Complessità e Teoria dell'Informazione

Prof.ssa Carla Piazza

Appunti di Claudio Desideri e Lara Vignotto – a.a. 2023/2024

Indice

I Teoria dell'Informazione **7**

1 Introduzione **9**
1.1 Probabilità, Entropia e Inferenza . . . 9
1.1.1 Teorema di Bayes . . . 9
1.1.2 Proprietà dell'Entropia . . . 9
1.1.2.1 Scomponibilità dell'Entropia . . . 11
1.1.3 Inferenza . . . 11

2 Compressione **13**
2.1 Il Teorema Della Codifica Sorgente . . . 13
2.2 Codici Simbolo . . . 13
2.2.1 Limite imposto dalla Decodificabilità Univoca . . . 16
2.2.2 Compressione Massima . . . 19
2.2.3 Shannon Code . . . 20
2.3 Codici Stream . . . 22
2.3.1 Un po' di Storia . . . 22
2.3.2 Lempel-Ziv . . . 22
2.3.2.1 LZ77 . . . 23

3 Codifica di Canale Rumoroso **25**
3.1 Variabili Aleatorie Dipendenti . . . 25
3.1.1 Divergenza e Disuguaglianza di Gibbs . . . 25
3.1.2 Entropia e Mutual Information . . . 26
3.1.2.1 Mutual Information . . . 27

4 Kolmogorov Complexity **29**
4.1 Nozioni Preliminari . . . 29
4.1.1 Macchine di Turing . . . 29
4.2 Complessità di Kolmogorov . . . 30
4.2.1 Complessità di Kolmogorov vs Entropia di Shannon . . . 34

II Complessità **35**

5 Introduzione **37**
5.1 Tesi di Church-Turing Estesa . . . 38

6 Macchine di Turing **39**
6.1 Definizioni . . . 39
6.2 Unlimited Register Machines . . . 40
6.2.1 URM + Prodotto . . . 41
6.3 Ulteriori Definizioni . . . 42
6.4 Macchine di Turing a k-nastri e Input/Output . . . 43
6.4.1 Complessità Spaziale . . . 44
6.5 Random Access Machines . . . 46
6.6 Macchine Nondeterministiche . . . 47

Introduction

Programma:

- Teoria dell'Informazione
- Teoria della Complessità
- Algoritmi su Grafi ecc ...

Parte I

Teoria dell'Informazione

Capitolo 1

Introduzione

1.1 Probabilità, Entropia e Inferenza

Claude Shannon, 1948, *A Mathematical Theory of Communication*.

Un messaggio è una sequenza di lettere (simboli) da un alfabeto. Qual è l'informazione in una frase (messaggio)? Come possiamo misurare la quantità di informazione? Dipende dal contesto.

Esempio Il messaggio è "Piove!". Qual è la quantità di informazione? Per il signor Muller, che vive a Vienna, dove piove spesso, la quantità di informazione è bassa. Per Fatima, che vive nel deserto, invece, è alta.

Si ha quindi che

- Bassa probabilità di un evento

...

1.1.1 Teorema di Bayes

Sappiamo che nel caso di due **eventi indipendenti**, la probabilità congiunta è

$$p(a_i, b_j) = p(a_i) \cdot p(b_j)$$

Nel caso, invece, di due **eventi dipendenti** si ha

$$p(a_i, b_j) = p(a_i|b_j) \cdot p(b_j) = p(b_j|a_i) \cdot p(a_i)$$

e quindi

$$p(a_i|b_j) = \frac{p(b_j|a_i) \cdot p(a_i)}{p(b_j)}$$

con $p(b_j)$ fattore di normalizzazione.

TODO: vedere al capitolo 2 del libro il prior, poterior, likelihood, ecc.

1.1.2 Proprietà dell'Entropia

Libro, pag. 33. Le proprietà della funzione di entropia sono:

- $H(P) \geq 0$ con uguaglianza iff $p_i = 1$ per un qualche i. In particolare, $H(P) = 0$ iff $\exists i \; p_i = 1$, ovvero quando c'è un evento certo l'entropia è nulla.
- L'entropia è massimizzata se la distribuzione p è uniforme.

Analizziamo meglio e dimostriamo la seconda proprietà.

Proprietà 1.1.1 (Entropia massima)

$$\mathcal{H}(P) \leq \log_2 |P|$$

con $|P|$ numero di eventi ($|X|$), e

$$\mathcal{H}\left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right) = \log_2 k$$

dove k è il numero di eventi, ovvero $|P|$.

Dimostrazione La dimostrazione è basata su proprietà di funzioni complesse, in particolare sulla diseguaglianza di Jensen, la quale è una disuguaglianza che lega il valore di una funzione convessa al valore della medesima funzione calcolata nel valor medio del suo argomento.

Sia $\boldsymbol{f(x) = -x \log_2 x}$ (cfr. definizione di entropia). Vogliamo controllare se è concava o convessa, calcoliamo quindi la sua derivata seconda:

$$f''(x) = -\frac{1}{x}$$

Sappiamo che $0 \leq x \leq 1$ perché è una probabilità, e quindi abbiamo che $-1/x < 0$: la funzione è concava. Prendiamo ora due punti x_1 e x_2, e un punto x tra i due.

TODO: disegno

Abbiamo che la media pesata di x_1 e x_2 è

$$x = \lambda x_1 + (1 - \lambda) x_2 \qquad \text{con} \quad 0 \leq \lambda \leq 1$$

con λ il peso. In particolare, se $\lambda = 1$ allora $x = x_1$, se $\lambda = 0$ allora $x = x_2$, e se $\lambda = 1/2$ allora x si troverà esattamente a metà tra x_1 e x_2. Se prendiamo la combinazione lineare di $f(x_1)$ e $f(x_2)$, otteniamo la seguente disuguaglianza:

$$\lambda f(x_1) + (1 - \lambda) f(x_2) \leq f(\lambda x_1 + (1 - \lambda) x_2)$$

Tale disuguaglianza si può generalizzare alla combinazione lineare di un qualsiasi numero di punti. Se $f''(x) \leq 0$, $\forall x \in [x_1, x_n]$ (ovvero $f''(x)$ è concava) si ha la disuguaglianza di Jensen:

$$\sum_{i=1}^{n} \lambda_i f(x_i) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i\right)$$

con $\lambda_i \geq 0$ e $\sum_{i=1}^{n} \lambda_i = 1$. La disuguaglianza cambia verso ($\geq$) quando la funzione è convessa, cioè $f''(x) \geq 0$. Ricordiamo che vogliamo arrivare alla combinazione lineare dove i punti hanno la stessa probabilità. Quindi con

$$\lambda_i = \frac{1}{k} \qquad |P| = |X| = k \qquad P = \{p_1, \ldots, p_k\}$$

scriviamo la disuguaglianza di Jensen come

$$\sum_{i=1}^{k} \underbrace{\frac{1}{k}}_{\lambda_i} \underbrace{(-p_i \log_2 p_i)}_{f(x_i)} \leq \underbrace{-\left(\sum_{i=1}^{k} \frac{1}{k} p_i\right) \log_2 \left(\sum_{i=1}^{k} \frac{1}{k} p_i\right)}_{f(\sum \lambda_i x_i)}$$

semplifichiamo perché $k > 0$

$$-\frac{1}{k} \sum_{i=1}^{k} p_i \log_2 p_i \leq -\frac{1}{k} \log_2 \frac{1}{k}$$

$$\mathcal{H}(P) \leq \log_2 k = \mathcal{H}\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$$

Ovvero l'entropia è massima per la distribuzione uniforme. □

Proprietà 1.1.2 (Entropia Congiunta) *Siano P, Q due distribuzioni, e x_i, y_j coppia di eventi tali che $x_i \in P$ e $y_j \in Q$. L'entropia congiunta di P, Q è:*

$$\mathcal{H}(P,Q) = -\sum_{i,j} p(x_i, y_j) \log_2(p(x_i, y_j))$$

Se x e y sono indipendenti (quindi la probabilità congiunta è il prodotto delle due probabilità) l'entropia è additiva:

$$\mathcal{H}(P,Q) = \mathcal{H}(P) + \mathcal{H}(Q)$$

La somma è possibile perché usiamo i logaritmi, e una delle loro proprietà è $\log(a \cdot b) = \log(a) + \log(b)$.

1.1.2.1 Scomponibilità dell'Entropia

Libro, pag. 33. Sia P una distribuzione (vettore) di probabilità, e X delle variabili.

$$\begin{aligned} P &= \{p_1,\ p_2,\ \ldots,\ p_n\} \\ X &= \{\underbrace{x_1}_{p_1}, \underbrace{x_2, \ldots, x_n}_{1-p_1}\} \end{aligned}$$

In questo contesto, la probabilità, ad esempio, del secondo evento x_2 è, normalizzata, pari a $p_2/1-p_1$, e quella dell'ultimo elemento x_n è $p_n/1-p_1$.

Esempio Abbiamo una moneta regolare. Al primo lancio esce H, e come risultati desiderati per il secondo e terzo lancio vogliamo T e T. Abbiamo quindi:

$$\underbrace{H}_{\substack{p_1=\frac{1}{2} \\ \frac{1}{2}}} \quad \underbrace{T \quad T}_{\substack{1-p_1=\frac{1}{2} \\ \frac{1}{4} \quad \frac{1}{4}}}$$

La quantità di informazione ricevuta da P è uguale a quella ricevuta dal processo in due passaggi.

$$\begin{aligned} \mathcal{H}(P) &= \sum p_i \log_2 \frac{1}{p_i} \\ &= \underbrace{\mathcal{H}(p_1, 1-p_1) + (1-p_1) \cdot \mathcal{H}\underbrace{\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_n}{1-p_1}\right)}_{\substack{p_i \text{ normalizzati la cui somma è 1} \\ 1-p_1 = p_2+p_3+\cdots+p_n}}}_{\text{si possono dividere in diversi punti, ottenendo, ad esempio, tante entropie}} \end{aligned}$$

TODO: vdere proprietà libro pag 33 sez 2.6?

1.1.3 Inferenza

TODO: capitolo 3, esercizio 3.8 pag 57

Capitolo 2

Compressione

TODO: capitolo 4, esercizio 4.1 pag 66

2.1 Il Teorema Della Codifica Sorgente

Capitolo 4 del libro di MacKay. In questo capitolo discuteremo come misurare il contenuto informativo del risultato di un esperimento aleatorio.

Studiamo $\mathcal{H}(\{p, 1-p\})$, con $0 \leq p \leq 1$

$$\begin{aligned} \mathcal{H}(\{p, 1-p\}) &= \\ &= p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \\ &= -p \log(p) - (1-p) \log(1-p) \\ &= \mathcal{H}(p) \end{aligned}$$

TODO: finire
LEZ 3
TODO: soluzione esercizio 4.1
TODO: muddy children puzzle

2.2 Codici Simbolo

Capitolo 5 del libro di MacKay. Nell'ultimo capitolo abbiamo visto una prova dello status fondamentale dell'entropia come misura del contenuto informativo medio. Abbiamo definito uno schema di compressione dei dati utilizzando codici a blocchi di lunghezza fissa. Abbiamo così verificato la possibilità di compressione dei dati, ma la codifica a blocchi definita nella dimostrazione non ha fornito un algoritmo pratico. In questo capitolo studieremo algoritmi pratici di compressione dei dati.

Immagina un guanto di gomma pieno d'acqua. Se comprimiamo due dita del guanto, qualche altra parte del guanto deve espandersi, perché il volume totale dell'acqua è costante (l'acqua è essenzialmente incomprimibile). Allo stesso modo, quando accorciamo le parole in codice per alcuni risultati, ci devono essere altre parole in codice che si allungano, se lo schema non è *lossy*. In questo capitolo scopriremo l'equivalente teorico dell'informazione del volume dell'acqua.

Definizione 2.2.1 (Alfabeti di input/output)

$$\begin{aligned} \textit{Alfabeto di input} \quad & \mathcal{A} = \{a_1, a_2, \ldots, a_k\} \\ \textit{Alfabeto di output} \quad & \mathcal{B} = \{b_1, b_2, \ldots, b_D\} \end{aligned}$$

Definizione 2.2.2 (Codice) *Sia $\mathcal{A}^*$ un messaggio (sequenza di caratteri) sull'alfabeto $\mathcal{A}$. Il* ***codice*** *c è*

$$c : \mathcal{A}^* \to \mathcal{B}^*$$

ovvero un messaggio dall'alfabeto $\mathcal{A}$ all'alfabeto $\mathcal{B}$ (iniettiva).

Con $\mathcal{A}^* = \bigcup_{n\in\mathbb{N}} A^n$, ovvero l'insieme di tutte le possibili stringhe che si possono creare utilizzando l'alfabeto $\mathcal{A}$, compresa la stringa vuota.

Si vuole comprimere il messaggio in modo da ottenere il messaggio più corto possibile. Per farlo, utilizziamo una codifica.

Definizione 2.2.3 (Codifica) *Una codifica è una funzione*

$$\varphi : \mathcal{A} \to \mathcal{B}^*$$

Inoltre

$$\varphi(\underbrace{x_1, x_2, \dots, x_m}_{\in\mathcal{A}^*}) = \varphi(x_1)\varphi(x_2)\dots\varphi(x_m)$$

Esempio 1 Alfabeti: $\mathcal{A} = \{a, b, c\}$, $\mathcal{B} = \{0, 1\}$; codifica: $\varphi(a) = 0$, $\varphi(b) = 10$, $\varphi(c) = 01$. È una buona codifica? No, perché è ambigua. Ad esempio

$$\varphi(ab) = 010 = \varphi(ca)$$

È iniettiva nella codifica ma non sul messaggio. Una codifica di questo tipo viene detta **not uniquely decodable** (non univocamente decodificabile).

Definizione 2.2.4 (Univocamente decodificabile) *Un codice $\varphi : \mathcal{A} \to \mathcal{B}^*$ è univocamente decodificabile (uniquely decodable) se*

$$\forall m_1, m_2 \in \mathcal{A}^* \qquad \varphi(m_1) \neq \varphi(m_2)$$

In questo corso non utilizzeremo codici non univocamente decodificabili.

Esempio 2 Alfabeti: $\mathcal{A} = \{a, b, c\}$, $\mathcal{B} = \{0, 1\}$; codifica: $\varphi(a) = 0$, $\varphi(b) = 01$, $\varphi(c) = 011$. Ad esempio, il messaggio *aabcbbca* viene codificato come $\varphi(aabcbbca) = 000101101010110$. È univocamente decodificabile (UD)?

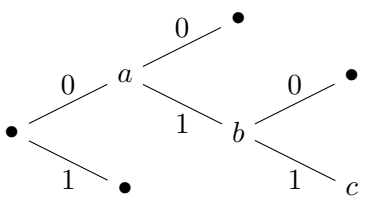
Sì, è UD con delay 1. Ad ogni 0, si controlla il carattere successivo: se è un altro 0, la lettera è a, altrimenti si prosegue fino al primo 1 per decidere se è b o c. Il delay 1 è riferito allo zero che si incontra, che significa l'inizio di un'altra lettera.

Esempio 3 Alfabeti: $\mathcal{A} = \{a, b, c\}$, $\mathcal{B} = \{0, 1\}$; codifica: $\varphi(a) = 00$, $\varphi(b) = 1$, $\varphi(c) = 10$. Ad esempio, il messaggio *bcbaca* viene codificato come $\varphi(bcbaca) = 1101001000$. È UD?

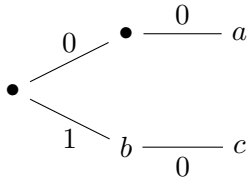
Sì, si controlla se c'è un numero pari o dispari di 0 dopo un 1: se è pari, si tratta di una b seguita da una o più a, se è dispari di una c, eventualmente seguita da una o più a. È UD con unbounded delay.

Abbiamo visto che nel caso di distribuzione uniforme, la quantità di informazione è pari all'entropia. Quanto è complesso computare una codifica/decodifica?

Poiché l'alfabeto di input è binario, per rappresentare una codifica si può utilizzare un albero binario. Quello per l'**Esempio 2** è il seguente:



Quello per l'**Esempio 3** è il seguente: TODO: fixare in modo che gli archi ad a e c siano inclinati



Quando si finisce in un nodo con un'etichetta ci si deve chiedere se la conclusione è che ci si può fermare. Il grado (fattore) di diramazione è pari alla cardinalità dell'alfabeto di output. Inoltre, se l'albero ha altezza h, tutte le codifiche hanno lunghezza h.

Ci chiediamo, qual è una codifica sicuramente UD e senza delay?

Definizione 2.2.5 (Codice prefisso)

$$\varphi : \mathcal{A} \to \mathcal{B}^* \text{ è un codice prefisso}$$
$$\Updownarrow$$
$$\forall a_i, a_j \in \mathcal{A} \quad \varphi(a_i) \in \mathcal{B}^* \text{ non è un prefisso di } \varphi(a_j) \in \mathcal{B}^*$$

Al posto di codice prefisso (prefix code) utilizzeremo il termine prefisso (prefix). è corretto?

Esempio 3 (vedi sopra) φ non è un prefisso, perché la codifica di b è 1, che è un prefisso della codifica di c, ovvero 10.

Lemma 2.2.1 *φ è un codice prefisso $\Rightarrow$ φ è UD senza delay*

Per memorizzare l'albero si utilizza CONSTANT SPACE, che è un sottoinsieme di LINEAR TIME. Ciò equivale a dire che è possibile computarlo con un automa.

Esempio 3 (vedi sopra) $\varphi(bbb) = 111$: il messaggio di input ha lunghezza 3, e viene codificato in un messaggio di lunghezza uguale (3).

$\varphi(aca) = 001000$: il messaggio di input ha lunghezza 3, e viene codificato in un messaggio di lunghezza 6.

Il numero di possibili messaggi di lunghezza 3 in output è $2^3 = 8$.

Definizione 2.2.6 (Lunghezza media di una codifica, EL (Expected Length)) *Siano $\varphi : \mathcal{A} \to \mathcal{B}^*$ una codifica, e P una distribuzione di probabilità sull'alfabeto di input $\mathcal{A}$*

$$EL(\varphi) = \sum_{i=1}^{n} p(a_i) \cdot |\varphi(a_i)|$$

con $|\varphi(a_i)| = l_i$ lunghezza della codifica di a_i.

Esempio Supponiamo che nell'**Esempio 3** le probabilità siano $p(a) = 1/2$, $p(b) = 1/4$, $p(c) = 1/4$. La lunghezza media è

$$EL(\varphi) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 2 = 1 + \frac{3}{4}$$

Immaginiamo una codifica φ^* diversa, per la quale $|\varphi^*(a)| = 1$, $|\varphi^*(b)| = 2$, $|\varphi^*(c)| = 2$. La lunghezza media è

$$EL(\varphi^*) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}$$

Abbiamo che $EL(\varphi) > EL(\varphi^*)$, quindi φ^* è migliore di φ.

Si vuole trovare la codifica con la minor EL sotto l'assunzione che la sorgente del messaggio non abbia memoria. Dati $\mathcal{A}$, $\mathcal{B}^*$, P, si vuole trovare la miglior codifica φ. Non è sufficiente considerare solo codici prefix-free.

Possiamo raggiungere il minimo di EL considerando i codici prefisso. EL è codificata da $\mathcal{H}(P)$. I codici possono essere asintotticamente ottimali, o ottimali.

2.2.1 Limite imposto dalla Decodificabilità Univoca

È possibile definire un codice $\varphi : \mathcal{A} \to \mathcal{B}^*$ che sia UD, date le lunghezze delle codifiche $l_1, \ldots, l_k$? Definiamo della terminologia:

$$k = |\mathcal{A}| \qquad D = |\mathcal{B}| (= 2 \text{ nel libro})$$

Teorema 2.2.2 (Disuguaglianza di Kraft-McMillan, o Teorema Inverso)

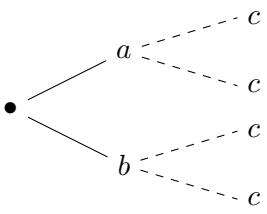
$$\varphi \text{ è } UD \quad \Rightarrow \quad \sum_{i=1}^{k} \frac{1}{D^{l_i}} \leq 1 \; = \; \sum_{i=1}^{k} D^{-l_i} \leq 1$$

Se > 1 non esiste un codice UD con tale lunghezza.

Esempio $\mathcal{A} = \{a, b, c\}$, $\mathcal{B} = \{0, 1\}$, $l_1 = 1$, $l_2 = 1$, $l_3 = 2$ (lunghezze delle codifiche di a, b, c). Applicando il teorema si ottiene

$$\frac{1}{2^1} + \frac{1}{2^1} + \frac{1}{2^2} > 1$$

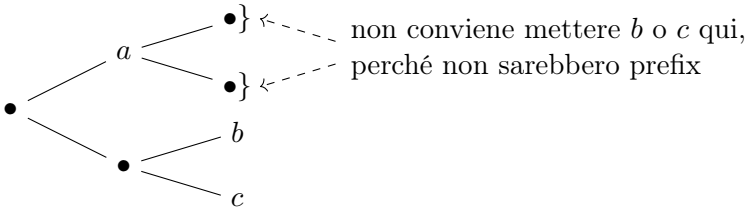
Infatti, non importa dove si sceglie di codificare la c, la codifica non è UD.



Esempio $\mathcal{A} = \{a, b, c\}$, $\mathcal{B} = \{0, 1\}$, $l_1 = 1$, $l_2 = 2$, $l_3 = 2$. Applicando il teorema si ottiene

$$\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^2} = 1$$

è UD (cfr. anche Teorema Diretto).



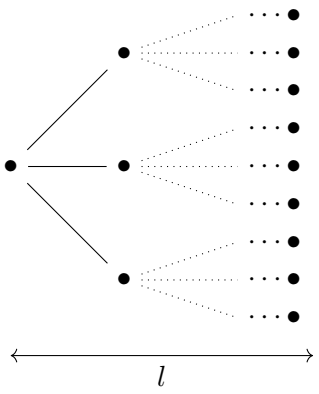
Teorema 2.2.3 (Teorema Diretto)

$$\sum_{i=1}^{k} D^{-l_i} \leq 1 \quad \Rightarrow \quad \exists \varphi \textit{ prefisso con lunghezze } l_1, \ldots, l_k$$

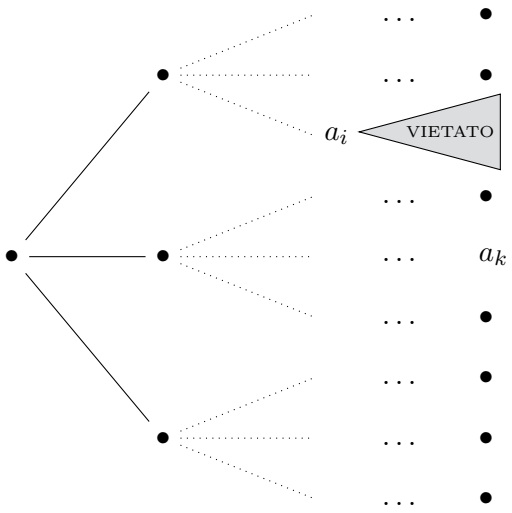
Prefisso è più forte di UD.

I due risultati (teoremi) affermano che, anche se ci limitiamo all'utilizzo di codici prefisso, non perdiamo alcuna potenza nella compressione. È sufficiente l'utilizzo dei codici prefisso, comprimono a sufficienza.

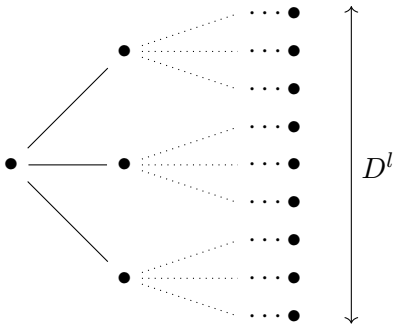
Dimostrazione Teorema Inverso, caso prefisso + Teorema Diretto Sia φ un codice prefisso e $l_1 \leq l_2 \leq \cdots \leq l_k = l$. Consideriamo l'albero D-ario (ogni nodo ha D figli) che rappresenta φ, di altezza l.



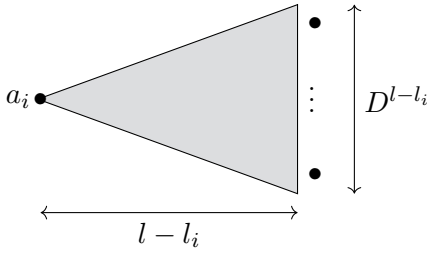
La codifica più lunga a_k è in una delle foglie. Supponiamo che la codifica a_i sia in uno dei nodi interni. Nessuno dei nodi interni (e in particolare nessuna delle foglie) del sottoalbero con a_i come radice può essere utilizzato per un'altra codifica.



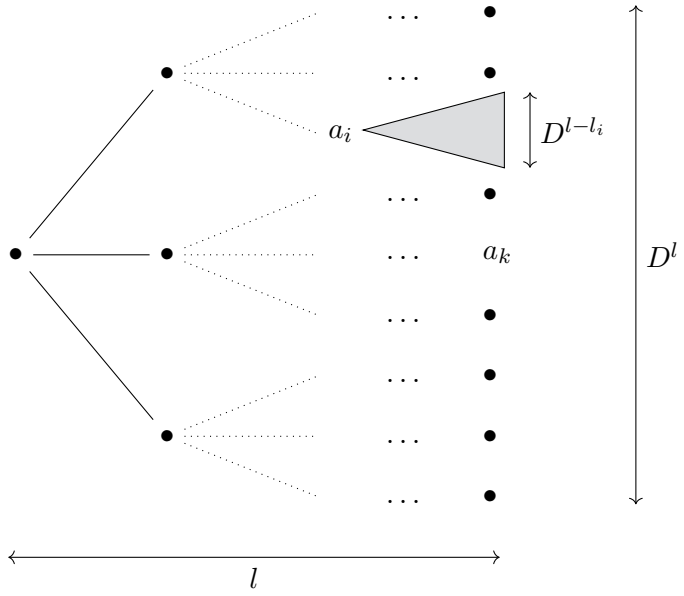
In un tale albero, il numero di foglie è pari a D^l.



Il numero di foglie di un sottoalbero che parte da un nodo interno a_i è D^{l-l_i}.



Quindi, riassumendo:



Scriviamo che la differenza tra il numero totale di foglie D^l e il numero di foglie dei sottoalberi (ovvero il numero di foglie vietate a causa dei sottoalberi creati da ogni a_i) è maggiore o uguale a 1:

$$D^l - \sum_{i=1}^{k-1} D^{l-l_i} \geq 1$$

$$D^l \left(1 - \sum_{i=1}^{k-1} D^{-l_i}\right) \geq 1$$

$$1 - \sum_{i=1}^{k-1} D^{-l_i} \geq D^{-l_k}$$

$$1 \geq \sum_{i=1}^{k} D^{-l_i}$$

□

Per il **Teorema Diretto**, si può leggere la dimostrazione "al contrario". In altre parole, vengono date le istruzioni per costruire l'albero, ovvero si disegna l'albero completo e si inizia ad etichettare. Più precisamente, si prende a_1, lo si mette a lunghezza l_1, e fino alle foglie si segna il resto dell'albero (il sottoalbero con radice a_1) come vietato. Si continua così per tutti gli a_i. □

Dimostrazione Teorema Inverso, caso φ UD Siano $l_1 \leq \cdots \leq l_k = l$ le lunghezze delle codifiche di $a_1, \ldots, a_k$. Consideriamo $\mathcal{N}(n,h)$ numero di stringhe su $\mathcal{A}^n$ (ovvero di lunghezza n) che hanno una codifica φ UD di lunghezza h. Sia $|\mathcal{B}^h| = D^h$ (numero di stringhe di lunghezza h su $\mathcal{B}$). Poiché φ è UD, $\mathcal{N}(n,h) \leq D^h$.

$$\sum_{i=1}^{k} D^{-l_i} = D^{-l_1} + D^{-l_2} + \cdots + D^{-l_k}$$

Studiamo la crescita di tale oggetto alla potenza di n, quando n va ad infinito. Questo perché, se la somma è > 1, allora la potenza va ad infinito; se la somma è < 1, allora la potenza va a 0 (è limitata); se la somma è $= 1$, allora la potenza va a 1.

$$\forall n \qquad \left(D^{-l_1} + D^{-l_2} + \cdots + D^{-l_k}\right)^n \qquad \text{(chiamiamola } \alpha^n\text{)}$$

Se si svolge l'elevamento a potenza di tale polinomio, si otterrà una serie di addendi del seguente tipo:

$$D^{-l_1 \cdot n} + \cdots + D^{(-l_1)+(-l_2)+(-l_1)+\ldots} + \cdots + D^{-l_k \cdot n}$$

dove $-l_1 \cdot n$ è la lunghezza della codifica della stringa $a_1 a_1 \dots a_1$, con n ripetizioni di a_1, ovvero $|\varphi(a_1 \dots a_1)| = l_1 \cdot n$. Allo stesso modo, $-l_k \cdot n$ è la lunghezza della codifica della stringa $a_k a_k \dots a_k$, con n ripetizioni di a_k. Un generico esponente all'interno, ad esempio $-(l_1 + l_2 + l_1 + \dots)$ è una somma di lunghezze di codifiche, e ammonta alla lunghezza di una generica stringa. Ad esempio, $-(10)$, se chiamo $10 = h$.

Si ha che D^{-h} si verifica nella somma esattamente $\mathcal{N}(n, h)$ volte (alcune delle quali saranno 0). Quindi, la somma $D^{-l_1 \cdot n} + \dots + D^{-l_k \cdot n}$ si può riscrivere come

$$\begin{aligned} \mathcal{N}(n,0)D^{-0} + \mathcal{N}(n,1)D^{-1} + \dots + \underbrace{\mathcal{N}(n, n \cdot l)}_{\geq 1} D^{-n \cdot l} &\leq D^0 D^{-0} + D^1 D^{-1} + \dots + D^{n \cdot l} D^{-n \cdot l} \\ 1 + 1 + \dots + 1 &\leq n \cdot l + 1 \\ \forall n \qquad \alpha^n &\leq \underbrace{l}_{costante} \cdot n \end{aligned}$$

Da cui si ricava

$$\alpha \leq 1$$

Quindi

$$\sum_{i=1}^{k} D^{-l_i} \leq 1$$

□

2.2.2 Compressione Massima

La sorgente che genera il messaggio del codice è stazionaria e senza memoria (Fig. 2.1). Poiché il

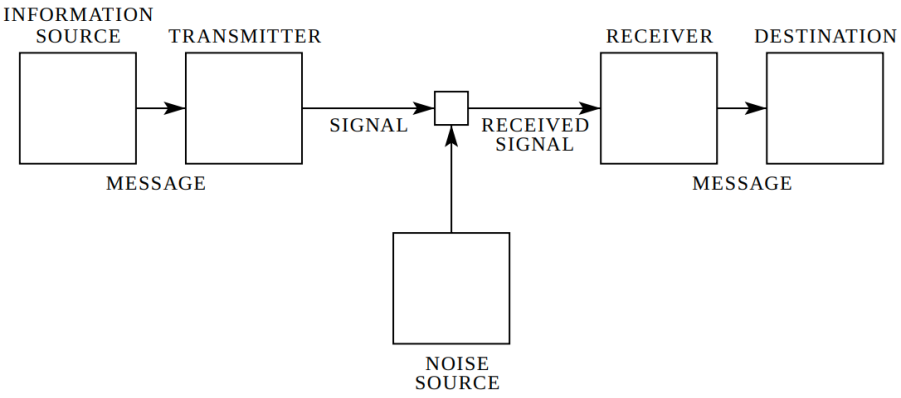


Figura 2.1: Diagramma schematico di un sistema di comunicazione, da C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, 1948.

messaggio codificato deve passare attraverso un canale, il quale ha una sua capacità e una sua velocità, lo si vuole comprimere il più possibile.

Ricordiamo che

$$P = \{p_1, \dots, p_k\} \qquad \mathcal{A} = \{a_1, \dots, a_k\} \qquad l_i = |\varphi(a_i)| \qquad EL(\varphi) = \sum_{i=1}^{k} p_i l_i$$

Teorema 2.2.4 (1° Shannon)

$$\varphi \text{ è } UD \quad \Rightarrow \quad EL(\varphi) \geq \mathcal{H}_D(P)$$

con

$$\mathcal{H}_D(P) = \sum_{i=1}^{k} p_i \cdot \log_D \frac{1}{p_i}$$

Dimostrazione

$$\begin{aligned} EL(\varphi) - \mathcal{H}_D(P) &= \sum_{i=1}^{k} p_i \underbrace{l_i}_{\log_D D^{l_i}} + \sum_{i=1}^{k} p_i \cdot \log_D p_i \\ &= \sum_{i=1}^{k} p_i \cdot \log_D (D^{l_i} \cdot p_i) \end{aligned}$$

Prima di proseguire, ricordiamo la seguente proprietà dei logaritmi su $\mathbb{N}$:

$$\log_e x \leq x - 1; \qquad -\log_e x \geq -(x-1)$$

e anche la proprietà dei logaritmi:

$$\log_b x = \frac{\log_c x}{\log_c b}$$

Continuiamo la dimostrazione:

$$\begin{aligned} EL(\varphi) - \mathcal{H}_D(P) &= \sum_{i=1}^{k} p_i \cdot \log_D (D^{l_i} \cdot p_i) \\ &= \frac{1}{log_e D} \sum_{i=1}^{k} p_i \cdot \log_e (D^{l_i} \cdot p_i) \\ &= -\frac{1}{\log_e D} \sum_{i=1}^{k} p_i \cdot \log_e \left(\frac{1}{D^{l_i} \cdot p_i} \right) \\ &\geq -\frac{1}{\log_e D} \sum_{i=1}^{k} p_i \cdot \left(\frac{1}{D^{l_i} \cdot p_i} - 1 \right) \\ &= -\frac{1}{\log_e D} \underbrace{\left(\sum_{i=1}^{k} \frac{1}{D^{l_i}} - \underbrace{1}_{\sum p_i} \right)}_{\leq 0} \\ &\geq 0 \end{aligned}$$

□

2.2.3 Shannon Code

Dal Teorema 1° Shannon abbiamo:

$$EL(\varphi) = \sum_{i=1}^{k} p_i l_i \;\geq\; \mathcal{H}_D(P) = \sum_{i=1}^{k} p_i \log_D \frac{1}{p_i}$$

La differenza sta in l_i e $\log_D \frac{1}{p_i}$, quindi vogliamo provare ad eguagliarli:

$$l_i = \log_D \frac{1}{p_i}$$

Ma $\log_D \frac{1}{p_i}$ non è necessariamente intero. Decidiamo quindi di considerare il suo primo intero più grande:

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil = \lceil -\log_D p_i \rceil$$

È sempre possibile definire un codice UD con tali lunghezze? Possiamo utilizzare Kraft-McMillan. Vogliamo controllare se

$$\sum_{i=1}^{k} D^{-\lceil -\log_D p_i \rceil} \stackrel{?}{\leq} 1$$

Sappiamo che

$$\lceil -\log_D p_i \rceil = -log_D p_i + \beta_i \qquad\qquad 0 \leq \beta_i < 1$$

Quindi possiamo scrivere

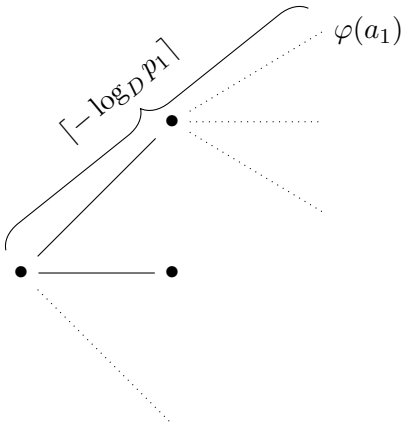
$$\begin{aligned}\sum_{i=1}^{k} D^{-(-log_D p_i + \beta_i)} &= \sum_{i=1}^{k} D^{log_D p_i} \cdot D^{\beta_i} \\ &= \sum_{i=1}^{k} p_i \cdot D^{-\beta_i} \\ &= \sum_{i=1}^{k} p_i \cdot \frac{1}{D^{\beta_i}}\end{aligned}$$

Ricordiamo che $0 \leq \beta_i < 1$ e $D > 1$, di conseguenza $\frac{1}{D^{\beta_i}} \leq 1$. Quindi

$$\sum_{i=1}^{k} p_i \cdot \frac{1}{D^{\beta_i}} \leq 1$$

$$\sum_{i=1}^{k} D^{-\lceil -\log_D p_i \rceil} \leq 1$$

Esiste quindi un prefix code con lunghezze $l_i = \lceil -\log_D pi \rceil$ definibile utilizzando una strategia greedy sull'albero D-ario.

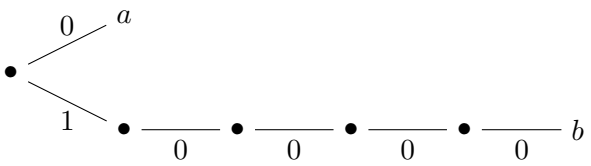


Esempio Siano $\mathcal{A} = \{a, b\}$, $\mathcal{B} = \{0, 1\}$, $P = \{1 - \frac{1}{32}, \frac{1}{32}\}$. Abbiamo che

$$l_2 = -\log_2 32 = 5 \qquad\qquad l_1 = \left\lceil -\log_2 \left(1 - \frac{1}{32}\right) \right\rceil = 1$$

Ciò significa che lo shannon code codifica l'alfabeto come

$$\varphi(a) = 0 \qquad\qquad \varphi(b) = 10000$$



TODO: finire lezione 5
TODO: osservazione su φ_{SF} in appunti lez 6

Definizione 2.2.7 (Efficienza (Efficiency of code))

$$Eff(\varphi) = \frac{\mathcal{H}_D(P)}{EL(\varphi)}$$

Eff è sempre ≤ 1, e ci dice quanto siamo vicini all'entropia, che non è sempre raggiungibile.

Ci chiediamo perché Huffmann (merge delle foglie) è ottimale, mentre Shannon-Fano (splitting dalla radice) non lo è?

2.3 Codici Stream

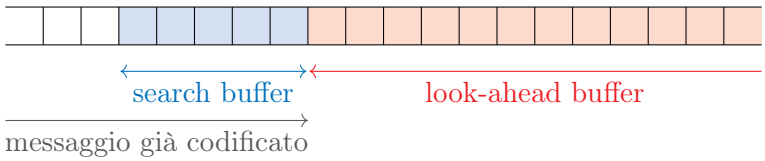
Capitolo 6 del libro di MacKay. In questo capitolo discuteremo uno schema di compressione dei dati. La codifica Lempel-Ziv è un metodo "universale", progettato secondo la filosofia per cui vorremmo un unico algoritmo di compressione che faccia un lavoro ragionevole per qualsiasi fonte. In effetti, per molte fonti della vita reale, le proprietà universali di questo algoritmo valgono solo nel limite di quantità di dati eccessivamente grandi, ma, comunque, la compressione Lempel-Ziv è ampiamente utilizzata e spesso efficace.

2.3.1 Un po' di Storia

TODO: scrivere la storia

2.3.2 Lempel-Ziv

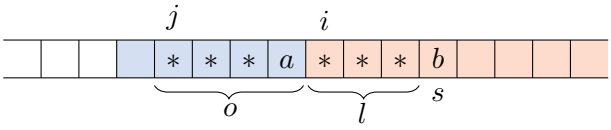
Immaginiamo un messaggio come uno stream (vettore) di caratteri su $\mathcal{A}^*$.



Cerchiamo il prefisso più lungo del look-ahead buffer che sia presente nel search buffer. Se lo troviamo, lo sostituiamo con un puntatore alla sua posizione nel search buffer e la sua lunghezza. Se non lo troviamo, scriviamo il carattere e passiamo al carattere successivo.

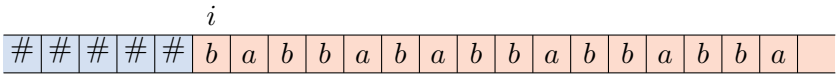
Ad ogni iterazione, l'algoritmo produce triple della forma (o, l, s), dove

- o è l'offset, ovvero la distanza tra i (primo carattere del look-ahead buffer) e j (primo carattere del prefisso nel search buffer)
- l è la lunghezza del prefisso
- s è il primo carattere mismatching

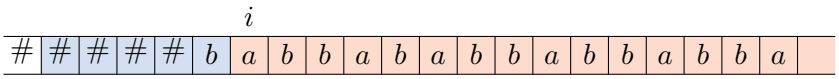


2.3.2.1 LZ77

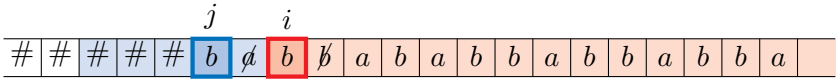
Abbiamo un search buffer di lunghezza 5, e la stringa *babbababbabbabba*. All'inizio, il search buffer è vuoto, e possiamo quindi immaginarlo con caratteri al di fuori dell'alfabeto.



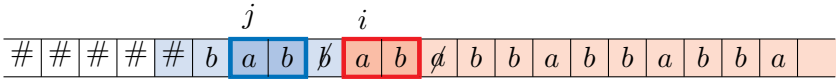
La prima tripla (o, l, s) della codifica è quindi $(0, 0, b)$. Proseguiamo spostando il search buffer di un carattere a destra:



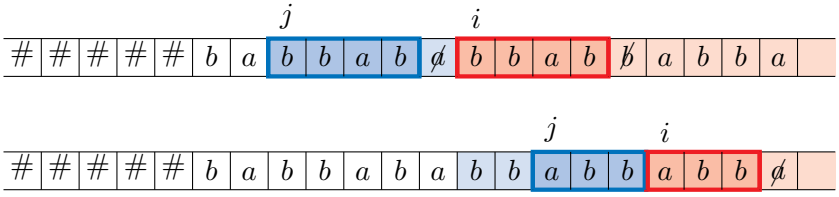
Il prefisso del look-ahead buffer inizia per a, e quindi non è presente nel search buffer. La seconda tripla è $(0, 0, a)$, e il search buffer viene spostato di un carattere a destra:



Confrontiamo nuovamente il prefisso del look-ahead buffer con il search buffer. Il prefisso inizia per b, che è presente anche nel search buffer. Continua con b, ma nel search buffer la b è seguita da una a. Il prefisso è quindi b, di lunghezza 1, la distanza tra j e i è 2, e il primo carattere mismatching è b. La terza tripla è quindi $(2, 1, b)$, e il search buffer viene spostato di due caratteri a destra:



Anche in questo caso cominciamo a guardare i caratteri del look-ahead buffer, e cerchiamo il più lungo prefisso presente anche nel search buffer. Il prefisso è ab (lunghezza 2), la distanza tra j e i è 3, e il primo carattere mismatching è una a. La tripla è quindi $(3, 2, a)$, e proseguiamo spostando il search buffer di tre caratteri a destra. Gli ultimi due passaggi sono i seguenti:



Corrispondenti alle triple $(5, 4, b)$ e $(3, 3, a)$ rispettivamente.

La codifica del messaggio *babbababbabbabba* con l'algoritmo LZ77 è quindi rappresentata dalla sequenza di triple $(0, 0, b)(0, 0, a)(2, 1, b)(3, 2, a)(5, 4, b)(3, 3, a)$.

L'esempio appena mostrato utilizza un messaggio piuttosto corto, e quindi non è possibile apprezzare la compressione. Per messaggi più lunghi, la dimensione delle triple è molto più piccola della dimensione del messaggio originale.

In quale caso LZ77 comprime molto? Immaginiamo di avere il messaggio *aaaaaaaaaaaaaaab*. Il primo mismatch si trova solo all'ultimo carattere, tutta la parte di ripetizione di a è il prefisso più lungo. Nel caso migliore, si può ottenere una compressione esponenziale.

Per questo tipo di codifica, il decodificatore deve essere a conoscenza dell'alfabeto e della lunghezza del search buffer. Più precisamente, deve avere un buffer di lunghezza almeno pari alla lunghezza del buffer del codificatore.

Note La sorgente che genera le lettere del messaggio è stazionaria e senza memoria. Le probabilità non cambiano nel tempo.

LZ esegue i cambiamenti sulle probabilità durante la codifica, mentre S e SF no. Questi ultimi due dovrebbero costruire un altro albero se le probabilità cambiano.

Capitolo 3

Codifica di Canale Rumoroso

3.1 Variabili Aleatorie Dipendenti

Capitolo 8 del libro di MacKay. Negli precedenti capitoli sulla compressione dei dati ci siamo concentrati sui vettori aleatori x provenienti da una distribuzione di probabilità estremamente semplice, ovvero la distribuzione separabile in cui ciascuna componente x_n è indipendente dalle altre.

In questo capitolo considereremo insiemi congiunti in cui le variabili aleatorie sono dipendenti. Questo materiale ha due motivazioni. Innanzitutto, i dati del mondo reale hanno correlazioni interessanti, quindi per eseguire una buona compressione dei dati, dobbiamo sapere come lavorare con modelli che includono dipendenze. In secondo luogo, un canale rumoroso con input x e output y definisce un insieme congiunto in cui x e y sono dipendenti (se fossero indipendenti, sarebbe impossibile comunicare sul canale) quindi la comunicazione sui canali rumorosi è descritto in termini di entropia degli insiemi congiunti.

3.1.1 Divergenza e Disuguaglianza di Gibbs

La divergenza di Kullback-Leibler è una misura della differenza tra due distribuzioni di probabilità. Siano P e Q due distribuzioni di probabilità su X. La divergenza di Kullback-Leibler tra P e Q è definita come

$$\begin{aligned} D(P||Q) &= \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \\ &= \sum_{x\in X} P(x)\left(\log\left(\frac{1}{Q(x)}\right) - \log\left(\frac{1}{P(x)}\right)\right) \end{aligned}$$

Notiamo che $\sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$ è il vaore atteso di $\log\left(\frac{P(x)}{Q(x)}\right)$ rispetto a P. Inoltre, $\log\left(\frac{P(x)}{Q(x)}\right)$ può essere espresso come $\log\left(\frac{1}{Q(x)}\right) - \log\left(\frac{1}{P(x)}\right)$, con $-P(x)\log\left(\frac{1}{P(x)}\right)$ entropia della distribuzione P.

Il significato della divergenza, in questo contesto, è "quanti bit in più devo usare per codificare una sorgente utilizzando Q al posto di P", con P distribuzione reale dei dati.

Esempio Se si hanno dei caratteri con probabilità 0, si utilizzerà per essi una codifica molto lunga, così da conservare codifiche brevi per altri caratteri con probabilità maggiori.

$$\begin{aligned} \mathcal{A} &= \{a, b, c, d, e\} \\ P &= \{\ldots, \ldots, \ldots, 0, 0\} \\ Q &= \{0.2, 0.2, 0.2, 0.2, 0.2\} \end{aligned}$$

Non conosciamo la distribuzione reale dei dati P, ma abbiamo dedotto Q.

Proprietà La divergenza non è simmetrica.

$$D(P||Q) \neq D(Q||P)$$

La divergenza è sempre non negativa, e vale 0 se e solo se $P = Q$.

$$D(P||Q) \geq 0 \qquad D(P||Q) = 0 \Leftrightarrow P = Q$$

Quest'ultima proprietà è una conseguenza della **disuguaglianza di Gibbs**.

3.1.2 Entropia e Mutual Information

Siano X, Y variabili aleatorie. Come abbiamo già visto, l'entropia congiunta è

$$\mathcal{H}(X,Y) = \sum_{\substack{x\in X \\ y\in Y}} p(x,y) \log\left(\frac{1}{p(x,y)}\right)$$

L'entropia è additiva sse X e Y sono indipendenti

$$\mathcal{H}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y) \quad \Leftrightarrow \quad p(x,y) = p(x)p(y)$$

L'entropia condizionale di X dato $Y = y$, ovvero se si conosce il valore di una delle due variabili, è l'entropia della distribuzione di probabilità $P(x|y)$

$$\mathcal{H}(X|Y=y) = \sum_{x\in X} p(x|y) \log\left(\frac{1}{p(x|y)}\right)$$

L'entropia condizionale di X dato Y è la media, su tutti i valori di Y, dell'entropia condizionale di X dato $Y = y$. L'incertezza del valore di X se si conosce il valore di Y è quindi

$$\begin{aligned}\mathcal{H}(X|Y) &= \sum_{y\in Y} p(y)\mathcal{H}(X|Y=y) \\ &= \sum_{\substack{x\in X \\ y\in Y}} p(x,y) \log\left(\frac{1}{p(x|y)}\right)\end{aligned}$$

***X* e *Y* indipendenti** Se X e Y sono indipendenti, allora

$$\begin{gathered}p(x,y) = p(x)p(y) \\ p(x|y) = p(x) \quad \Rightarrow \quad \mathcal{H}(X,Y) = \mathcal{H}(X)\end{gathered}$$

(Dimostrare che se X e Y sono indipendenti, allora $\mathcal{H}(X,Y) = \mathcal{H}(X)$)

***X* e *Y* dipendenti** Conosciuta anche come *chain rule* per l'entropia. L'entropia congiunta, $\mathcal{H}(X,Y)$, l'entropia condizionale, $\mathcal{H}(X|Y)$ o $\mathcal{H}(Y|X)$, e l'entropia marginale, $\mathcal{H}(X)$ o $\mathcal{H}(Y)$, sono legate dalla seguente relazione

$$\begin{aligned}\mathcal{H}(X,Y) &= \mathcal{H}(X) + \mathcal{H}(Y|X) \\ &= \mathcal{H}(Y) + \mathcal{H}(X|Y)\end{aligned}$$

Abbiamo un analogo al caso delle probabilità, solo con l'addizione al posto della moltiplicazione, in quanto passiamo alla funzione logaritmo.

3.1.2.1 Mutual Information

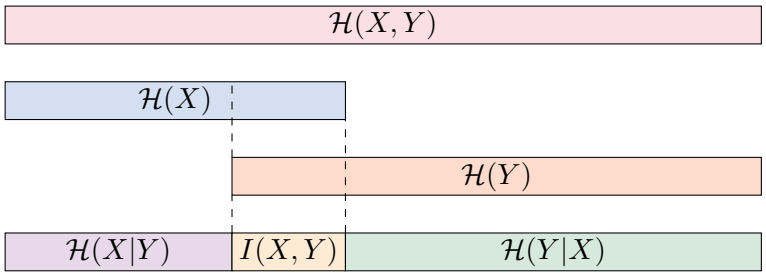
La **mutual information** è una misura della dipendenza tra due variabili aleatorie. È definita come

$$\begin{aligned} I(X,Y) &= \mathcal{H}(X) - \mathcal{H}(X|Y) \\ &= \mathcal{H}(Y) - \mathcal{H}(Y|X) \end{aligned}$$

Misura la riduzione media dell'incertezza su X se si conosce il valore di Y, o viceversa. Vale che $I(X,Y) \geq 0$, e in particolare $I(X,Y) = 0$ se X e Y sono indipendenti. Inoltre, $I(X,Y)$ è massimo quando l'incertezza su X se si conosce Y (o viceversa) è nulla:

$$X = Y \quad \Rightarrow \quad \mathcal{H}(X|Y) = 0 \quad \Rightarrow \quad I(X,Y) = \mathcal{H}(X)$$

Rappresentazione visuale $I(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y)$



Teorema 3.1.1 *La mutual information è pari alla divergenza tra la distribuzione congiunta e il prodotto delle distribuzioni marginali*

$$I(X,Y) = D(XY||X \otimes Y)$$

con XY distribuzione congiunta $p(X,Y)$, e $X \otimes Y$ prodotto delle distribuzioni marginali $p(X) \cdot p(Y)$.

Se X e Y sono indipendenti, allora $p(X,Y) = p(X) \cdot p(Y)$, e quindi $I(X,Y) = 0$.

Capitolo 4

Kolmogorov Complexity

Prerequisiti al capitolo 2 del libro di Papadimitriou. Argomenti al capitolo 8 del libro di T. Cover, nel libro di Li e Vitanyi, al capitolo 3 del libro di Papadimitriou.

4.1 Nozioni Preliminari

La complessità di Kolmogorov è una nozione di complessità algoritmica. Questo argomento è un "ponte" tra le due parti di questo corso.

Mentre negli anni '40 Shannon (Princeton, Bell Labs, MIT), Fano (PoliTO, MIT), e Huffmann (MIT, UCAL Santa Cruz) lavoravano nella parte occidentale del mondo, dall'altra parte della cortina di ferro, nel 1965 Kolmogorov (Moscow University) analizzava la nozione di casualità (randomness) delle stringhe.

Ad esempio, se si riceve la stringa 0100110100111010, o la stringa 1111111111111111, quale delle due è più casuale? Poiché la sorgente che le genera è stazionaria e senza memoria, entrambe le stringhe sono equiprobabili. Kolmogorov non era soddisfatto di questa nozione di casualità.

Intuitivamente, la sua definizione di complessità di un oggetto è la lunghezza del programma più corto in grado di generarlo.

4.1.1 Macchine di Turing

Una macchina di Turing è un modello di calcolo inventato da Alan Turing nel 1936. È possibile immaginarla come un nastro, o registro, con un numero infinito di celle. In ogni cella si possono scrivere dei simboli dall'alfabeto Σ, o dei simboli speciali. In un dato momento, con un puntatore è possibile leggere il contenuto di una cella, eventualmente cambiarlo, cambiare lo stato del puntatore, e spostarsi a destra o a sinistra di una cella (o rimanere fermi).

Definizione 4.1.1 (Macchina di Turing) *Una macchina di Turing è una tupla $\mathcal{M}(K, \Sigma, \delta, s)$, dove*

- *K è un insieme finito di stati, di cui $s \in K$ è quello iniziale*
- *Σ è un alfabeto finito, e $\triangleright, \sqcup \in \Sigma$. Inoltre, $\Sigma \cap K = \emptyset$*
- *δ è la funzione di transizione, definita come $\delta : K \times \Sigma \to K \cup \{yes, no, halt\} \times \Sigma \times \{\rightarrow, \leftarrow, -\}$*

$\forall q \in K$ con $\delta(q, \triangleright) = (q', \triangleright, \rightarrow)$.

Il simbolo $\triangleright$ si legge *start*, $\sqcup$ si legge *blank*, e $\rightarrow, \leftarrow, -$ indicano rispettivamente il movimento a destra (*right*), a sinistra (*left*), o l'immobilità del puntatore (*stay*).

Esempio: raddoppia un numero Siano $\Sigma = \{0, 1, \triangleright, \sqcup\}$, $K = \{s\}$. Le funzioni di transizione per la macchina di Turing che raddoppia un numero sono

$$\delta(s, \triangleright) = (s, \triangleright, \rightarrow)$$

$$\delta(s, 0) = (s, 0, \rightarrow)$$
$$\delta(s, 1) = (s, 1, \rightarrow)$$
$$\delta(s, \sqcup) = (halt, 0, -)$$

Partendo dalla stringa 1011 e dalla macchina di Turing appena definita, si ottiene la stringa 10110.

4.2 Complessità di Kolmogorov

Definizione 4.2.1 (Macchine di Turing Universali) *Una macchina di Turing universale è una macchina di Turing $\mathcal{U}$ che prende in input un'altra macchina di Turing $\mathcal{M}$ e un input x, e simula l'esecuzione di $\mathcal{M}$ su x.*

$$\mathcal{U}(\mathcal{M}; x) = \mathcal{M}(x)$$

La tesi di Church-Turing afferma che ogni funzione calcolabile può essere calcolata da una macchina di Turing. Di conseguenza, le macchine di Turing universali devono esistere. Intuitivamente, una macchina di Turing universale può essere vista come un compilatore.

Definizione 4.2.2 (Kolmogorov Complexity) *La complessità di Kolmogorov di una stringa $x \in \Sigma^*$ è la lunghezza della macchina di Turing $\mathcal{M}$ più corta tale che $\mathcal{U}(\mathcal{M}) = x$.*

$$K_{\mathcal{U}}(x) = \min_{\mathcal{M}:\mathcal{U}(\mathcal{M})=x} |\mathcal{M}|$$

con $\mathcal{U}$ una macchina di Turing universale fissata.

Il nastro della macchina universale $\mathcal{U}$ è composto da due parti: la prima parte contiene la codifica binaria della macchina $\mathcal{M}$, e la seconda parte contiene l'input x.

Esempio Macchina di Turing $\mathcal{M}$ che non prende nulla in input e produce in output la stringa 0110.

Per descrivere $\mathcal{M}$, è necessario specificare K, Σ, δ, s, con $\delta(s, \triangleright) = (q_1, \triangleright, \rightarrow)$, $\delta(q_1, \sqcup) = (q_2, 0, \rightarrow)$, $\delta(q_2, \sqcup) = (q_3, 1, \rightarrow)$, $\delta(q_3, \sqcup) = (q_4, 1, \rightarrow)$, $\delta(q_4, \sqcup) = (halt, 0, -)$. Per codificare $\mathcal{M}$ in binario, bisogna specificare in binario gli argomenti descritti precedentemente.

Se si considerano le macchine di Turing che producono 0110 in output e non prendono nulla in input, la lunghezza della più corta è la complessità di Kolmogorov della stringa 0110.

Definizione 4.2.3 (Complessità di Kolmogorov Condizionale)

$$K_{\mathcal{U}}(x|y) = \min_{\mathcal{M}:\mathcal{U}(\mathcal{M};y)=x} |\mathcal{M}|$$

La macchina $\mathcal{M}$ conosce qualcosa della stringa x che deve produrre in output: $\mathcal{U}(\mathcal{M}; y) = \mathcal{M}(y) = x$. Quindi

$$K_{\mathcal{U}}(x) \geq K_{\mathcal{U}}(x|y)$$

La complessità di Kolmogorov condizionale è la lunghezza della macchina di Turing più corta che produce in output x se si conosce y. La conoscenza di y dà informazioni su x, e quindi aiuta a costruire una macchina $\mathcal{M}$ più corta.

Esempio Sia $000\ldots0$ una sequenza di m 0, che vogliamo produrre in output. Se non conosciamo m, si ha qualcosa del tipo

$$m \begin{cases} \texttt{print } 0 \\ \texttt{print } 0 \\ \ldots \\ \texttt{print } 0 \end{cases}$$

Ma, dato m, si ha

```
for (i = 1 to m){
  print 0
}
```

In questo esempio abbiamo quindi $K_{\mathcal{U}}(x||x|)$.

Cosa succederebbe se cambiassimo da $\mathcal{U}$ ad un'altra macchina di Turing universale $\mathcal{A}$?

Teorema 4.2.1 *Siano $\mathcal{U}$ e $\mathcal{A}$ due macchine di Turing universali. Consideriamo $K_{\mathcal{U}}(x|y)$ e $K_{\mathcal{A}}(x|y)$. Allora*

$$K_{\mathcal{U}}(x|y) \leq K_{\mathcal{A}}(x|y) + c_{\mathcal{AU}}$$

con $c_{\mathcal{AU}}$ costante che dipende solo da $\mathcal{A}$ e $\mathcal{U}$.

Dimostrazione Sia $P_{\mathcal{A}}$ il programma più corto per $\mathcal{A}$ che produce x dato y:

$$K_{\mathcal{A}}(x|y) = |P_{\mathcal{A}}|$$

Questo significa

$$\mathcal{A}(P_{\mathcal{A}}; y) = P_{\mathcal{A}}(y) = x$$

Sia $\mathcal{U}$ una macchina di Turing universale. Allora $\mathcal{U}$ con in input $\mathcal{C}_{\mathcal{A}}$ (codifica di $\mathcal{A}$) seguito da $P_{\mathcal{A}}$ e y produce in output x:

$$\mathcal{U}(\underbrace{\mathcal{C}_{\mathcal{A}}; P_{\mathcal{A}};}_{\substack{\text{programma per } \mathcal{U} \text{ che,} \\ \text{dato } y, \\ \text{produce } x}} y) = \mathcal{A}(P_{\mathcal{A}}; y) = P_{\mathcal{A}}(y) = x$$

Quindi

$$\begin{aligned} K_{\mathcal{U}}(x|y) &\leq |\mathcal{C}_{\mathcal{A}}; P_{\mathcal{A}}| \\ &= |\mathcal{C}_{\mathcal{A}}| + |P_{\mathcal{A}}| + 1 \\ &= K_{\mathcal{A}}(x|y) + \underbrace{|\mathcal{C}_{\mathcal{A}}| + 1}_{c_{\mathcal{AU}}} \end{aligned}$$

□

Corollario 4.2.1.1

$$|K_{\mathcal{U}}(x|y) - K_{\mathcal{A}}(x|y)| \leq c_{\mathcal{AU}} \qquad \forall x, y$$

Quando $y = \varepsilon$

$$|K_{\mathcal{U}}(x) - K_{\mathcal{A}}(x)| \leq c_{\mathcal{AU}} \qquad \forall x$$

D'ora in poi, quindi, sapendo che cambia solo una costrante, si ometterà la macchina di Turing universale, e si scriverà semplicemente $K(x)$, o $K(x|y)$.

Esempio Vogliamo produrre in output la stringa $x = 010$, e abbiamo il programma P

```
print 0
print 1
print 0
```

Codificandolo in binario, si ha, ad esempio, $\text{bin}(P) = 300$. Si può concludere che la complessità di Kolmogorov di x è al massimo 300:

$$K(010) \leq 300$$

Per poter scrivere $=$ invece che $\leq$, bisognerebbe dimostrare che tutti i programmi di lunghezza 299 non producono in output la stringa 010. Ma almeno uno di essi potrebbe non terminare, e è impossibile dimostrarlo.

Dall'esempio precedente abbiamo visto come non sia possibile calcolare la complessità di Kolmogorov di una stringa. Tuttavia, possiamo calcolare la complessità di Kolmogorov di una stringa con una certa precisione dando dei *bounds*.

Teorema 4.2.2 (Teorema sui Limiti alla Complessità di Kolmogorov sulle Stringhe)

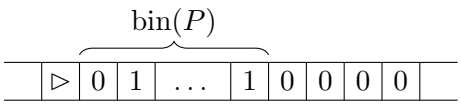
$$K(x) \leq |x| + c$$

Nel libro di T. Cover viene utilizzato un alfabeto con solo due simboli, $\Sigma = \{0, 1\}$. Questo dà un risultato diverso (più debole) rispetto a quello che si raggiunge utilizzando l'alfabeto $\Sigma = \{0, 1, \sqcup\}$.

Dimostrazione (intuizione) Il programma che produce in output $x = x_1 x_2 \dots x_n$ è della forma

$$\text{bin}(P) \begin{cases} \texttt{print } x_1 \\ \texttt{print } x_2 \\ \dots \\ \texttt{print } x_n \end{cases}$$

con $\text{bin}(P)$ della forma $010011 \dots 11$. Dando $\text{bin}(P)$ in input ad una macchina di Turing universale $\mathcal{U}$ utilizzando l'alfabeto $\Sigma = \{0, 1\}$, si avrà



C'è quindi un problema: utilizzando 0 invece di $\sqcup$, come si può capire dove finisce $\text{bin}(P)$? Una possibile soluzione è quella di raddoppiare ogni bit, e codificare lo stop con una coppia "non valida", ad esempio 10. Ad esempio, se $\text{bin}(P) = 010011$, si avrà

$$0\,0\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1\underbrace{1\,0}_{\text{stop}}$$

Utilizzando questo metodo, quando si vuole x in output si può settare il primo bit (dopo $\triangleright$) a 0 o a 1, dando due significati diversi alla stringa che li segue. In particolare, se la prima cifra è

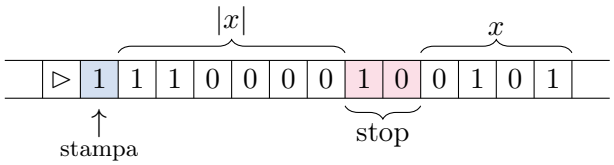
- 0, segue un generico programma P (ovvero 0 = esegui)
- 1, segue la codifica di x con il raddoppio dei bit (ovvero 1 = stampa)

Così facendo, con l'alfabeto $\Sigma = \{0, 1\}$, si ha

$$K(x) \leq 2|x| + 3$$

con 3 pari alla somma del primo bit (0 o 1) e della coppia "non valida" (10) che indica lo stop.

Analizziamo un'altra soluzione con un esempio. Sia $x = 0101$, e $|x| = 4$. Codifichiamo la lunghezza di x in binario come $|x| = 4 = 100$, e raddoppiamo i bit, ottenendo 110000. Quindi si avrà



In questo caso

$$K(x) \leq 2\log|x| + 2 + 1 + |x|$$

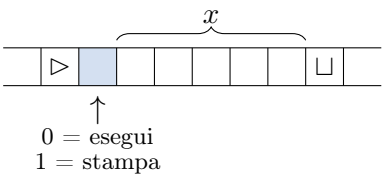
Quando x è molto lunga, $2\log|x|$ è molto più piccolo di $|x|$, e quindi si può trascurare. Quindi questa proposta è più corta (migliore) della precedente.

$$K(x) \leq 2\log|x| + 2 + 1 + |x| \leq 2|x| + 3$$

Un'ulteriore soluzione è quella di utilizzare il trucco precedente sulla lunghezza della lunghezza di x, e così via. In generale, si ha

$$K(x) \leq \cdots \leq \cdots \leq 2\log^*|x| + |x| + c$$

Ma ciò che il teorema afferma è più forte. Utilizzando l'alfabeto $\Sigma = \{0, 1, \sqcup\}$, si ha



Quindi

$$K(x) \leq |x| + 1$$

con $1 = c$ nel caso generico. □

Quante stringhe possono avere complessità di Kolmogorov minore di un dato valore k? Per stringhe di lunghezza 0 (stringa vuota) esiste 1 macchina di Turing (di lunghezza 0) che la produce in output. Per stringhe di lunghezza 1, esistono 2 macchine di Turing: quella la cui codifica è 0, e quella la cui codifica è 1, e così via.

lunghezza stringa	numero di macchine
0	0
1	2
⋮	⋮
h	2^h
⋮	⋮
$k-1$	2^{k-1}

In generale

$$\sum_{i=0}^{k-1} 2^i = 2^k - 1$$

ovvero al massimo $2^k - 1$ stringhe possono avere complessità di Kolmogorov minore di k.

Teorema 4.2.3

$$\exists x \qquad K(x) \geq |x|$$

Dimostrazione Sia $|x| = k$. Allora ci sono 2^k stringhe di lunghezza k. Ma esistono al massimo $2^k - 1$ stringhe di lunghezza minore di k. Quindi esiste almeno una stringa di lunghezza k che ha complessità di Kolmogorov almeno pari a k. □

Abbiamo trovato un limite superiore e inferiore alla complessità di Kolmogorov di una stringa.

$$|x| \leq K(x) \leq |x| + c$$

4.2.1 Complessità di Kolmogorov vs Entropia di Shannon

Da Complessità a Entropia Consideriamo

$$\varphi_K(x) = \text{bin}(\mathcal{M})$$

con $\varphi_K(x)$ codifica di Kolmogorov di x, e $\mathcal{M}$ la più corta macchina di Turing che produce x in output (con input vuoto). $\varphi_K(x)$ è UD. $\mathcal{U}$ decodifica $\varphi_K(x)$, ovvero invio un programma, io e il ricevente abbiamo la stessa macchina di Turing universale (il "compilatore"), e il ricevente riceve in programma.

$$|\varphi_K(x)| = K(x)$$

Consideriamo tutti i possibili messaggi di lunghezza n

$$EL_n(\varphi_K(x)) = \sum_{x \in \Sigma^n} p(x)K(x) = E_n(K(x))$$

con $E_n(K(x))$ complessità di Kolmogorov media su tutti i messaggi di lunghezza n. Per il Teorema di Shannon

$$EL_n(\varphi_K(x)) \geq \mathcal{H}(P^n) = n\mathcal{H}(P)$$

con P distribuzione di probabilità sull'alfabeto Σ. Quindi

$$E_n(K(x)) \geq n\mathcal{H}(P)$$

La complessità di Kolmogorov media di un singolo carattere è

$$\frac{E_n(K(x))}{n} \geq \mathcal{H}(P)$$

Da Entropia a Complessità Ogni codice $\varphi(x)$ può essere interpretato come un particolare programma che produce in output la stringa (codificata) x. Quindi

$$K(x) \leq |\varphi(x)| + \text{Decoder } \varphi$$

Per lo Shannon code abbiamo visto che

$$EL_n(\varphi) \leq n\mathcal{H}(P) + 1$$

Ciò significa che

$$E_n(K(x)) \leq n\mathcal{H}(P) + 1 + K(P)$$

con $K(P) = |\text{Decoder } \varphi|$ complessità di Kolmogorov del Decoder (posso inviare il codice più corto possibile che produce in output il Decoder). Quindi

$$\frac{E_n(K(x))}{n} \leq \mathcal{H}(P) + \underbrace{\frac{1 + K(P)}{n}}_{\substack{\text{perché} \\ \text{consideriamo} \\ \text{lo Shannon} \\ \text{code}}}$$

Parte II

Complessità

Capitolo 5

Introduzione

Libro di Papadimitriou main reference.

In questa parte utilizzeremo come modello di computazione le **macchine di Turing** (MdT). Esistono diversi modelli di MdT: macchine di Turing multinastro, macchine di Turing input/output, macchine di Turing con oracolo, macchine di Turing nondeterministiche. Le MdT verranno utilizzate per confrontare i diversi risultati di complessità che possiamo ottenere.

Ci concentreremo sia su **complessità temporale** (time complexity) che **spaziale** (space complexity). Il focus non sarà sulla complessità di un dato algoritmo, ma sulla complessità di un problema. I **problemi** possono essere classificati come di decisione (decision problems), di funzione (function problems), o di ottimizzazione (optimization problems).

- **Decision problem** P : inputs $\to \{yes, no\}$
- **Function problem** computare una data funzione, ad esempio l'ordinamento di una lista
- **Optimization problem** tra tutti i possibili output, si vuole trovare quello che minimizza o massimizza una funzione di costo.

Esempio Sia $G = (V, E)$ un grafo, e $u, v \in V$ due nodi.

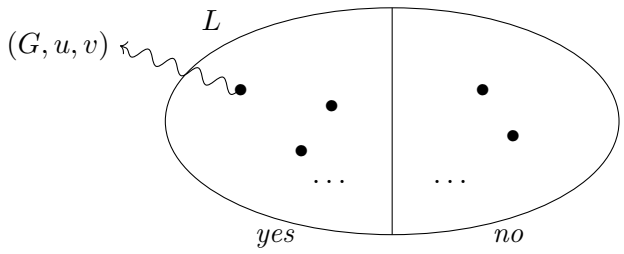
– decidere se esiste un cammino da u a v è un problema di decisione

– trovare un cammino da u a v è un problema di funzione

– trovare il cammino più corto da u a v è un problema di ottimizzazione

In questo corso ci concentreremo sui problemi di decisione. Se si ha una soluzione per un problema di funzione o di ottimizzazione, si possiede automaticamente una soluzione per il problema di decisione.

Immaginiamo tutti gli input possibili al problema dell'esempio precedente come ad un insieme infinito di tuple (G, u, v). Questo insieme si può dividere in due: il sottoinsieme dei *yes* di tutte le codifiche binarie di triple (G, u, v) tali che esiste un cammino in G da u a v, e, inversamente, il sottoinsieme *no*.



La codifica binaria di una tripla è una stringa del tipo $1011\ldots$. Più precisamente, è una stringa sull'alfabeto $\Sigma = \{0, 1\}$. L'insieme di tutte le possibili stringhe binarie è Σ^*. Questo insieme è quindi il linguaggio L sottoinsieme di Σ^*, ovvero $L \subseteq \Sigma^*$.

$$L = \{\text{bin}(G, u, v) | ln\ G\ u \to v\}$$

Esempio Consideriamo interi rappresentati in binario. Vogliamo decidere se un dato intero x è divisibile per 4.

$$\text{bin}(x) = 10\ldots11$$

in questo caso non è divisibile per 4. Un numero binario è divisibile per 4 se e solo se i due bit meno significativi sono 0.

$$\text{bin}(x) = x_n, x_{n-1}, \ldots, x_1, x_0 \quad \Leftrightarrow \quad x_0 = 0 \text{ and } x_1 = 0$$

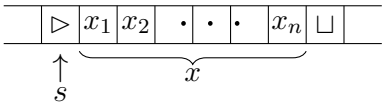
Il linguaggio indicato da questo problema di decisione è

$$L = \{x \in \{0,1\}^* | x = x_n, x_{n-1}, \ldots, x_1, x_0 \wedge x_0 = 0 \wedge x_1 = 0\}$$

Esempio: palindromo Decidere se una stringa è palindroma, con $\Sigma = \{0,1\}$.

$$x_1, x_2, x_3, \ldots, x_3, x_2, x_1$$

Ad esempio, $x = 101$ è palindroma, mentre $x = 1010$ non lo è. Cerchiamo il linguaggio $L = \{x | x \text{ è palindroma}\}$. Utilizziamo una macchina di Turing.



Si parte dallo stato s e si vuole finire nello stato p solo quando x è palindroma. Per decidere se x è palindroma, si può leggere x_1, ricordarne il valore nello stato del puntatore, e poi confrontarlo con x_n. Se sono uguali, si ripete lo stesso procedimento con x_2 e x_{n-1}, e così via. Se si arriva a x_n e x_1 senza aver trovato una discrepanza, allora x è palindroma. Se invece si trova una discrepanza, allora x non è palindroma. Le transizioni sono le seguenti:

$$\begin{aligned}\delta(s, \triangleright) &= (q, \triangleright, \rightarrow)\\ \delta(q, 1) &= (q_1, \triangleright, \rightarrow)\\ \delta(q, 0) &= (q_0, \triangleright, \rightarrow)\end{aligned}$$

TODO: finire di scrivere le transizioni Questa macchina eseguirà un numero quadratico di passi per controllare se la stringa x è palindroma: $O(|x|^2)$.

Se si vuole controllare in C (o in un altro linguaggio) se una stringa è palindroma, si può scrivere un programma che confronta il primo e l'ultimo carattere, poi il secondo e il penultimo, e così via, eseguendo un numero lineare di passi. La complessità è $O(|x|)$. Questo è un esempio di come la complessità di un problema dipenda dal modello di computazione utilizzato.

5.1 Tesi di Church-Turing Estesa

La tesi di Church-Turing afferma che ogni cosa che può essere computata, può essere computata da una macchina di Turing.

La versione estesa afferma che tutti i modelli (ragionevoli) di calcolo sono correlati polinomialmente. Questo significa che se un problema è risolvibile in tempo polinomiale in un modello di computazione, allora è risolvibile in tempo polinomiale in ogni modello di computazione.

In altre parole, la tesi di Church-Turing estesa afferma che la complessità computazionale di un problema è indipendente dal modello di calcolo utilizzato per risolverlo.

$$\underset{\text{problema}}{P} \rightarrow \text{MdT } O(f(n)) \rightarrow \underset{\text{modello}}{M} \; O(p(f(n)))$$

Ma è vera anche la direzione contraria. TODO: ???

Capitolo 6

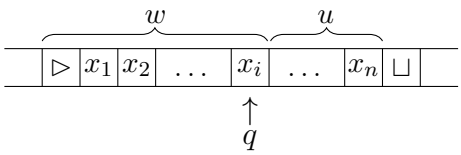
Macchine di Turing

6.1 Definizioni

Definizione 6.1.1 (Configurazione) *Una configurazione è una tripla (q, w, u), con*

- $q \in K \cup \{yes, no, halt\}$
- $w, u \in \Sigma^*$

Ad esempio, graficamente, una configurazione è



Definizione 6.1.2 (Configurazione Iniziale) *La configurazione iniziale su una stringa x è una tripla*

$$(1, \triangleright, x)$$

Definizione 6.1.3 (Configurazioni Finali) *Le configurazioni finali su una stringa x sono una tripla*

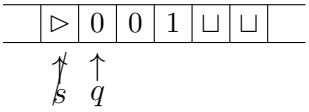
$$(H, w, u)$$

dove $H \in \{yes, no, halt\}$.

Definizione 6.1.4 (Passo di Computazione)

$$(q, w, u) \xrightarrow{\delta} (q', w', u')$$

Ad esempio, il passo di computazione è $(s, \triangleright, 001) \to (q, \triangleright 0, 01)$



Eseguito applicando $\delta(s, \triangleright) = (q, \triangleright, \to)$.

Definizione 6.1.5 (Time Complexity per una MdT $\mathcal{M}$ sull'input x) *$\mathcal{M}$ ha time complexity t su x se dopo esattamente t passi si raggiunge una configurazione finale.*

$$(s, \triangleright, x) \underbrace{\to \cdots \to}_{t \text{ passi}} (H, w, u)$$

Indicata in breve con $(s, \triangleright, x) \to^t (H, w, u)$.

$\mathcal{M}$ *ha time complexity* $f : \mathbb{N} \to \mathbb{N}$ *se,* $\forall x \in \Sigma^*$, $(s, \triangleright, x) \to^t (H, w, u)$ *con* $t \leq f(|x|)$.

La dimensione dell'input (bit length dell'input) è $|x|$. Questa è una complessità nel caso peggiore ($\leq$). Non stiamo utilizzando la notazione big-O.

6.2 Unlimited Register Machines

Una Unlimited Register Machine (URM) è una macchina di Turing con un numero illimitato di registri.

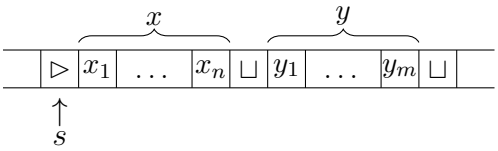
R_0	r_0
R_1	r_1
	...
R_m	r_m
	...

Ogni registro contiene un numero naturale. Quindi, il contenuto del registro R_m sarà $r_m \in \mathbb{N}$. Le operazioni possibili sono:

- **incremento** $S(i)$: $r_i := r_i + 1$
- **azzeramento** $Z(i)$: $r_i := 0$
- **trasferimento** $T(i, j)$: $r_j := r_i$, ovvero trasferisco il contenuto del registro R_i nel registro R_j
- **jump** $J(i, j, k)$: se $r_i = r_j$ allora salta all'istruzione k, altrimenti prosegue con l'istruzione successiva

Esempio Dati $x, y \in \mathbb{N}$, decidere se $x = y$.

MdT Si può utilizzare una macchina di Turing che contiene la rappresentazione binaria dei due interi, separati da un separatore.



Questa macchina richiede, nel caso peggiore, un numero quadratico di passi per terminare. La complessità è $\Theta(|x|^2)$.

URM Possiamo utilizzare una URM con x e y rispettivamente nei registri R_0 e R_1.

R_0	x
R_1	y

Alla fine, scriveremo 1 in R_0 se $x = y$, 0 altrimenti. Le istruzioni sono le seguenti:

1. $J(0, 1, 4)$
2. $Z(0)$
3. $J(0, 0, 100)$
4. $Z(0)$

5. $S(0)$

In questo caso, la complessità si può calcolare in due modi.

Definizione 6.2.1 (Time Complexity su URM)

- Uniform cost criterium *(criterio del costo uniforme): numero di istruzioni eseguite.*
- Logarithmic cost criterium *(criterio del costo logaritmico): ogni istruzione ha un costro proporzionale al numero di cifre coinvolte.*

Quindi, per questa macchina, la complessità è

- utilizzando il criterio del costo uniforme: $\Theta(1)$
- utilizzando il criterio del costo logaritmico: $\Theta(|x| + |y|)$

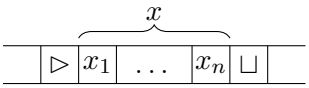
Nel secondo caso, ci si avvicina al costo per la macchina di Turing.

Mentre le macchine di Turing sono un modello di computazione sequenziale, nelle URM si ha l'istruzione *jump*. In altre parole:

- **MdT** 1 bit di informazione in ogni cella → tempo: numero di passi
- **URM** registri, un intero di lunghezza arbitraria (più bit) in ogni registro → tempo: numero di istruzioni (uniform time complexity)

Esempio Computare $x + 1$, $x \in \mathbb{N}$.

MdT Si ha una macchina di Turing che contiene x in binario.



Nel caso peggiore $x = 111\ldots1$, quindi la complessità è lineare $\Theta(n)$.

URM Si ha una URM con x nel registro R_0. È suffieciente una singola istruzione $S(0)$, quindi la complessità è $\Theta(1)$.

6.2.1 URM + Prodotto

Cambiamo il modello di computazione URM, considerando URM + prodotto. Oltre alle istruzioni $S(i)$, $Z(i)$, $T(i,j)$, e $J(i,j,k)$, aggiungiamo l'istruzione $P(i)$, che esegue l'operazione $r_i := r_i * r_i$.

Esempio di programma per URM + prodotto Dati $x, y \in \mathbb{N}$, decidere se $x = y$.

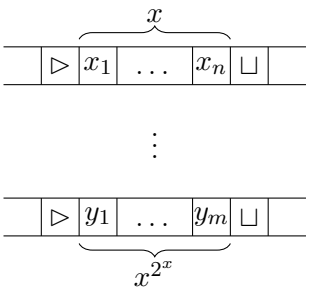
1. $J(1,2,5)$
2. $P(0)$
3. $S(2)$
4. $J(3,3,1)$

Pertendo da un input di x in R_0, x in R_1, e 0 in tutti gli altri registri. Si avrà:

R_0	x		R_0	x^2		R_0	$(x^2)^2$		R_0	$(x^4)^2$				R_0	x^{2^i}
R_1	x		R_1	x		R_1	x		R_1	x				R_1	x
R_2	0	$\rightarrow$	R_2	1	$\rightarrow$	R_2	2	$\rightarrow$	R_2	3	$\rightarrow$	$\ldots$	$\rightarrow$	R_2	2^i
R_3	0		R_3	0		R_3	0		R_3	0				R_3	0
R_4	$\vdots$		R_4	$\vdots$		R_4	$\vdots$		R_4	$\vdots$				R_4	$\vdots$

Il numero di istruzioni è lineare $\Theta(n)$.

MdT Se si eseguisse la stessa computazione su una macchina di Turing, si avrebbe



Quindi $\Omega(\log(x^{2^x})) = \Omega(2^x \log(x))$.

Questo risultato sembra contraddire la tesi di Church-Turing estesa, che afferma che tuttti i modelli **ragionevoli** di computazione sono correlati polinomialmente. Ma cosa significa *ragionevole*? Non si può avere una operazione che fa crescere "troppo" l'input (nell'esempio, il prodotto), si deve utilizzare il creiterio logaritmico.

In altre parole, se l'algoritmo utilizza operazioni che in un numero polinomiale di passi fanno crescere l'input esponenzialmente, e queste sono utilizzate un numero di volte che dipende dalla dimensione dell'input, allora si deve utilizzare un criterio logaritmico. Quando non si è sicuri della potenza delle operazioni della macchina, il costo di ogni singola operazione dev'essere proporzionale al numero di bit manipolati.

istruzione	uniform	logarithmic
$S(i)$	$\Theta(1)$	$\Theta(\log(r_i))$
$Z(i)$	$\Theta(1)$	$\Theta(1)$
$T(i,j)$	$\Theta(1)$	$\Theta(\log(r_i))$
$J(i,j,k)$	$\Theta(1)$	$\Theta(\min(\log(r_i), \log(r_j)))$
$P(i)$	$\Theta(1)$	$\Theta((\log(r_i))^2)$

Con r_i contenuto del registro i. In particolare per $P(i)$, nella moltiplicazione di un numero x per se stesso si ha $x_1, x_2, \ldots, x_n \times x_1, x_2, \ldots, x_n$. Si hanno x^n bit operazioni, quindi $O((\log(x))^2)$.

6.3 Ulteriori Definizioni

Come abbiamo visto, nei problemi di decisione si ha un input $x \in \Sigma^*$ e un output in $\{\text{yes}, \text{no}\}$. Possiamo definire un linguaggio L come l'insieme di tutte le stringhe che hanno output yes.

$$L \subseteq (\Sigma \backslash \{\sqcup\})^*$$

Un problema P è una funzione

$$P : \Sigma^* \to \{\text{yes}, \text{no}\}$$

Definizione 6.3.1 (Decidibilità di un Linguaggio da una MdT)

Una macchina di Turing $\mathcal{M}$ decide un linguaggio L

$$\Updownarrow$$

$$\forall x \in (\Sigma \backslash \{\sqcup\})^* \begin{cases} x \in L \to \mathcal{M}(x) = yes \\ x \notin L \to \mathcal{M}(x) = no \end{cases}$$

Il linguaggio L si dice ***ricorsivo****.*

Definizione 6.3.2 (Accettazione di un Linguaggio da una MdT)

Una macchina di Turing $\mathcal{M}$ accetta un linguaggio L

$$\Updownarrow$$

$$\forall x \in (\Sigma \backslash \{\sqcup\})^* \begin{cases} x \in L \to \mathcal{M}(x) = yes \\ x \notin L \to \mathcal{M}(x) \uparrow \ \textit{(non termina)} \end{cases}$$

Il linguaggio L si dice ***ricorsivamente enumerabile****.*

Teorema 6.3.1

$$L \text{ è ricorsivo } \Rightarrow \ L \text{ è ricorsivamente enumerabile}$$

Esempio Trovare un linguaggio L tale che L è ricorsivamente enumerabile ma non ricorsivo.
Nell'halting problem abbiamo

$$\mathcal{U}(\mathcal{M}; x) = \mathcal{M}(x)$$

L'halting language

$$H = \{(\text{bin}(\mathcal{M}); x) \mid \mathcal{M}(x) \downarrow\}$$

è ricorsivamente enumerabile ma non ricorsivo. Infatti, se $\mathcal{M}$ termina su x, allora $\mathcal{U}(\mathcal{M}; x) = \mathcal{M}(x) =$ yes, altrimenti $\mathcal{U}(\mathcal{M}; x) \uparrow$. Questo è un risultato qualitativo.

Esempio Sia

$$L = \{\text{bin}(\mathcal{M}) \mid \forall x \ \mathcal{M}(x) \downarrow \text{ in al massimo 100 passi}\}$$

L è ricorsivo. Infatti, la macchina $\mathcal{M}$ può eseguire al massimo 100 spostamenti a destra sul nastro. Quindi, tutte le macchine che terminano in al massimo 100 passi accettano input $\forall x \in |\Sigma|^n$ con $n \leq 100$.

Definizione 6.3.3 (Computazione di Funzioni) *Sia f una funzione $f : (\Sigma \backslash \sqcup)^* \to \Sigma^*$. Una macchina di Turing $\mathcal{M}$ computa f se*

$$\forall x \in (\Sigma \backslash \sqcup)^* \qquad \mathcal{M}(x) \downarrow \ \textit{e alla fine } f(x) \textit{ è sul nastro}$$

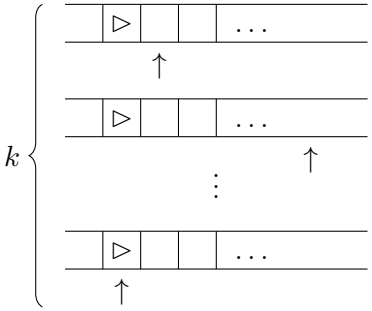
La funzione f è detta ***ricorsiva****, o* ***computabile****.*

6.4 Macchine di Turing a k-nastri e Input/Output

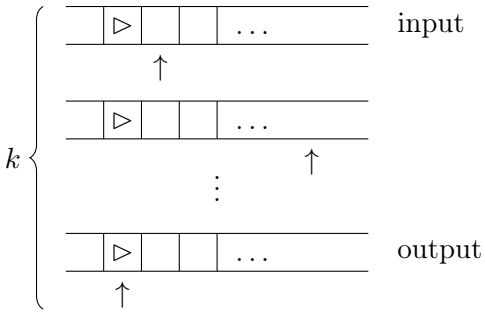
Definizione 6.4.1 (Macchina di Turing a k-nastri) *Una macchina di Turing a k-nastri è una tupla $\mathcal{M} = (K, \Sigma, \delta, s)$ con K, Σ, s definite come per una macchina di Turing, e*

$$\delta : K \times \Sigma \to (K \cup \{yes, no, halt\}) \times (\Sigma \times \{\leftarrow, \rightarrow, -\})^k$$

Una macchina di Turing a k-nastri è una macchina di Turing con un numero limitato di nastri, che possono essere utilizzati in parallelo. La funzione δ cambia perché si ha un puntatore per nastro.



Definizione 6.4.2 (Macchina di Turing a k-nastri con Input/Output) *Una macchina di Turing a k-nastri con I/O è una macchina di Turing a k-nastri con un nastro di input e un nastro di output. Il nastro di input è di sola lettura, il nastro di output è di sola scrittura.*



Definizione 6.4.3 (Configurazione e Configurazione Iniziale) *Siano $w_i, u_i \in \Sigma^*$ stringhe. Una configurazione è una tupla*

$$(q, w_1, u_1, w_2, u_2, \ldots, w_k, u_k) \to (q', w_1', u_1', w_2', u_2', \ldots, w_k', u_k')$$

Una configurazione iniziale su input x è una tupla

$$(s, \triangleright, x, \triangleright, \varepsilon, \ldots, \triangleright, \varepsilon)$$

6.4.1 Complessità Spaziale

Definizione 6.4.4 (Complessità Spaziale per una MdT a k-nastri su input x) *Si ha che*

$$(s, \triangleright, x) \to^* (H, w_1, u_1, \ldots, w_k, u_k)$$

con $H = \{halt, yes, no\}$. Lo spazio utilizzato è

$$\sum_{i=1}^{k} |w_i| + |u_i|$$

Definizione 6.4.5 (Complessità Spaziale per una MdT a k-nastri con I/O) *Si ha che*

$$(s, \triangleright, x) \to^* (H, w_1, u_1, \ldots, w_k, u_k)$$

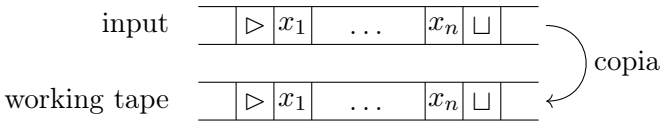
con $H = \{halt, yes, no\}$. Lo spazio utilizzato è

$$\sum_{i=1}^{k-1} |w_i| + |u_i|$$

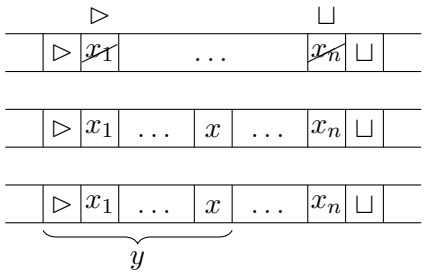
dove $\delta(q, \sigma_1, \ldots, \sigma_k) = (q', \sigma_1', \ldots, \sigma_k', \to)$.

Definizione 6.4.6 (Classi di Complessità Spaziale) *L è decidibile in spazio $f(n)$ se esiste una macchina di Turing a k-nastri con I/O $\mathcal{M}$ che decide L e, $\forall x$, $\mathcal{M}$ utilizza uno spazio al massimo $f(|x|)$.*

Esempio: palindromo $L = \{x | x \text{ è palindroma}\}$. Si vuole trovare la macchina più efficiente in termini di spazio. La seguente macchina è efficiente nel tempo:



perché ha TIME $\Theta(n)$ e SPACE $\Theta(n)$. Mentre la seguente macchina è efficiente nello spazio:



con SPACE $\Theta(\log n)$.

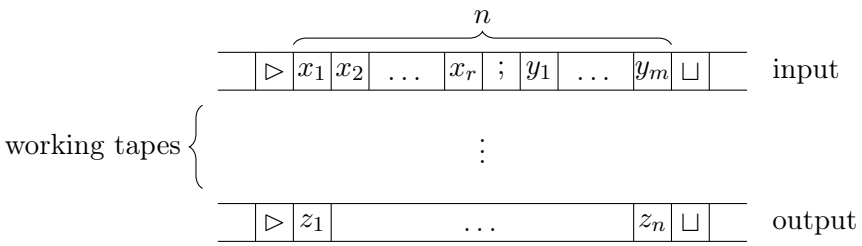
Definizione 6.4.7

$$TIME(f(n)) = \{L \mid L \text{ può essere deciso in tempo } f(n)\}$$
$$SPACE(f(n)) = \{L \mid L \text{ può essere deciso in spazio } f(n)\}$$

In altre parole, SPACE($f(n)$) è l'insieme di tutti i linguaggi che possono essere decisi in tempo $f(n)$ da una macchina di Turing a k-nastri con I/O. Per ogni input x tale che $|x| = n$, la macchina utilizza spazio al più $f(n)$.

Proprietà 6.4.1 *Se esiste una macchina di Turing che decide L in tempo $f(n)$, e $f(n) \geq n$, allora esiste una macchina di Turing con I/O che decide L in tempo $O(f(n))$.*

Esempio Calcola $x + y$.



Questo ha spazio lineare $\Theta(n)$ (molto male).

Definizione 6.4.8 (Classe P) *Definiamo la classe P come*

$$P = \bigcup_{h \in \mathbb{N}} TIME(n^h)$$

ovvero l'unione di tutti i problemi che possono essere risolti in tempo polinomiale.

La classe P ci piace così tanto perché abbiamo la tesi di Church-Turing estesa. Questa classe è **invariante** rispetto alla scelta del modello di computazione. Possiamo definire la classe EXP

$$\text{EXP} = \bigcup_{h \in \mathbb{N}} \text{TIME}(2^{n^h})$$

La classe $\mathbb{L}$, PSPACE, e EXPSPACE

$$\mathbb{L} = \text{SPACE}(\log n)$$

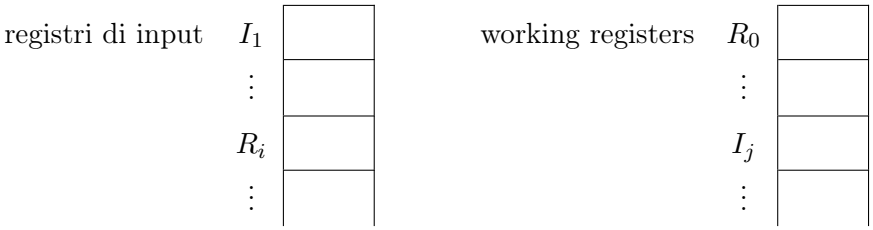
$$\text{PSPACE} = \bigcup_{h \in \mathbb{N}} \text{SPACE}(n^h)$$
$$\text{EXPSPACE} = \bigcup_{h \in \mathbb{N}} \text{SPACE}(2^{n^h})$$

Proprietà 6.4.2

$$TIME(f(n)) \subseteq SPACE(f(n))$$

6.5 Random Access Machines

Capitolo 2.6 del libro. Le random access machine (RAM), sono un modello di computazione sequenziale, composte da registri di input e registri di lavoro. Ogni registro contiene un intero.



Le operazioni possibili sono:

- `READ` j: $r_0 := i_j$
- `READ` $\uparrow j$: $r_0 := i_{r_j}$ (vai al registro R_j, leggine il contenuto h, vai al registro I_h, copiane il contenuto in R_0)
- `STORE` j: $r_j := r_0$
- `STORE` $\uparrow j$
- `LOAD` j: $r_0 := r_j$
- `LOAD` $\uparrow j$
- `LOAD` $= j$: $r_0 := j$
- `ADD` j: $r_0 := r_0 + r_j$
- `ADD` $\uparrow j$: $r_0 := r_0 + r_{r_j}$
- `ADD` $= j$
- `SUB` j
- ...
- `HALF`: $r_0 := \left\lfloor \frac{r_0}{2} \right\rfloor$ (tolgo da r_0 l'ultimo bit)
- `JUMP` j: $k := j$ (contatore)
- `JPOS` j: if $r_0 > 0$ then $k := j$
- `JNEG` j
- `JZERO` j
- `HALT`

Il libro dimostra che

Teorema 6.5.1 *RAM con complessità temporale uniforme e macchine di Turing con k-nastri sono correlate polinomialmente.*

In particolare

$$\underbrace{\text{MdT}}_{f(n)} \to^{\text{simula}} \underbrace{\text{RAM}}_{O(f(n))}$$
$$\underbrace{\text{RAM}}_{f(n)} \to^{\text{simula}} \underbrace{\text{MdT a 7-nastri}}_{O((f(n))^3)}$$

Ad esempio, quando si sommano due numeri, si ottiene al massimo 1 bit in più dell'input maggiore.

6.6 Macchine Nondeterministiche

Si hanno

- Macchine deterministiche $\mathcal{M} = (K, \Sigma, \delta, s)$, con δ **funzione**

  $$\delta : K \times \Sigma^k \to (K \cup \{\text{yes}, \text{no}, \text{halt}\}) \times \Sigma^k \times \{\leftarrow, \rightarrow, -\}^k$$

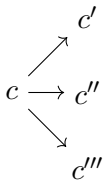
  la cui configurazione è del tipo

  $$c \to c'$$

- Macchine nondeterministiche $\mathcal{M} = (K, \Sigma, \Delta, s)$, con Δ **relazione**

  $$\Delta \subseteq K \times \Sigma^k \times (K \cup \{\text{yes}, \text{no}, \text{halt}\}) \times \Sigma^k \times \{\leftarrow, \rightarrow, -\}^k$$

  quindi con una o più possibili transizioni. La configurazione $(q, u_1, w_1, \ldots, q_k, w_k)$ è del tipo



Esempio: Reachability Problem Dato un grafo diretto $G = (V, E)$, e due nodi $u, v \in V$, decidere se u raggiunge v (se esiste un cammino da u a v). Studiamo la complessità in tempo e spazio di questo problema utilizzando sia un modello deterministico che nondeterministico.

Modello deterministico Un possibile algoritmo per risolvere questo problema è BFS(G, u). Si costruisce un albero con radice u, e ad ogni livello si aggiungono i nodi raggiungibili da u in un passo. Se v è raggiungibile da u, allora v sarà raggiunto da u in un numero di passi $\leq |V|$. Quindi, la complessità in tempo è $O(|V| + |E|)$, e in spazio $O(|V|)$.