepsilon)}$$

dove X è la variabile casuale associata all'estrazione di un simbolo d'ingresso $X_k^n(i)$ nella costruzione del codice casuale e Y è la variabile casuale associata alla ricezione di un simbolo d'uscita $Y_k^n(i)$ (dato che, per definizione del codice casuale, i simboli di ingresso $X_1^n(i), \ldots, X_n^n(i)$ sono indipendenti e identicamente distribuiti, e dato che il canale è senza memoria, anche i simboli d'uscita $Y_1^n(i), \ldots, Y_n^n(i)$ sono indipendenti e identicamente distribuiti). Quindi,

$$\mathbb{P}(\widehat{W} \neq i \mid W = i) \leq \varepsilon + \sum_{j \neq i} 2^{-n(I(X,Y)-3\varepsilon)} \leq \varepsilon + M_n \, 2^{-n(I(X,Y)-3\varepsilon)} = \varepsilon + \left\lceil 2^{nR} \right\rceil 2^{-n(I(X,Y)-3\varepsilon)} \ .$$

Scegliamo ora $p(x)$ in modo che $I(X,Y) = C$. Dato che $R = C - \delta$ e $\left\lceil 2^{nR} \right\rceil \leq 2^{nR} + 1$,

$$\mathbb{P}(\widehat{W} \neq i \mid W = i) \leq \varepsilon + \big(2^{nR} + 1\big) 2^{-n(C-3\varepsilon)} = \varepsilon + 2^{-n(\delta-3\varepsilon)} + 2^{-n(C-3\varepsilon)} \ .$$

Dato che $C > \delta > 3\varepsilon$, per C e δ fissati e per ogni $\varepsilon > 0$ esiste un n_ε tale che per ogni $n \geq n_\varepsilon$ abbiamo che $2^{-n(\delta-3\varepsilon)} + 2^{-n(C-3\varepsilon)} \leq \varepsilon$. Abbiamo quindi dimostrato che per questi n vale che

$$\mathbb{P}\left(\widehat{W} \neq W \mid W = i\right) = \frac{1}{M_n} \sum_{i=1}^{M_n} \mathbb{P}\left(\widehat{W} \neq W \mid W = i\right) \leq \frac{1}{M_n} \sum_{i=1}^{M_n} 2\varepsilon = 2\varepsilon \ .$$

Per terminare la dimostrazione, dobbiamo dimostrare che esiste un codice specifico K per il quale vale la disuguaglianza sopra. A questo scopo, sia $\mathbb{P}(\widehat{W} \neq W \mid K)$ la probabilità di errore di decodifica quando è stato usato il codice K. Si osservi che

$$\min_{K \in \mathcal{K}_n} \mathbb{P}\left(\widehat{W} \neq W \mid K\right) \leq \sum_K \mathbb{P}\left(\widehat{W} \neq W \mid K\right) P(K) = \mathbb{P}\left(\widehat{W} \neq W\right) \leq 2\varepsilon \ .$$

Allora

$$K_n^* = \operatorname*{argmin}_{K \in \mathcal{K}_n} \mathbb{P}\left(\widehat{W} \neq W \mid K\right)$$

soddisfa $\mathbb{P}(\widehat{W} \neq W \mid K_n^*) \leq 2\varepsilon$.

Ora, dato che l'enunciato del teorema richiede che $\lambda^{(n)}(K_n) = 0$ per $n \to \infty$, dobbiamo mettere in relazione il maggiorante su $\mathbb{P}(\widehat{W} \neq W \mid K_n^*)$ con $\lambda^{(n)}$. Abbiamo

$$\frac{1}{M_n} \sum_{i=1}^{M_n} \mathbb{P}\left(\widehat{W} \neq i \mid W = i,\, K_n^*\right) = \mathbb{P}\left(\widehat{W} \neq W \mid K_n^*\right) \leq 2\varepsilon \ .$$

Rinumeriamo i messaggi in modo che

$$\mathbb{P}\left(\widehat{W} \neq 1 \mid W = 1,\, K_n^*\right) \leq \cdots \leq \mathbb{P}\left(\widehat{W} \neq M_n \mid W = M_n,\, K_n^*\right) \ .$$

Sia I l'insieme dei primi $\lceil M_n/2 \rceil$ indici e J l'insieme degli ultimi $\lfloor M_n/2 \rfloor$ indici. Allora, dato che

$$\sum_{i \in I} \mathbb{P}\left(\widehat{W} \neq i \mid W = i,\, K_n^*\right) + \sum_{j \in J} \mathbb{P}\left(\widehat{W} \neq j \mid W = j,\, K_n^*\right) \leq 2M_n\varepsilon \tag{1}$$

dev'essere $\max\limits_{i \in I} \mathbb{P}\left(\widehat{W} \neq i \mid W = i,\, K_n^*\right) \leq 4\varepsilon$, altrimenti avremmo che per almeno $\lceil M/2 \rceil$ messaggi i —ovvero almeno uno in I e tutti quelli in J— vale che $\mathbb{P}(\widehat{W} \neq i \mid W = i,\, K_n^*) > 4\varepsilon$, il che implicherebbe

$$\sum_{i=1}^{M_n} \mathbb{P}\left(\widehat{W} \neq i \mid W = i,\, K_n^*\right) > \frac{M_n}{2} 4\varepsilon = 2M_n\varepsilon$$

contraddicendo la (1). Sia quindi K_n' il codice di tipo $\left(\lceil 2^{nR-1} \rceil, n\right)$ ottenuto considerando le prime $\lceil M/2 \rceil$ parole di codice di K_n^*. Per costruzione, questo codice ha tasso $R_n' \geq R - \frac{1}{n}$ e soddisfa $\lambda^{(n)}(K_n') \leq 4\varepsilon$. Si consideri ora la sequenza di codici $K_1', K_2', \ldots$ costruiti come descritto sopra. Tale sequenza ha tasso asintotico $\lim_{n\to\infty} R_n' \geq \lim_{n\to\infty}\left(R - \frac{1}{n}\right) = R$. Inoltre, per ogni ε piccolo a piacere trovo n_ε tale che per ogni $n \geq n_\varepsilon$ vale $\lambda^{(n)}(K_n') \leq 4\varepsilon$. Per definizione di limite, questo è equivalente a $\lim_{n\to\infty} \lambda^{(n)}(K_n') = 0$. □