

Appunti di Reti di Telecomunicazioni

Capitolo 2 – Catene di Markov

CATENE DI MARKOV TEMPO-DISCRETE	2
<i>Definizioni introduttive</i>	2
<i>Probabilità di transizione ad un passo</i>	3
<i>Catene di Markov omogenee</i>	4
<i>Matrice delle probabilità di transizione ad un passo</i>	4
Proprietà	5
<i>Matrice delle probabilità di transizione ad n passi</i>	6
<i>Probabilità di stato e probabilità asintotiche</i>	8
Esempio	10
Esempio	12
Catene di Markov completamente regolari	14
Esempio	14
Esempio di catena di Markov periodica.....	15
<i>Metodo di calcolo delle probabilità asintotiche</i>	16
Esempio	17
Esempio: sistema multiprocessore	18
<i>Definizioni varie sulle catene di Markov</i>	22
<i>Esempio: protocollo Slotted Aloha</i>	24
Ulteriori osservazioni sul protocollo.....	27
Teorema 1 (per catene di Markov omogenee e irriducibili).....	29
Teorema 2 (per catene di Markov omogenee, irriducibili e non periodiche)	29
Caso particolare.....	29
CATENE DI MARKOV TEMPO-CONTINUE	29
<i>Introduzione</i>	29
Il processo telegrafico casuale	30
<i>Le probabilità asintotiche</i>	30
Introduzione	30
Le probabilità di transizione ad un passo	31
Tempo di permanenza in uno stato	33
Le frequenze di transizione di stato	34
Determinazione delle probabilità asintotiche.....	36
Esempio: calcolatore con multiprogrammazione	39

Catene di Markov tempo-discrete

Definizioni introduttive

Sappiamo che un **processo stocastico** è un processo per cui, dato un certo spazio degli eventi S relativo ad un determinato fenomeno, noi associamo, a ciascun elemento di S , una funzione (a valori discreti o continui) del tempo t . Allora diamo la seguente definizione:

Def. Un processo stocastico si dice "**markoviano**" quando, scelto un certo istante t di osservazione, l'evoluzione del processo, a partire da t , dipende solo da t mentre non dipende in alcun modo dagli istanti precedenti

Detto anche in altri termini, un processo è markoviano quando, dato l'istante di osservazione, solo tale istante determina l'evoluzione futura del processo, mentre tale evoluzione non dipende dal passato.

Volendo tradurre questa definizione in formule, possiamo procedere nel modo seguente: consideriamo gli istanti $t_1 < t_2 < \dots < t_k < t_{k+1}$; in corrispondenza di ciascuno di tali istanti, estraiamo dal processo una variabile aleatoria; indichiamo tali variabili aleatorie con $X(t_1), X(t_2), \dots, X(t_k), X(t_{k+1})$; supponiamo inoltre che il processo sia a valori discreti, il che significa che tali variabili aleatorie possono assumere solo valori (o "*stati*") che sono in numero finito oppure infinito numerabile. Ciascuna di queste variabili aleatorie contiene in pratica in se le caratteristiche del processo nell'istante in cui è stata estratta: per esempio, le caratteristiche del processo all'istante t_{k+1} sono contenute nella variabile $X(t_{k+1})$; queste caratteristiche possono o meno dipendere da quelle che sono state le caratteristiche del processo negli istanti precedenti; un legame esistente tra queste caratteristiche è senz'altro rappresentato dal valore di

$$P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k \cap X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1)$$

Infatti, questa quantità rappresenta la probabilità che il sistema, all'istante t_{k+1} , si trovi nello stato x_{k+1} , noti che siano gli stati x_1, x_2, \dots, x_k in cui si trovava negli istanti precedenti.

Adesso supponiamo che l'istante di osservazione prescelto sia l'istante t_k : allora, noi diremo che il nostro processo è markoviano se accade che

$$P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k \cap X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1) = P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k)$$

ossia se lo stato del sistema ad un qualsiasi istante successivo a quello di osservazione dipende solo dallo stato del sistema all'istante di osservazione.

Questa è la definizione relativa al caso in cui il processo è a valori discreti: in particolare, un processo markoviano a valori discreti prende il nome di "**catena di Markov**" ed è quello di cui ci occuperemo in questo capitolo.

N.B. Esiste tuttavia una analoga definizione relativa ai processi stocastici a valori continui: noi diremo che un processo stocastico a valori continui è markoviano se e solo se risulta verificata la relazione

$$\begin{aligned} P(X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_k \cap X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1) = \\ = P(X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_k) \end{aligned}$$

Proviamo per esempio a vedere se il *processo telegrafico casuale*, che è un noto processo stocastico tempo-continuo a valori discreti, è una catena di Markov: la risposta è evidentemente affermativa, in quanto sappiamo che i cambiamenti di stato in tale processo sono condizionati all'avvenire di eventi di Poisson e ogni evento di Poisson dipende solo dall'evento immediatamente precedente e non da tutti gli altri ⁽¹⁾.

Ad ogni modo, il processo telegrafico casuale rientra nella categorie delle **catene di Markov tempo-continue** (brevemente **CTMC**), in quanto sappiamo che la realizzazioni del processo telegrafico casuale sono funzioni continue del tempo. Viceversa, noi siamo per il momento interessati alle cosiddette **catene di Markov tempo-discrete** (brevemente **DTMC**), tali cioè che le loro realizzazioni siano funzioni discrete del tempo, ossia assumano dei valori solo in istanti di tempo prefissati (che possono essere o meno multiple di una quantità fissa).

Quindi, riepilogando, diciamo che *una catena di Markov tempo-discreta è un processo stocastico, tempo-discreto e a valori discreti, tale che l'evoluzione del processo a partire dall'istante di osservazione dipende solo dallo stato in cui si trova il sistema nell'istante di osservazione.*

Il fatto che si tratti di processi stocastici tempo-discreti ci consente di semplificare le nostre notazioni: infatti, anziché indicare con $X(t_1), X(t_2), \dots, X(t_N)$, le variabili estratte dal processo nei vari istanti, le indichiamo semplicemente con X_1, X_2, \dots, X_N .

Probabilità di transizione ad un passo

Al fine di caratterizzare ancora meglio una catena di Markov, valutiamo la seguente quantità:

$$P(X_N = x_N \cap X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0)$$

Si tratta della probabilità che il processo (o "*sistema*") assuma il valore (o "*stato*") x_0 all'istante $t=t_0$, il valore x_1 all'istante t_1 e così via fino al valore x_N all'istante t_N .

Essendo una *probabilità assoluta* (dove l'evento considerato è dato dall'intersezione di più eventi), possiamo valutarla facendo uso delle *probabilità condizionate*: la indichiamo, per comodità, con P_C , e otteniamo allora che

$$P_C = P(X_N = x_N | X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0) P(X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0)$$

¹ Ricordiamo, a tal proposito, quanto dimostrato nel corso di Teoria dei Segnali: il tempo di attesa ed il tempo di arrivo, in un processo telegrafico casuale, sono delle variabili aleatorie con distribuzione esponenziale, ossia prive di memoria.

Essendo il processo per ipotesi markoviano, possiamo semplificare quella probabilità condizionata, restringendo la condizione al solo istante precedente quello di osservazione:

$$P_C = P(X_N = x_N | X_{N-1} = x_{N-1}) P(X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0)$$

Adesso facciamo lo stesso discorso con $P(X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0)$: otteniamo che

$$\begin{aligned} P_C &= P(X_N = x_N | X_{N-1} = x_{N-1}) P(X_{N-1} = x_{N-1} | X_{N-2} = x_{N-2} \dots \cap X_0 = x_0) P(X_{N-2} = x_{N-2} \cap \dots \cap X_0 = x_0) = \\ &= P(X_N = x_N | X_{N-1} = x_{N-1}) P(X_{N-1} = x_{N-1} | X_{N-2} = x_{N-2}) P(X_{N-2} = x_{N-2} \cap \dots \cap X_0 = x_0) \end{aligned}$$

Proseguendo in questo modo, arriviamo al seguente risultato finale:

$$P_C = P(X_N = x_N | X_{N-1} = x_{N-1}) P(X_{N-1} = x_{N-1} | X_{N-2} = x_{N-2}) \dots P(X_1 = x_1 | X_0 = x_0) P(X_0 = x_0)$$

La generica quantità $P(X_k = x_k | X_{k-1} = x_{k-1})$ prende il nome di **probabilità di transizione ad un passo**: essa rappresenta la probabilità che il sistema passi dallo stato x_{k-1} allo stato x_k nell'intervallo di tempo $[t_{k-1}, t_k]$. Ovviamente, quindi, per **passo** si intende l'intervallo di tempo che intercorre tra due cambiamenti di stato successivi.

Possiamo dunque scrivere che

$$P(X_N = x_N \cap X_{N-1} = x_{N-1} \cap \dots \cap X_1 = x_1 \cap X_0 = x_0) = P(X_0 = x_0) \prod_{k=1}^N P(X_k = x_k | X_{k-1} = x_{k-1})$$

Catene di Markov omogenee

La generica probabilità di transizione ad un passo $P(X_k = x_k | X_{k-1} = x_{k-1})$ può o meno dipendere dall'ampiezza $t_k - t_{k-1}$ del passo stesso nonché. Se dipende solo da tale ampiezza, noi diciamo che la catena di Markov (tempo-discreta) è **omogenea**; in caso contrario, si parla ovviamente di catena di Markov **non omogenea**.

Da questo momento in poi, supporremo di avere a che fare solo con catene di Markov tempo-discrete omogenee.

Matrice delle probabilità di transizione ad un passo

Consideriamo ancora una volta la generica probabilità di transizione ad un passo

$$P(X_k = x_k | X_{k-1} = x_{k-1})$$

Gli stati x_k e x_{k-1} sono da ricercarsi tra tutti i possibili stati che il sistema può assumere; per esempio, se si trattasse di un processo telegrafico casuale, avremo

solo +1 o -1, mentre, più in generale, possiamo avere un numero di stati finito oppure infinito numerabile (a seconda di quali sono i valori assumibili da parte del processo, ossia da parte delle realizzazioni che costituiscono il processo).

Per semplificare le nostre notazioni, poniamo $j=x_k$ e $i=x_{k-1}$ e poniamo inoltre

$$p_{ij} = P(X_k = j | X_{k-1} = i)$$

Questa è dunque la probabilità che il processo si trovi nello stato j all'istante t_k quando all'istante t_{k-1} si trovava nello stesso i .

E' chiaro che, a seconda di quanti sono gli stati che il processo può assumere, potremo avere un numero finito o infinito numerabile di probabilità di transizione ad un passo. E' conveniente allora costruire, con tali probabilità, una matrice fatta nel modo seguente:

$$[P] = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0m} \\ p_{10} & p_{11} & \cdots & p_{1m} \\ \cdots & & & \\ p_{m0} & p_{m1} & \cdots & p_{mm} \end{bmatrix}$$

Ogni elemento della matrice corrisponde ad una diversa probabilità di transizione ad un passo: in particolare

- gli elementi della riga i° corrispondono alle probabilità di transizione ad un passo con stato iniziale i ;
- gli elementi della colonna j° corrispondono alle probabilità di transizione ad un passo con stato finale j .

La dimensione della matrice è m , dove m è il numero di stati possibili per il sistema: se tale numero è finito, la matrice è *determinata*; se invece esso è infinito numerabile, allora la matrice è *indefinita*.

La matrice $[P]$ prende il nome di **matrice delle probabilità di transizione ad un passo**.

Proprietà

Si può dimostrare in modo del tutto intuitivo la seguente proprietà di cui gode la matrice $[P]$:

$$\sum_{j=0}^m p_{ij} = 1 \quad i = 0, 1, \dots, m$$

Questa proprietà dice in pratica che la somma degli elementi su di una riga è pari ad 1. Si dice allora che $[P]$ è una **matrice stocastica**.

La dimostrazione è immediata: essendo $p_{ij} = P(X_k = j | X_{k-1} = i)$, abbiamo che

$$\sum_{j=0}^m p_{ij} = \sum_{j=0}^m P(X_k = j | X_{k-1} = i) = P(X_k = 0 | X_{k-1} = i) + P(X_k = 1 | X_{k-1} = i) + \dots + P(X_k = m | X_{k-1} = i)$$

La sommatoria in questione rappresenta dunque la probabilità che il sistema, trovandosi nel generico stato i all'istante t_{k-1} , passi in uno qualsiasi degli stati (incluso i stesso) all'istante successivo ed è ovvio che tale probabilità valga 1.

Matrice delle probabilità di transizione ad n passi

Abbiamo detto che il generico elemento $p_{ij} = P(X_k = j | X_{k-1} = i)$ della matrice delle probabilità di transizione ad un passo indica la probabilità che il sistema passi dallo stato i allo stato j nell'intervallo di tempo $[t_{k-1}, t_k]$. In modo del tutto analogo, è anche possibile valutare la probabilità che il sistema si trovi nello stato j all'istante t_k dopo che all'istante t_{k-2} si trovava nello stato i :

$$p_{ij}(2) = P(X_k = j | X_{k-2} = i)$$

Si tratta cioè della probabilità che il sistema, trovandosi nello stato i all'istante t_{k-2} , dopo due passi, quindi dopo un intervallo di tempo di ampiezza $t_k - t_{k-2}$, si trovi nello stato j .

In generale, è possibile definire la cosiddetta **probabilità di transizione ad n passi** come la seguente probabilità:

$$p_{ij}(n) = P(X_k = j | X_{k-n} = i)$$

Si tratta cioè della probabilità che il sistema, all'istante t_k , si trovi nello stato j dopo che, n passi prima, cioè n intervalli di tempo prima, si trovava nello stato i .

Per queste nuove probabilità di transizione valgono le stesse considerazioni fatte per quelle ad un unico passo, nel senso che, a seconda di quanti sono gli stati che il processo può assumere, noi potremo avere un numero finito o infinito numerabile di probabilità di transizione. E' dunque ancora una volta conveniente costruire, con tali probabilità, una matrice fatta nel modo seguente:

$$[P(n)] = \begin{bmatrix} p_{00}(n) & p_{01}(n) & \dots & p_{0m}(n) \\ p_{10}(n) & p_{11}(n) & \dots & p_{1m}(n) \\ \dots & \dots & \dots & \dots \\ p_{m0}(n) & p_{m1}(n) & \dots & p_{mm}(n) \end{bmatrix}$$

Questa, in analogia a prima, prende il nome di **matrice delle probabilità di transizione ad n passi**.

Vediamo adesso se è possibile trovare un legame tra questa matrice e la matrice di transizione ad un passo $[P]$ (che poi, in modo formalmente più corretto, dovrebbe essere indicata con $[P(1)]$).

Supponiamo per il momento che sia $n=2$: il generico elemento della matrice $[P(2)]$ è

$$p_{ij}(2) = P(X_k = j | X_{k-2} = i)$$

Questa probabilità condizionata è esprimibile mediante la nota formula

$$p_{ij}(2) = \frac{P(X_k = j \cap X_{k-2} = i)}{P(X_{k-2} = i)}$$

Possiamo adesso applicare il *teorema delle probabilità totali* per riscrivere in altra forma il numeratore di quella frazione: applicando il teorema con riferimento al passo $[t_{k-2}, t_{k-1}]$, abbiamo che

$$p_{ij}(2) = \frac{\sum_n P(X_k = j \cap X_{k-2} = i \cap X_{k-1} = n)}{P(X_{k-2} = i)}$$

Esprimendo ancora una volta le probabilità al numeratore mediante le probabilità condizionate, abbiamo inoltre che

$$p_{ij}(2) = \frac{\sum_n P(X_k = j | X_{k-2} = i \cap X_{k-1} = n) P(X_{k-1} = n \cap X_{k-2} = i)}{P(X_{k-2} = i)}$$

Avendo a che fare con un processo di Markov, quella probabilità condizionata si può semplificare, limitando la condizione al solo istante t_{k-1} :

$$p_{ij}(2) = \frac{\sum_n P(X_k = j | X_{k-1} = n) P(X_{k-1} = n \cap X_{k-2} = i)}{P(X_{k-2} = i)}$$

Ricordando che la generica probabilità di transizione ad un passo è $p_{ij} = P(X_k = j | X_{k-1} = i)$, possiamo riscrivere quella relazione come

$$p_{ij}(2) = \frac{\sum_n p_{nj} P(X_{k-1} = n \cap X_{k-2} = i)}{P(X_{k-2} = i)}$$

Ora esprimiamo il termine $P(X_{k-1} = n \cap X_{k-2} = i)$ ancora una volta mediante le probabilità condizionate:

$$p_{ij}(2) = \frac{\sum_n p_{nj} P(X_{k-1} = n | X_{k-2} = i) P(X_{k-2} = i)}{P(X_{k-2} = i)}$$

Il termine $P(X_{k-2}=1)$ al numeratore non dipende dall'indice n della sommatoria, per cui possiamo portarlo fuori e semplificarlo con il denominatore: resta dunque

$$p_{ij}(2) = \sum_n p_{nj} P(X_{k-1} = n | X_{k-2} = i)$$

Infine, il termine $P(X_{k-1} = n | X_{k-2} = i)$ non è altro che p_{in} , per cui

$$p_{ij}(2) = \sum_n p_{nj} p_{in}$$

Questa relazione dice in pratica che il generico termine $p_{ij}(2)$ della matrice $[P(2)]$ è pari al prodotto tra i seguenti vettori:

$$p_{ij}(2) = \sum_n p_{nj} p_{in} = \sum_n p_{in} p_{nj} = [p_{i1} \mid p_{i2} \mid \dots \mid p_{im}] \begin{bmatrix} p_{1j} \\ p_{2j} \\ \dots \\ p_{mj} \end{bmatrix}$$

Il vettore $[p_{i1} \mid p_{i2} \mid \dots \mid p_{im}]$ non è altro che la i° riga della matrice $[P]$, mentre il vettore $[p_{1j} \mid p_{2j} \mid \dots \mid p_{mj}]^T$ è la j° colonna della matrice $[P]$. Si deduce quindi che la matrice $[P(2)]$ non è altro che il prodotto righe*colonne tra la matrice $[P]$ e se stessa, ossia

$$[P(2)] = [P] * [P] = [P]^2$$

Questo risultato è del tutto generale, nel senso che si trova

$$[P(n)] = [P]^n$$

Quindi, nota la matrice delle probabilità di transizione ad 1 solo passo, noi possiamo calcolarci la matrice delle probabilità di transizione per un qualsiasi numero di passi.

Probabilità di stato e probabilità asintotiche

Sia X_k la variabile aleatoria che fornisce lo stato del processo all'istante t_k : si definisce **probabilità asintotica** la quantità

$$\pi_j = \lim_{k \rightarrow \infty} P(X_k = j)$$

Che cosa rappresenta questa quantità? Il termine $P(X_k=j)$ è una **probabilità di stato**, ossia la probabilità che il sistema, nell'istante t_k , si trovi nello stato j ; di conseguenza, è chiaro che π_j rappresenta la probabilità che il sistema, a regime,

ossia per $k \rightarrow \infty$, si trovi nello stato j . Il nostro scopo diventa adesso quello di valutare quanto vale questa probabilità asintotiche.

La prima cosa da dire è che non sempre queste proprietà esistono; sussiste allora la seguente definizione: *una catena di Markov si dice "regolare" quando esistono le probabilità asintotiche.*

Nell'ipotesi che la catena da noi considerata sia regolare, andiamo a valutare prima le probabilità di stato e successivamente quelle asintotiche.

Intanto, applicando il teorema delle probabilità totali e, successivamente, le formule delle probabilità condizionate, possiamo scrivere che

$$P(X_k = j) = \sum_i P(X_k = j \cap X_{k-1} = i) = \sum_i P(X_k = j | X_{k-1} = i) P(X_{k-1} = i)$$

Adesso, ricordando che la generica probabilità di transizione ad un passo è $p_{ij} = P(X_k = j | X_{k-1} = i)$, possiamo riscrivere quella relazione come

$$P(X_k = j) = \sum_i p_{ij} P(X_{k-1} = i) \quad \mathbf{(a)}$$

Con lo stesso identico discorso, ma scegliendo una diversa partizione con cui applicare il teorema delle probabilità totali, noi possiamo esprimere il termine $P(X_k=j)$ anche nella forma

$$P(X_k = j) = \sum_i p_{ij} P(X_0 = i) \quad \mathbf{(b)}$$

Vediamo come possiamo utilizzare le relazioni (a) e (b).

La relazione (a) indica che il termine $P(X_k=j)$ è dato dal seguente prodotto:

$$P(X_k = j) = \sum_i p_{ij} P(X_{k-1} = i) = [p_{1j} \mid p_{2j} \mid \dots \mid p_{mj}] \begin{bmatrix} P(X_{k-1} = 0) \\ P(X_{k-1} = 1) \\ \dots \\ P(X_{k-1} = m) \end{bmatrix}$$

Il vettore riga $[p_{1j} \mid p_{2j} \mid \dots \mid p_{mj}]$ è la j° colonna della matrice di transizione ad un passo $[P]$, mentre il vettore per cui è moltiplicato può essere chiamato col nome di **vettore delle probabilità di stato del processo all'istante t_{k-1}** , in quanto esso contiene le probabilità che il sistema, all'istante t_{k-1} , si trovi in uno qualsiasi degli stati permessi.

Poniamo allora $\mathbf{p_j(k)} = \mathbf{P(X_k=j)}$, per cui la relazione di prima diventa

$$\mathbf{p_j(k)} = [p_{1j} \mid p_{2j} \mid \dots \mid p_{mj}] \begin{bmatrix} p_0(k-1) \\ p_1(k-1) \\ \dots \\ p_m(k-1) \end{bmatrix} \quad j = 0, 1, \dots, m$$

Al variare di j in $\{0,1,\dots,m\}$, abbiamo a primo membro un vettore di dimensione $m+1$ ed al secondo membro il prodotto della matrice $[P]$ per un vettore colonna di dimensione $m+1$: allora, ponendo

$$[p(k)] = \begin{bmatrix} p_0(k) \\ p_1(k) \\ \dots \\ p_m(k) \end{bmatrix} \quad \text{e} \quad [p(k-1)] = \begin{bmatrix} p_0(k-1) \\ p_1(k-1) \\ \dots \\ p_m(k-1) \end{bmatrix}$$

possiamo riscrivere quella relazione nella forma

$$\boxed{[p(k)] = [P][p(k-1)]}$$

Con un discorso del tutto analogo, fatto a partire dalla relazione (b), si trova facilmente che

$$\boxed{[p(k)] = [P]^k [p(0)]}$$

Queste relazioni ci dicono quanto segue:

- in primo luogo, dato il vettore delle probabilità di stato relativo ad un certo istante, possiamo ottenere quello relativo all'istante successivo semplicemente moltiplicando a sinistra per la matrice di transizione ad un passo;
- in modo analogo, la seconda relazione dice che, *noto il vettore delle probabilità di stato all'istante iniziale t_0 , possiamo conoscere il vettore delle probabilità di stato di un istante successivo generico t_k semplicemente moltiplicando a sinistra per $[P]^k$.*

A questo punto, dato il vettore $[p(k)]$ delle probabilità di stato all'istante t_k , le probabilità asintotiche si possono calcolare facendo il limite per $k \rightarrow \infty$.

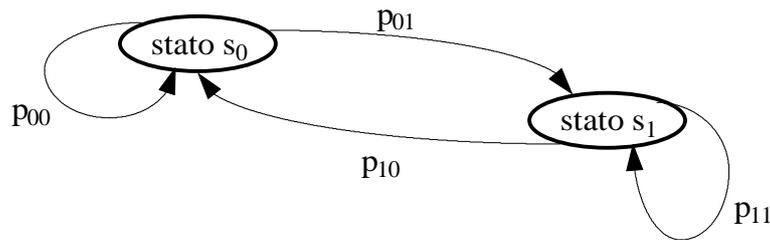
Esempio

Supponiamo di avere una catena di Markov omogenea a due soli stati, che indichiamo con s_0 e s_1 . Supponiamo anche di conoscere la matrice delle probabilità di transizione ad un passo (che sarà ovviamente di dimensione 2×2 in quanto 2 sono gli stati possibili per la catena):

$$[P] = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Si nota immediatamente che la somma degli elementi su ciascuna riga di questa matrice è pari ad 1.

Da un punto di vista schematico, possiamo dare di questa matrice una rappresentazione grafica in termini di **grafo orientato**:



Questo grafo consente di rappresentare in modo molto rapido le probabilità che il sistema, nell'arco di un solo passo, ossia di un solo intervallo di tempo, passi da uno stato all'altro tra quelli consentiti. Facciamo anche osservare come questo passo sia assolutamente generico, in quanto siamo sempre nell'ipotesi che la catena di Markov sia omogenea.

Proviamo allora a calcolare la matrice delle probabilità di transizione a 2 passi: possiamo evidentemente utilizzare la relazione

$$[P(n)] = [P]^n$$

Per $n=2$ noi otteniamo che

$$[P(2)] = [P] * [P] = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix} \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 5/4 & 3/4 \\ 3/4 & 5/4 \end{bmatrix}$$

La matrice $[P(2)]$ può anche essere riscritta nella forma

$$[P(2)] = \begin{bmatrix} 1 + \left(\frac{1}{2}\right)^2 & 1 - \left(\frac{1}{2}\right)^2 \\ 1 - \left(\frac{1}{2}\right)^2 & 1 + \left(\frac{1}{2}\right)^2 \end{bmatrix}$$

Questa forma è utile in quanto si trova, applicando ripetutamente la formula $[P(n)] = [P]^n$, che

$$[P(n)] = \begin{bmatrix} 1 + \left(\frac{1}{2}\right)^n & 1 - \left(\frac{1}{2}\right)^n \\ 1 - \left(\frac{1}{2}\right)^n & 1 + \left(\frac{1}{2}\right)^n \end{bmatrix}$$

Adesso supponiamo che il vettore delle probabilità di stato, all'istante $t=t_0$, sia il seguente:

$$[p(0)] = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$$

Questo vettore indica, in pratica, che la catena, all'istante $t=t_0$, si trova nello stato s_0 con probabilità $1/3$ e nello stato s_1 con probabilità $2/3$.

Sulla base di questo dato possiamo andare a calcolare il vettore delle probabilità di stato all'istante generico $t=t_n$: infatti, applicando la formula $[p(n)] = [P]^n [p(0)]$, si trova che

$$[p(n)] = \frac{1}{6} \begin{bmatrix} 3 - \left(\frac{1}{2}\right)^n \\ 3 + \left(\frac{1}{2}\right)^n \end{bmatrix}$$

Si nota dunque che il vettore delle probabilità di stato relativo all'istante generico $t=t_n$ dipende dal passo n , il che è indice del fatto che la catena non sia stazionaria ⁽²⁾.

Al contrario, si può verificare che, se il vettore delle probabilità di stato relative all'istante iniziale $t=t_0$ fosse

$$[p(0)] = [p_0] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

risulterebbe

$$[p(n)] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

e quindi la catena sarebbe in questo caso stazionaria.

Per quanto riguarda le probabilità asintotiche, sono evidenti i loro valori:

- cominciamo dal primo caso, quello cioè in cui la catena non è stazionaria: calcolando il limite, per $n \rightarrow \infty$, del vettore delle probabilità di stato al generico passo n , abbiamo che

$$[p(\infty)] = \lim_{n \rightarrow \infty} [p(n)] = \lim_{n \rightarrow \infty} \begin{bmatrix} 3 - \left(\frac{1}{2}\right)^n \\ 3 + \left(\frac{1}{2}\right)^n \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

- nel secondo caso, invece, in cui $[p(n)] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$, è chiaro che

$$[p(\infty)] = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} = [p(n)].$$

Esempio

Consideriamo una catena di Markov a 3 stati, che indichiamo con s_1, s_2 ed s_3 . Supponiamo anche qui di conoscere la matrice di transizione ad un passo:

² Ricordiamo che un processo stocastico è stazionario quando le sue caratteristiche sono costanti nel tempo, ossia quando non dipendono dal tempo stesso. Nell'esempio considerato, ciò non avviene, in quanto le probabilità di stato per l'istante t_n dipendono dal passo n , ossia sostanzialmente da $t_n - t_{n-1}$.

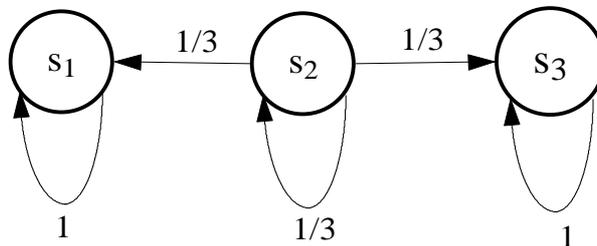
$$[P] = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1 \end{bmatrix}$$

La struttura di questa matrice ci indica una caratteristica particolare della catena: si nota infatti che

$$p_{12} = p_{13} = p_{31} = p_{32} = 0$$

Dire che $p_{12}=0$ significa dire che non è possibile che il sistema, nell'arco di un passo generico, passi dallo stato 1 allo stato 2. Stesso discorso, ovviamente, per le altre tre probabilità. Questo fatto indica che gli stati s_1 ed s_3 sono **stati assorbenti**, cioè sono tali che il sistema, una volta arrivato in uno dei due, non può più uscirvi.

Questa cosa risulta molto più chiara se rappresentiamo graficamente la matrice [P]:



Premesso, questo, andiamo a valutare la matrice delle probabilità di transizione ad n passi: utilizzando la solita relazione $[P(n)] = [P]^n$, si trova quanto segue:

$$[P(n)] = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) & \left(\frac{1}{3} \right)^n & \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) \\ 0 & 0 & 1 \end{bmatrix}$$

Adesso andiamo a calcolare il vettore delle probabilità di stato per il generico istante n : dobbiamo usare la relazione $[p(n)] = [P]^n [p(0)]$, per cui ci serve conoscere il vettore $[p(0)]$, ossia il vettore delle probabilità di stato iniziali. Supponiamo che sia

$$[p(0)] = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Facendo i conti, si trova che

$$[p(n)] = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) & \left(\frac{1}{3} \right)^n & \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 1/3 \\ \frac{1}{3} \left(1 - \left(\frac{1}{3} \right)^n \right) + \left(\frac{1}{3} \right)^{n+1} \\ 1/3 \end{bmatrix}$$

Calcolando adesso il limite per $n \rightarrow \infty$, abbiamo che

$$[p(\infty)] = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

ossia che le probabilità asintotiche coincidono con le probabilità di stato iniziali.

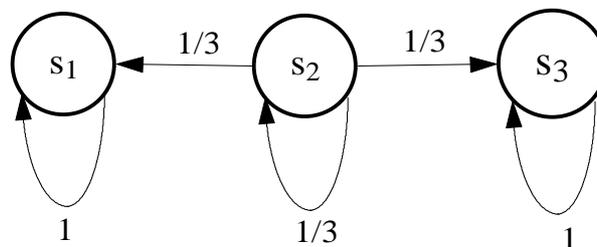
Catene di Markov completamente regolari

Abbiamo detto che una catena di Markov è *regolare* quando ammette le probabilità asintotiche. Dagli esempi precedenti si nota come, spesso, le probabilità asintotiche dipendano dalle probabilità di stato iniziali. Sussiste allora la seguente definizione: *una catena di Markov si dice **completamente regolare** quando ammette le probabilità asintotiche e queste NON dipendono dalle condizioni iniziali.*

Si dimostra (noi non lo facciamo), inoltre, il seguente teorema: *condizione necessaria e sufficiente affinché una catena di Markov sia completamente regolare è che la matrice delle probabilità di transizione ad n passi $[P(n)]$, per $n \in \mathbb{N}$ abbia tutte le righe uguali; in questo caso, ciascuna delle suddette righe rappresenta proprio la soluzione asintotica.*

Esempio

Consideriamo nuovamente la catena di Markov dell'ultimo esempio trattato, la cui matrice delle probabilità di transizione ad 1 passo aveva la seguente rappresentazione grafica:



Abbiamo trovato che la matrice delle probabilità di transizione ad n passi è

$$[P(n)] = \begin{bmatrix} \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) & 0 & \frac{0}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) \\ \left(\frac{1}{3} \right)^n & \left(\frac{1}{3} \right)^n & \frac{1}{2} \left(1 - \left(\frac{1}{3} \right)^n \right) \\ 0 & 0 & 1 \end{bmatrix}$$

Per $n \rightarrow \infty$ (cioè per un numero infinito di passi), questa matrice diventa

$$[P(\infty)] = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}$$

Questa matrice non dipende dal passo n .

Assumendo, come condizioni iniziali, quelle del vettore $[p(0)] = [1/3 \ 1/3 \ 1/3]^T$, abbiamo trovato che il vettore delle probabilità di stato, al generico passo n , è

$$[p(n)] = [P(n)][p(0)] = [P]^n [p(0)] = \begin{bmatrix} 1/3 \\ \frac{1}{3} \left(1 - \left(\frac{1}{3} \right)^n \right) + \left(\frac{1}{3} \right)^{n+1} \\ 1/3 \end{bmatrix}$$

Calcolando il limite per $n \rightarrow \infty$, otteniamo $[p(\infty)] = [1/3 \ 1/3 \ 1/3]^T$, ossia un vettore coincidente con quello delle condizioni iniziali.

Se però cambiassimo il vettore delle probabilità iniziali, prendendo ad esempio $[p(0)] = [1/2 \ 1/3 \ 1/6]^T$, troveremmo un diverso vettore $[p(\infty)]$, il che ci dice che la catena non è completamente regolare, ma solo regolare: essa ammette soluzione asintotica, ma tale soluzione dipende dalle condizioni iniziali.

Esempio di catena di Markov periodica

Consideriamo una catena di Markov (omogenea e tempo-discreta) avente 4 stati diversi. Supponiamo che la matrice delle probabilità di transizione ad un passo sia la seguente:

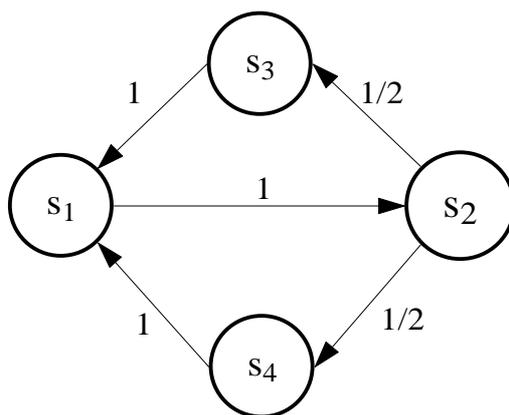
$$[P] = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Una catena di questo tipo, ossia avente una matrice $[P]$ di questo tipo, gode della proprietà di essere **periodica**: questo significa che, se il sistema si trova, in un certo istante, in un certo stato, tornerà sempre in tale stato dopo un certo numero costante di passi.

In altre parole, lo stato del sistema si ripete periodicamente: dato un certo stato ad un certo istante, il sistema torna nello stesso stato dopo un numero finito e costante di passi.

Nel nostro caso, dato un certo stato in cui il sistema si trova in un certo istante, ci vogliono sempre 3 passi perché il sistema torni nello stesso stato.

La cosa risulta evidente rappresentando ancora una volta graficamente la matrice $[P]$:



Supponiamo ad esempio di partire dallo stato s_1 : l'unica possibilità è quella di passare nello stato s_2 ; dallo stato s_2 , c'è un 50% di possibilità di passare ad s_3 ed un altro 50% di passare ad s_4 ; in ogni caso, però, sia da s_3 sia da s_4 , l'unica possibilità è di passare ad s_1 , per cui ci sono voluti 3 passi per tornare allo stato di partenza e ce ne vorranno evidentemente altri 3 per tornarvi ancora.

E' facile verificare che lo stesso risultato si ottiene partendo da uno qualsiasi degli altri 3 stati.

Il fatto che la catena sia periodica di "periodo 3" implica come conseguenza fondamentale che

$$[P] = [P(4)] = [P(7)] = [P(10)] = \dots$$

ossia in generale che

$$[P(k)] = [P(k + 3)] \quad \text{con } k = 0, 1, \dots, \infty$$

La seconda conseguenza è evidentemente il fatto per cui non esista una soluzione asintotica, il che significa che tale catena non è regolare.

Metodo di calcolo delle probabilità asintotiche

Fino ad ora abbiamo effettuato, ove possibile, il calcolo delle probabilità (di stato) asintotiche calcolando prima le probabilità di stato al generico passo n e poi calcolando il limite per $n \rightarrow \infty$. Esiste invece un metodo molto più immediato per il calcolo di queste probabilità. Tale metodo è applicabile solo alle catene di Markov che siano completamente regolari: la *regolarità* serve a garantire l'esistenza delle probabilità asintotiche, mentre la *completa regolarità* serve a garantire l'indipendenza di tali probabilità dalle probabilità di stato iniziali. Vediamo di che si tratta.

Abbiamo detto che il vettore delle probabilità di stato al generico istante t_n è legato a quello dell'istante precedente dalla relazione

$$[p(n)] = [P][p(n-1)]$$

Per calcolare le probabilità asintotiche, dobbiamo calcolare il limite per $n \rightarrow \infty$: quindi

$$\lim_{n \rightarrow \infty} [p(n)] = \lim_{n \rightarrow \infty} [P][p(n-1)] = [P] \lim_{n \rightarrow \infty} [p(n-1)]$$

E' abbastanza ovvio che, se poniamo $[\pi] = \lim_{n \rightarrow \infty} [p(n)]$, risulta anche $[\pi] = \lim_{n \rightarrow \infty} [p(n-1)]$, per cui la relazione di prima diventa

$$\boxed{[\pi] = [P] \cdot [\pi]}$$

Questa relazione è un sistema di equazioni lineari le cui incognite sono le probabilità asintotiche $[\pi_1, \pi_2, \dots, \pi_N]$, dove ovviamente N, che è il numero degli stati possibili per il sistema, può essere sia finito sia infinito numerabile.

Allora, a livello analitico, possiamo determinare le probabilità asintotiche della nostra catena laddove esistono, risolvendo semplicemente quel sistema. A livello numerico, invece, ossia usando un PC, tale procedimento è ovviamente valido solo quando N è un numero finito.

Osserviamo inoltre come le N equazioni rappresentate da quella relazione matriciale non siano però sufficienti per la determinazione delle probabilità asintotiche: infatti, si trova che una di esse deve essere necessariamente una combinazione lineare delle altre, il che significa che quel sistema fornisce ∞^1 soluzioni.

Allora, per determinare le probabilità asintotiche, si usa una ulteriore equazione che è la cosiddetta **condizione di normalizzazione**: essa dice che quelle probabilità asintotiche sono legate dalla relazione

$$\boxed{\sum_{i=1}^N \pi_i = 1}$$

ed è ovvio che sia così, in quanto, a regime, il sistema dovrà necessariamente trovarsi in uno degli stati possibili.

Esempio

Supponiamo di avere una catena di Markov tempo-discreta (DTMC) a 2 soli stati e supponiamo che la sua matrice delle probabilità di transizione ad un passo sia la seguente:

$$[P] = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Questa catena è stata già considerata in precedenza ed analizzata con altre tecniche.

Per trovare le probabilità asintotiche, dobbiamo risolvere il sistema rappresentato dalla relazione

$$\begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}$$

e dalla condizione di normalizzazione.

Scritto in forma scalare, il sistema risulta essere il seguente:

$$\begin{cases} \pi_1 = \frac{3}{4}\pi_1 + \frac{1}{4}\pi_2 \\ \pi_2 = \frac{1}{4}\pi_1 + \frac{3}{4}\pi_2 \\ 1 = \pi_1 + \pi_2 \end{cases}$$

Risolvendo, si trova

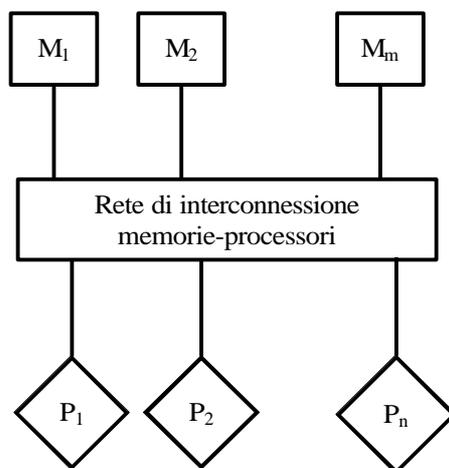
$$\begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

e questo è lo stesso risultato trovato in precedenza a partire dalle probabilità di stato.

Esempio: sistema multiprocessore

Negli esempi precedenti abbiamo sempre supposto di conoscere a priori la matrice delle probabilità di stato della catena di Markov in esame. In quest'altro esempio, invece, procediamo in modo diverso: partiamo da un sistema concreto, verifichiamo che esso sia modellabile tramite una catena di Markov e, in caso affermativo, andiamo a valutare i parametri caratteristici della catena.

Il sistema che prendiamo in esame è un sistema multiprocessore che schematizziamo nel modo seguente:



Abbiamo dunque n **processori** P_1, P_2, \dots, P_n , i quali, tramite una opportuna *rete di interconnessione*, sono in grado di accedere ad una qualsiasi tra m **memorie** M_1, M_2, \dots, M_m . Questa rete di interconnessione è in grado di accettare una qualsiasi combinazione di richieste dei processori verso le memorie, per cui non ci sono limiti da questo punto di vista.

Così come accade nei sistemi reali, si assume la presenza di un *clock* che sincronizzi le varie operazioni: ciò significa che il tempo è discretizzato.

Facciamo l'ipotesi che ciascun processore faccia le proprie richieste di accesso alla memoria indipendentemente da ciò che fanno gli altri processori e supponiamo inoltre che, per ciascun processore, siano fisse le probabilità di richiedere l'accesso ad una determinata memoria.

Quando 2 o più processori richiedono l'accesso ad una stessa memoria (situazione di **collisione**), è ovvio che solo una richiesta possa essere soddisfatta; possiamo allora fare due ipotesi sulla *politica di gestione* di una generica collisione: supponiamo in primo luogo, che, mentre una delle richieste pervenute viene soddisfatta, le altre vengano poste in attesa di essere servite; in secondo luogo, supponiamo che un processore, che non abbia visto esaudita la propria richiesta di accesso a memoria (cioè quindi che non abbia ricevuto i dati richiesti o non abbia potuto scrivere i dati inviati), rimanga bloccato fin quando la richiesta non viene esaudita.

Queste sono dunque le ipotesi sul funzionamento del sistema in esame. Il passo successivo, come detto, consiste nel capire se un siffatto sistema può essere modellato come una catena di Markov.

A questo scopo, il procedimento da seguire è il seguente: bisogna stabilire come definire lo *stato* del sistema e successivamente bisogna verificare se l'evoluzione dello stato dipenda solo dall'istante di osservazione oppure no; in caso affermativo, il sistema sarà una catena di Markov.

E' chiaro allora che è opportuno ricercare una definizione dello *stato* del sistema tale da poter rientrare nella classe delle catene di Markov. Per semplicità, mettiamoci nel caso in cui $m=n=2$, per cui abbiamo solo due processori e due memorie. Con questa posizione (che peraltro è generalizzabile), possiamo pensare di definire lo stato del sistema in base al numero di richieste di servizio per ciascuna memoria:

- stato $\mathbf{s}_1=(1,1)$: abbiamo 1 richiesta di servizio per ciascuna memoria;
- stato $\mathbf{s}_2=(0,2)$: abbiamo 2 richieste di servizio entrambe per M_2 , per cui siamo in una situazione di collisione;
- stato $\mathbf{s}_3=(2,0)$: abbiamo 2 richieste di servizio entrambe per M_1 , per cui siamo ancora in una situazione di collisione;

Nell'individuare questi 3 stati abbiamo evidentemente escluso il caso in cui non ci siano richieste di servizio a memoria: il motivo, derivante dalle situazioni reali, è che il numero di richieste a memoria da parte di un processore è generalmente molto elevato, per cui è plausibile supporre che, in ciascun istante discreto, ciascun processore faccia una propria richiesta di accesso alla memoria.

Per continuare la nostra analisi, possiamo definire le probabilità che ciascuna memoria riceva una richiesta di servizio:

q_1 = probabilità che ci sia 1 richiesta di accesso alla memoria M_1 ;

q_2 = probabilità che ci sia 1 richiesta di accesso alla memoria M_2 ;

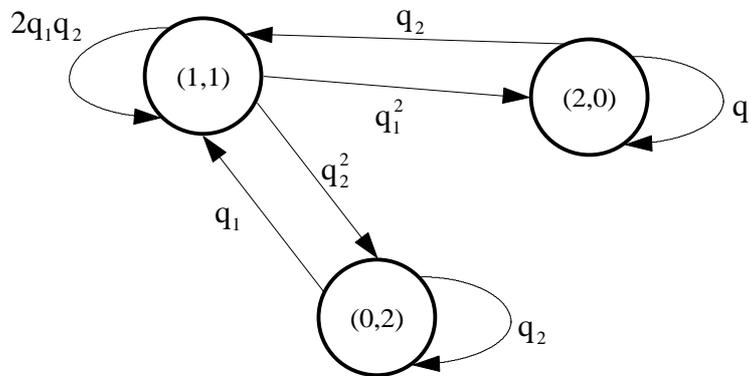
E' evidente che queste probabilità prescindono da quale processore faccia la richiesta di servizio.

Adesso possiamo tracciare il diagramma di transizione di stato del sistema così modellato: dobbiamo cioè valutare le probabilità che il sistema passi da uno stato all'altro tra quelli individuati.

Ad esempio, consideriamo lo stato (1,1) in cui ciascuna memoria ha una richiesta di servizio, il che significa, in altre parole, che i processori hanno fatto richieste diverse. Per passare, ad esempio, nello stato (2,0), è necessario che, nell'istante successivo, ci siano due richieste entrambe alla memoria M_1 e questo avviene con probabilità $q_1 \cdot q_1 = q_1^2$; analogamente, per passare da (1,1) a (0,2), è necessario che ci siano due richieste entrambe alla memoria M_2 e questo avviene con probabilità

$q_2 \cdot q_2 = q_2^2$. Situazione diversa si presenta, invece, per passare da (2,0) a (1,1): infatti, dire che il sistema si trova nello stato (2,0) equivale a dire che un processore è stato servito, ossia sta accadendo alla memoria M_1 , mentre l'altro è in attesa. Allora, il passaggio da (2,0) a (1,1) si ha solo se il processore che è sotto servizio cessa di accedere alla memoria M_1 e richiede l'accesso ad M_2 . Quindi, la probabilità di transizione da (2,0) a (1,1) è q_2 , ossia la probabilità che ci sia una richiesta per M_2 . Analogamente, per passare da (0,2) a (1,1), ci dovrà essere una sola richiesta per M_1 , per cui la probabilità è q_1 . In base a queste considerazioni, si capisce anche perché non è possibile una transizione da (2,0) a (0,2) o viceversa: infatti, se una memoria ha ricevuto due richieste, per cui una viene esaudita mentre l'altra è in attesa, all'istante successivo ci sarà comunque una richiesta alla stessa memoria, mentre l'altra potrebbe riguardare la stessa memoria oppure l'altra.

Mettendo insieme queste considerazioni, si ricava il seguente diagramma di transizione di stato:



Si nota che la probabilità che il sistema, essendo nello stato (1,1), rimanga in tale stato vale $2q_1q_2$: infatti, per passare da (1,1) ad (1,1) è necessario che i due processori, una volta esaurito il servizio richiesto, facciano nuovamente due richieste distinte: si tratta dunque della probabilità congiunta di due eventi, di cui il primo corrisponde a P_1 che richiede servizio ad M_1 e P_2 che richiede servizio ad M_2 ed il secondo corrispondente a P_1 che richiede servizio ad M_2 e P_2 che richiede servizio ad M_1 ; entrambi gli eventi hanno probabilità q_1q_2 , per cui la probabilità congiunta è la loro somma, ossia appunto $2q_1q_2$.

A questo punto, possiamo dedurre facilmente che il sistema è una catena di Markov: infatti, l'evoluzione del sistema dal passo generico N al passo successivo $N+1$ dipende solo dal passo N , in quanto ciascun processore effettua la propria richiesta (passo $N+1$) di accesso a memoria solo in base a quello che è successo nell'ultima richiesta (passo N).

Abbiamo dunque a che fare con una catena di Markov tempo-discreta, per la quale possiamo immediatamente costruire la matrice di transizione ad un passo. Infatti, direttamente per ispezione del diagramma tracciato poco fa, abbiamo che

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{21} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 2q_1q_2 & q_2^2 & q_1^2 \\ q_1 & q_2 & 0 \\ q_2 & 0 & q_1 \end{bmatrix} \quad \begin{array}{l} \longleftarrow s_1 = (1,1) \\ \longleftarrow s_2 = (0,2) \\ \longleftarrow s_3 = (2,0) \end{array}$$

Questa matrice ci consente di calcolare, se esiste, la soluzione asintotica del sistema, ossia le probabilità di stato asintotiche (cioè per $t \rightarrow \infty$). Dobbiamo infatti verificare se ammette soluzione il sistema rappresentato dalle seguenti equazioni:

$$\begin{cases} [\pi] = [\pi]P \\ \sum_{i=1}^3 \pi_i = 1 \end{cases}$$

Scrivendo le 4 equazioni per esteso, abbiamo quanto segue:

$$\begin{cases} \pi_1 = \pi_1 \cdot 2q_1q_2 + \pi_2 \cdot q_1 + \pi_3 \cdot q_2 \\ \pi_2 = \pi_1 \cdot q_2^2 + \pi_2 \cdot q_2 + \pi_3 \cdot 0 \\ \pi_3 = \pi_1 \cdot q_1^2 + \pi_2 \cdot 0 + \pi_3 \cdot q_1 \\ \pi_1 + \pi_2 + \pi_3 = 1 \end{cases}$$

Facciamo adesso l'ipotesi che le probabilità di accesso a memoria siano uguali, per cui $q_1=q_2=0.5$. Con questa posizione si verifica facilmente che il sistema ammette soluzione e tale soluzione è unica, il che significa che la catena è completamente regolare. Facendo i conti, si trova che

$$\begin{cases} \pi_1 = \frac{1}{2} \\ \pi_2 = \frac{1}{4} \\ \pi_3 = \frac{1}{4} \end{cases}$$

Come è ovvio che sia, la somma di queste tre probabilità è pari ad 1.

L'utilità dei calcoli fatti è non tanto nel calcolo delle probabilità asintotiche, quanto nel fatto che queste ultime ci consentono di valutare alcuni indici di prestazione del sistema. Ad esempio, possiamo calcolare il numero di processori che, mediamente, in ciascun ciclo di clock, stanno elaborando dati, a fronte del rimanente numero di processori che invece sono in attesa di essere serviti. Il discorso da fare è semplice. Intanto, la variabile N che ci dà il numero di processori in elaborazione durante il generico ciclo di clock è una variabile aleatoria. Noi dobbiamo calcolare la media di questa variabile aleatoria: $E[N]$. Per definizione, tale media è la somma pesata dei valori assumibili da N , dove i coefficienti di peso sono le varie probabilità:

$$E[N] = \sum_k k \cdot p(N = k)$$

In questo caso, i casi possibili sono 3: se il sistema si trova nello stato $s_1=(1,1)$, entrambi i processori stanno elaborando i dati, per cui il numero di processori in elaborazione è 2; se invece il sistema si trova nello stato $s_2=(0,2)$ o nello stato $s_3=(2,0)$, allora c'è un solo processore in elaborazione e l'altro è in attesa. Di conseguenza, possiamo scrivere che

$$E[N] = \sum_k k \cdot p(N = k) = 2 \cdot p(s_1) + 1 \cdot p(s_2) + 1 \cdot p(s_3)$$

A noi interessa ovviamente una valutazione in condizioni di regime, per cui usiamo le probabilità asintotiche calcolate poco fa:

$$E[N] = 2 \cdot \pi_1 + 1 \cdot \pi_2 + 1 \cdot \pi_3 = 1.5$$

Il risultato ottenuto è peraltro abbastanza intuitivo: infatti, in qualunque stato il sistema si trovi, c'è sempre un processore sotto servizio, per cui $E[N]$ deve essere maggiore di 1; analogamente, in qualunque stato il sistema si trovi, non ci potranno mai essere 2 processori sotto servizio, per cui $E[N]$ deve essere minore di 2.

E' interessante osservare che i risultati ottenuti dipendono dai valori che abbiamo assegnato a q_1 e q_2 , che in particolare sono state poste uguali. Un modo possibile di procedere è quello di massimizzare $E[N]$ rispetto a uno dei due valori (in modo da massimizzare l'operatività del sistema), dal che risulterà poi fissato anche l'altro visto che $q_1 + q_2 = 1$.

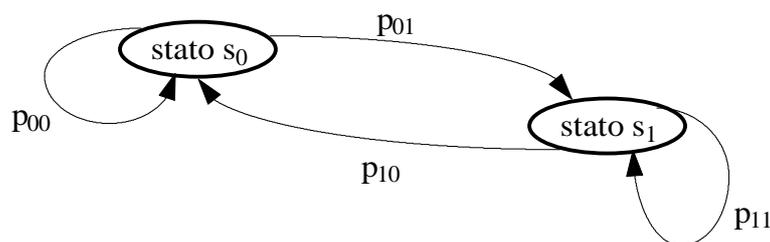
Definizioni varie sulle catene di Markov

Per completare il nostro studio sulle catene di Markov tempo-discrete, forniremo adesso, senza dimostrarli, una serie di risultati abbastanza importanti.

Negli esempi precedenti, abbiamo visto che è possibile dare una rappresentazione grafica molto comoda della matrice delle probabilità di transizione ad un passo: per esempio, se la catena di Markov è a due soli stati, per cui tale matrice è del tipo

$$[P] = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

la sua rappresentazione grafica è



Sono qui indicati i due stati possibili per la catena e le probabilità che la catena passi da uno stato all'altro in un solo passo.

Sulla base di questa rappresentazione grafica è possibile comprendere meglio la seguente definizione:

Def. Una catena di Markov si dice **irriducibile** quando, dato uno stato in cui la catena si trova, da esso è possibile raggiungere uno qualsiasi di tutti gli altri stati

I termini concreti, ciò significa che nessuna delle probabilità di transizione ad un passo è nulla: il sistema, partendo da un certo stato, può passare in qualunque altro stato in un unico passo.

Sempre riguardo gli stati di una catena di Markov, consideriamo adesso una serie di importanti quantità: la prima, che indichiamo con $\mathbf{f}_j(\mathbf{n})$, è definita come la probabilità che il primo ritorno allo stato s_j avvenga n passi dopo aver lasciato lo stesso stato s_j .

Quindi, in altre parole, supponendo che la catena, in un certo istante, si trovi nello stato s_j , la quantità $\mathbf{f}_j(\mathbf{n})$ rappresenta la probabilità che ci vogliano n passi (ossia n intervalli di tempo) perché la catena ritorni nello stato s_j .

Indichiamo invece con \mathbf{f}_j la probabilità che ci sia un ritorno allo stato s_j . In altre parole, mentre con $\mathbf{f}_j(\mathbf{n})$ noi indichiamo quanti passi sono necessari perché si ritorni allo stato di partenza, con \mathbf{f}_j non ci preoccupiamo del numero di passi necessari, ma ci preoccupiamo solo del ritorno allo stato s_j .

E' abbastanza intuitivo comprendere come $\mathbf{f}_j(\mathbf{n})$ e \mathbf{f}_j sia legate dalla relazione

$$\mathbf{f}_j = \sum_{n=1}^{\infty} \mathbf{f}_j(\mathbf{n})$$

Infatti, è chiaro che il ritorno allo stato s_j può avvenire dopo 1 passo, dopo 2 passi, dopo 3 passi fino all'infinito (il che equivarrebbe a dire che non avviene mai). Evidentemente, essendo \mathbf{f}_j una probabilità, abbiamo solo due casi:

- se $\mathbf{f}_j=1$, allora noi siamo certi che ci sia un ritorno allo stato s_j , a prescindere poi dal numero di passi necessari; si dice allora che lo stato s_j è uno **stato ricorrente**;
- se, invece, $\mathbf{f}_j < 1$, allora c'è una certa probabilità che ci sia questo ritorno ma c'è anche una probabilità non nulla che esso non ci sia; in questo caso, lo stato s_j è detto **stato transitorio**.

Osserviamo che, per una catena irriducibile, nella quale cioè da ogni stato sia sempre possibile raggiungere uno stato qualsiasi in un unico passo, risulta ovviamente $\mathbf{f}_j=1$.

Un altro concetto legato a quanto appena detto è quello del **tempo medio di ritorno allo stato s_j** : infatti, mentre prima ci siamo occupati della probabilità di tornare nello stato s_j oppure della probabilità che per tornare a s_j ci volessero n passi, può anche interessarci calcolare il numero medio di passi necessari per tornare allo stato s_j , partendo ovviamente sempre da s_j . Tale numero medio di passi prende appunto il nome di *tempo medio di ritorno allo stato s_j* ed è evidentemente definito mediante la relazione

$$M_j = \sum_{n=1}^{\infty} n \mathbf{f}_j(\mathbf{n})$$

Anche in questo caso, i casi possibili sono due:

- se $M_j = \infty$, significa che ci vogliono infiniti passi per tornare allo stato s_j , ossia significa, in pratica, che non c'è possibilità di ritorno a tale stato; si dice allora che lo stato s_j è uno **stato ricorrente nullo**;
- se, invece, $M_j < \infty$, allora c'è un numero medio finito di passi necessari per tornare ad s_j ; in questo caso, lo stato s_j è detto **stato ricorrente non nullo**.

L'ultima definizione che diamo è la seguente: dato sempre il generico stato s_j , si definisce **tempo di permanenza** in tale stato l'intervallo di tempo durante il quale la catena permane in tale stato, ossia, in altre parole, l'intervallo di tempo che intercorre tra l'istante in cui la catena passa in tale stato e l'istante in cui cambia stato. Si tratta evidentemente di una variabile aleatoria, che indichiamo con T_j , per la quale si trovano le seguenti caratteristiche statistiche:

$$P(T_j = n) = p_{jj}^n (1 - p_{jj})$$

$$E[T_j] = \frac{1}{1 - p_{jj}}$$

La quantità $P(T_j = n)$ è la probabilità che il sistema, una volta giunto nello stato j , vi rimanga per $n (\geq 1)$ passi consecutivi. E' evidente che la variabile aleatoria T_j risulta essere una variabile aleatoria geometrica (quindi senza memoria), da cui scaturisce l'espressione riportata per la sua media.

Generalmente, l'unità di misura del tempo di permanenza in uno stato è lo **slot**, che poi è comunque sinonimo di *passo*.

Esempio: protocollo Slotted Aloha

Consideriamo una *rete di telecomunicazioni* nella quale è a disposizione, di un certo numero di utenti, un'unica risorsa trasmissiva, ad esempio un unico *canale radio*. Il generico utente, quando ha bisogno di usare tale canale, deve seguire delle precise regole. L'insieme delle regole per la gestione dell'accesso al canale è detto **protocollo**. In questo esempio, consideriamo un particolare tipo di protocollo, detto **slotted Aloha**, le cui regole principali sono le seguenti:

- intanto, ogni utente può trasmettere solo in determinati **slot temporali**, di durata predefinita, e non quando vuole;
- in secondo luogo, quando due o più utenti vogliono usare lo stesso slot temporale per accedere al canale, si ha una **collisione**, per cui nessuno degli utenti può trasmettere.

Per una analisi più quantitativa del sistema, possiamo fare delle ipotesi semplificative:

- supponiamo che ogni utente, una volta preparati i dati da trasmettere, li possa inviare una sola volta: questo significa che o li trasmette nel primo slot temporale che risulta disponibile, oppure deve rinunciare;
- supponiamo inoltre che il numero di utenti sia ∞ e che essi generino un traffico in ingresso (cioè le richieste di accesso al canale) di tipo poissoniano, con intensità λ .

Sulla base di queste ipotesi, ha senso chiedersi se il sistema può essere modellato (e quindi poi studiato) come una catena di Markov tempo-discreta (DTMC).

Per stabilire questo, dobbiamo per prima cosa caratterizzare il generico *stato* del sistema, dopo di che potremo verificare se l'evoluzione del sistema, a partire da un

dato istante, dipenda solo dall'istante stesso oppure anche da ciò che è avvenuto prima.

E' evidente che, dal punto di vista del generico slot temporale messo a disposizione, le situazioni possibili sono solo 3:

- la prima è quella per cui nessun utente ha necessità di trasmettere, per cui non ci sono richieste di servizio e quindi lo slot è vuoto;
- la seconda situazione è invece quella in cui un solo utente ha fatto richiesta di servizio e quindi ha occupato lo slot per la trasmissione dei propri dati;
- l'ultima situazione è infine quella in cui ci sono state 2 o più richieste di accesso allo stesso slot temporale, per cui c'è stata una collisione e nessuna trasmissione è stata effettuata.

Possiamo dunque associare, a ciascuna di queste possibilità, uno stato del sistema:

s_0 = slot temporale libero

s_1 = slot temporale occupato da un solo utente

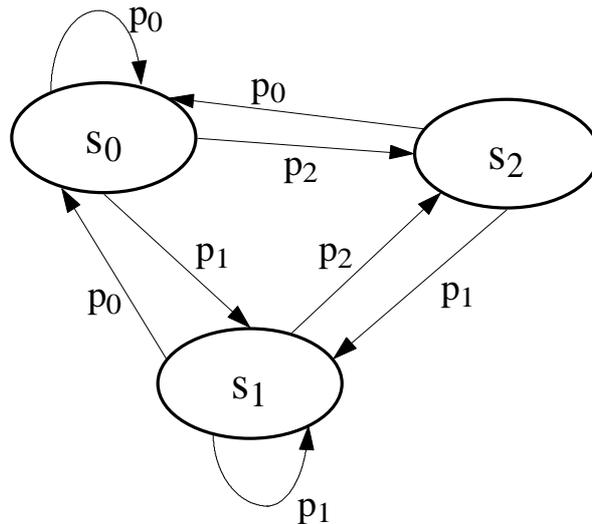
s_2 = collisione

Fatto questo, dobbiamo chiederci se l'evoluzione del sistema, a partire da un dato istante, dipenda solo dallo stato in tale istante oppure dipenda anche dalla storia passata: avendo supposto che il traffico in ingresso sia poissoniano, cioè senza memoria, deduciamo che l'evoluzione del processo non dipende dalla storia passata, per cui concludiamo che il processo è sicuramente markoviano. E', in particolare, una catena di Markov in quanto gli stati sono solo 3.

Possiamo adesso valutare le probabilità di transizione ad 1 passo:

- supponiamo, ad esempio, che il sistema si trovi nello stato s_0 , per cui lo slot temporale è libero; sia p_1 la probabilità che si passi nello stato s_1 : questa probabilità è pari alla probabilità che ci sia un'unica richiesta di servizio, cioè un unico arrivo, durante lo slot (che supponiamo sia di durata generica T_c), per cui, essendo il traffico in ingresso di tipo poissoniano, possiamo scrivere che $p_1 = \lambda T_c e^{-\lambda T_c}$
- supponiamo adesso che il sistema si trovi nello stato s_1 , per cui lo slot temporale è occupato da un utente: la probabilità p_0 di passare nello stato s_0 è la probabilità che l'utente che sta trasmettendo smetta di farlo in un intervallo di tempo T_c e che non ci siano altre richieste durante tale intervallo, per cui $p_0 = e^{-\lambda T_c}$;

Con discorsi di questo tipo, è possibile costruire il seguente diagramma delle probabilità di transizione ad un passo:



E' evidente il significato della probabilità p_2 : è la probabilità che si passi dallo stato s_0 (slot libero) allo stato s_2 (collisione), ossia la probabilità che, nell'intervallo di tempo T_c , ci siano 2 o più richieste di accesso al canale, ed è quindi anche la probabilità di passare da s_1 (slot occupato) ad s_2 . Dato che le richieste di servizio (o gli arrivi) possono essere 0,1 o più di 1, abbiamo che $p_2=1-p_0-p_1$.

La rappresentazione analitica della matrice delle probabilità di transizione ad 1 passo, ricavabile direttamente dal grafo, è dunque la seguente:

$$[P] = \begin{bmatrix} p_0 & p_1 & p_2 \\ p_0 & p_1 & p_2 \\ p_0 & p_1 & p_2 \end{bmatrix} = \begin{bmatrix} p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & 1-p_0-p_1 \end{bmatrix}$$

Essa ha una struttura molto particolare: infatti, è facile verificare che

$$[P] = [P]^2 = [P]^3 = \dots = [P]^N = [P]^\infty$$

Abbiamo cioè trovato che $[P]$ coincide con la matrice $[P]^\infty$ delle probabilità di transizione ad ∞ passi. Non solo, ma osserviamo anche che le righe di $[P]^\infty$ sono tutte uguali tra loro. Possiamo allora applicare un risultato enunciato in precedenza, in base al quale la catena di Markov è completamente regolare e le soluzioni asintotiche sono proprio gli elementi di una qualsiasi delle righe di $[P]^\infty$: concludiamo quindi che le probabilità asintotiche sono

$$\begin{aligned} \pi_0 &= p_0 \\ \pi_1 &= p_1 \\ \pi_2 &= p_2 = 1-p_0-p_1 \end{aligned}$$

A questo punto, possiamo sfruttare la nostra conoscenza delle catene di Markov tempo-discrete per determinare alcuni indici di prestazione del sistema in esame. Ad esempio, possiamo calcolare l'**efficienza di utilizzo** del canale radio, *definita come numero medio di utenti che trasmettono in uno slot*. Si tratta, in pratica, del traffico smaltito dal sistema nell'unità di tempo.

Indicata questa efficienza con S , è evidente che dobbiamo considerare solo i casi in cui lo slot è libero (stato s_0) oppure è occupato da 1 solo utente (stato s_1) e le relative probabilità: abbiamo perciò che

$$S = 1 \cdot p_1 + 0 \cdot p_0 = p_1 = \lambda T_C e^{-\lambda T_C}$$

Dato che λ è il numero medio di richieste di servizio nell'unità di tempo e T_C è la durata di uno slot temporale, deduciamo che il prodotto λT_C rappresenta il numero medio di utenti che richiedono servizio nell'intervallo di tempo T_C . Possiamo allora porre $G = \lambda T_C$ e scrivere che

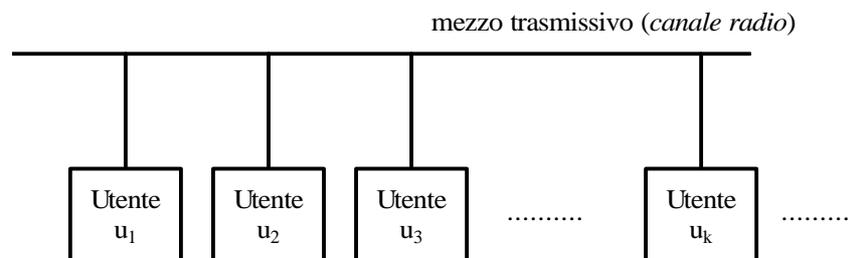
$$S = G e^{-G}$$

Mentre T_C è un parametro che si può fissare a priori, il valore di λ non è, in generale, costante, ma dipende dal traffico. Si tratta, perciò, di una ulteriore variabile aleatoria. Può avere senso, allora, diagrammare S in funzione di G , ossia in funzione di λ (proprio perché T_C si ritiene fissato). Il diagramma che si ottiene indica che il massimo dell'efficienza si ottiene in corrispondenza di $G=1$ (ossia in presenza di una sola richiesta, mediamente, di accesso al canale in ogni T_C) e vale $1/e$, che corrisponde quindi a circa il 36%.

Concludiamo, quindi, che il protocollo *slotted Aloha* garantisce una efficienza di utilizzo del canale che, al più, è del **36%**. Si tratta di un valore abbastanza basso.

Ulteriori osservazioni sul protocollo

Il protocollo *slotted Aloha* appena studiato è un tipico protocollo ad accesso multiplo, nel quale cioè ci sono più utenti (che si sono supposti in numero ∞) che possono fare richiesta di accesso all'unico canale trasmissivo disponibile:



Dato che il canale può supportare la trasmissione da parte di un utente per volta, abbiamo detto che si parla di **collisione** quando ci sono due o più richieste contemporanee di accesso al canale stesso. La collisione, per le ipotesi fatte, è "distruttiva", nel senso che, quando si verifica, nessun utente può trasmettere³.

Il modo con cui un utente può sapere se la sua trasmissione è andata a buon fine oppure no, è univoco: deve aspettare, per un certo tempo, un riscontro da parte della stazione ricevente; se questo riscontro arriva, significa che la stazione ricevente ha ricevuto i dati e quindi che la trasmissione è stata effettuata (salvo ovviamente errori durante la trasmissione); se, invece, il riscontro non arriva dopo un certo tempo massimo, la stazione mittente deduce che c'è stata una collisione

³ In effetti, non sempre accade questo: ad esempio, ci sono sistemi in cui, in presenza di collisione, riesce a trasmettere solo l'utente che ha fornito un contributo di potenza maggiore: si tratta dei cosiddetti **sistemi con cattura**.

che ha impedito la trasmissione. In questo caso, la stazione mittente deve aspettare un tempo casuale e poi deve riprovare. E' ovvio il motivo per cui il tempo di attesa, prima del nuovo tentativo, deve essere casuale: se questo tempo fosse deterministico, tutte le sorgente coinvolte nella collisione riprovarebbero nello stesso istante, per cui si avrebbe una continua collisione. La casualità, invece, con cui ogni stazione sceglie il tempo di attesa garantisce una buona probabilità di riuscita della trasmissione⁴.

Nello studio del sistema condotto poco fa, abbiamo trovato che la rete di telecomunicazione è modellabile tramite una catena di Markov la cui matrice di transizione ad un passo coincide con la matrice $[P]^\infty$ ed è

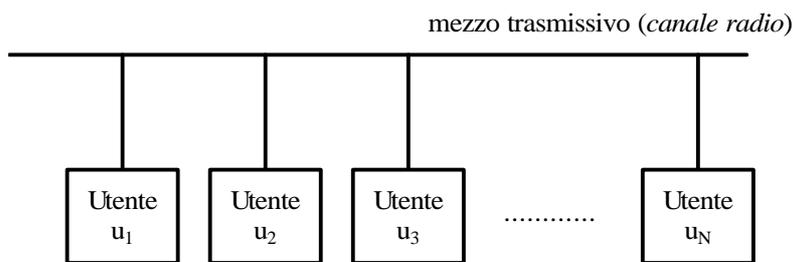
$$[P] = \begin{bmatrix} p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & 1-p_0-p_1 \end{bmatrix}$$

Osservando l'uguaglianza delle righe di questa matrice, abbiamo dedotto che la catena di Markov è completamente regolare e che le soluzioni asintotiche sono proprio gli elementi di una qualsiasi delle righe di $[P]^\infty$:

$$\begin{aligned} \pi_0 &= p_0 \\ \pi_1 &= p_1 \\ \pi_2 &= p_2 = 1-p_0-p_1 \end{aligned}$$

Sulla base di questi risultati, abbiamo calcolato l'efficienza di utilizzo del canale radio, che è risultata avere il valore massimo di circa il 36%, che è un valore molto basso.

Adesso vediamo come cambiano questi risultati se supponiamo che gli utenti della rete non siano più in numero infinito, ma in numero finito N:



L'ipotesi è sempre che ciascun utente si comporti indipendentemente dagli altri.

Supponiamo inoltre che il tempo di generazione di un pacchetto, da parte del generico utente, sia una variabile casuale con distribuzione esponenziale con parametro (intensità) λ/N .

Si può verificare che anche in questo caso risulta $p_1 = \lambda T_c e^{-\lambda T_c}$ e inoltre che il traffico in uscita (**throughput**) del sistema coincide con l'efficienza di utilizzo del canale.

⁴ Si tenga presente che la collisione avviene solo quando ci sono due richieste contemporanee di accesso al canale e non quando un utente chiede l'accesso mentre il canale è già occupato: in questo caso, non c'è collisione.

Teorema 1 (per catene di Markov omogenee e irriducibili)

Supponiamo di avere una catena di Markov omogenea e irriducibile. E' possibile dimostrare, per una catena di questo tipo, che i suoi stati sono o tutti transitori o tutti ricorrenti nulli o tutti ricorrenti non nulli.

In altre parole, se noi osserviamo che uno stato della catena è transitorio o ricorrente nullo o ricorrente non nullo, possiamo star certi che anche tutti gli altri sono transitori o ricorrenti nulli o ricorrenti non nulli.

Inoltre, se la catena è anche periodica, si dimostra che tutti gli stati hanno lo stesso periodo (il che significa che è costante, per ciascun passo, il numero minimo di stati necessari per tornare nel passo considerato).

Teorema 2 (per catene di Markov omogenee, irriducibili e non periodiche)

Data una catena di Markov omogenea, irriducibile e non periodica, si dimostra che le probabilità (di stato) asintotiche $\pi_j = \lim_{n \rightarrow \infty} P(X_n = j)$ esistono sempre e sono indipendenti dalle probabilità di stato iniziali.

In particolare, si dimostra quanto segue:

- se tutti gli stati sono ricorrenti nulli, non è possibile ricavare una soluzione asintotica, dato che sono necessari infiniti passi per tornare nello stato di partenza, ossia tale stato non viene mai raggiunto;
- se tutti gli stati sono ricorrenti non nulli, allora $\pi_j = \frac{1}{M_j}$

Caso particolare

Se il numero di stati della catena omogenea, irriducibile e non-periodica, è finito, si tratta di stati tutti ricorrenti non nulli.

Catene di Markov tempo-continue

Introduzione

Riprendiamo velocemente la definizione data di “processo stocastico markoviano”: *un processo stocastico è markoviano quando solo l'istante scelto di osservazione determina l'evoluzione futura del processo, mentre non contano gli istanti precedenti quello di osservazione.*

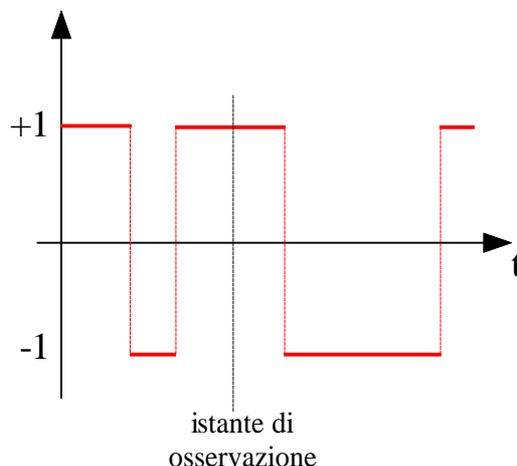
Abbiamo detto che si definisce in particolare **catena di Markov** un processo markoviano che goda della proprietà di essere a valori discreti: a questi valori discreti noi diamo spesso il nome di **stati** del processo (o anche “*stati del sistema*”).

Le catene di Markov, come tutti i processi stocastici, possono essere di due tipi: prima ci siamo occupati delle catene di Markov tempo-discrete, mentre adesso ci occupiamo delle **catene di Markov tempo-continue** (brevemente **CTMC**).

Il processo telegrafico casuale

Un esempio tipico di catena di Markov tempo-continua è quello che noi conosciamo come **processo telegrafico casuale**: si tratta di un processo stocastico, continuo nel tempo, che può assumere solo gli "stati" +1 e -1. Verifichiamo effettivamente che si tratti di una catena di Markov.

Consideriamo una generica realizzazione di questo processo:



Fissiamo un generico istante t che prendiamo come istante di osservazione: in tale istante, il sistema si troverà in uno dei due possibili stati e supponiamo che si tratti dello stato +1. Ci interessa verificare se l'evoluzione successiva del processo venga influenzata o meno dagli stati che il sistema ha assunto prima dell'istante di osservazione. Per fare questo ci ricordiamo di una importante caratteristica del processo telegrafico casuale: si definisce *tempo di attesa* del processo il tempo che intercorre tra l'istante considerato ed il primo cambio di stato che si verifica dopo tale istante; si tratta di una variabile aleatoria e sappiamo in particolare che essa ha distribuzione esponenziale: questo perché, per ipotesi, i cambiamenti di stato del processo telegrafico casuale avvengono quando si verifica un evento di Poisson e la distribuzione del processo di Poisson è proprio di tipo esponenziale.

La caratteristica della distribuzione esponenziale è quella di essere *senza memoria*: ciò significa, in questo caso, che la probabilità che il sistema cambi stato dipende effettivamente solo dallo stato in cui il sistema si trova all'istante di osservazione, mentre non dipende dagli istanti precedenti.

Possiamo perciò concludere che certamente il processo telegrafico casuale è un esempio di catena di Markov tempo-continua.

Le probabilità asintotiche

Introduzione

Così come abbiamo fatto per le catene di Markov tempo-discrete, l'obiettivo principale che ci poniamo è quello di valutare le **probabilità asintotiche** delle catene di Markov tempo-continue: l'espressione della generica probabilità asintotica è

$$\pi_j = \lim_{t \rightarrow \infty} P(X(t) = j)$$

ed essa rappresenta la probabilità che, a regime, il sistema si trovi nello stato j .

Le probabilità di transizione ad un passo

Indichiamo con $X(t)$ la nostra generica catena di Markov tempo-continua. Fissiamo degli istanti successivi $t_1 < t_2 < \dots < t_k < t_{k+1}$; in corrispondenza di ciascuno di tali istanti, estraiamo dal processo una variabile aleatoria; indichiamo tali variabili aleatorie con $X(t_1), X(t_2), \dots, X(t_k), X(t_{k+1})$; valutiamo allora quanto vale la quantità

$$P(X(t_1) = x_1 \cap X(t_2) = x_2 \cap \dots \cap X(t_k) = x_k \cap X(t_{k+1}) = x_{k+1})$$

dove ovviamente $x_1, x_2, \dots, x_k, x_{k+1}$ sono alcuni tra i possibili stati del sistema.

Esprimendola in termini di probabilità condizionate, abbiamo che

$$\begin{aligned} &P(X(t_1) = x_1 \cap X(t_2) = x_2 \cap \dots \cap X(t_k) = x_k \cap X(t_{k+1}) = x_{k+1}) = \\ &= P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k \cap \dots \cap X(t_2) = x_2 \cap X(t_1) = x_1) P(X(t_k) = x_k \cap \dots \cap X(t_2) = x_2 \cap X(t_1) = x_1) \end{aligned}$$

Adesso, sfruttando il fatto che abbiamo a che fare con una catena di Markov, possiamo semplificare quella probabilità condizionata, restringendola solo all'istante t_k (cioè l'istante di osservazione):

$$\begin{aligned} &P(X(t_1) = x_1 \cap X(t_2) = x_2 \cap \dots \cap X(t_k) = x_k \cap X(t_{k+1}) = x_{k+1}) = \\ &= P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k) P(X(t_k) = x_k \cap \dots \cap X(t_2) = x_2 \cap X(t_1) = x_1) \end{aligned}$$

Adesso ripetiamo lo stesso discorso per il termine $P(X(t_k) = x_k \cap \dots \cap X(t_2) = x_2 \cap X(t_1) = x_1)$: esprimendolo mediante le probabilità condizionate, abbiamo che

$$\begin{aligned} &P(X(t_1) = x_1 \cap X(t_2) = x_2 \cap \dots \cap X(t_k) = x_k \cap X(t_{k+1}) = x_{k+1}) = \\ &= P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k) P(X(t_k) = x_k | X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1) P(X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1) \end{aligned}$$

e quindi, usando la *proprietà di Markov*, possiamo scrivere che

$$\begin{aligned} &P(X(t_1) = x_1 \cap X(t_2) = x_2 \cap \dots \cap X(t_k) = x_k \cap X(t_{k+1}) = x_{k+1}) = \\ &= P(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k) P(X(t_k) = x_k | X(t_{k-1}) = x_{k-1}) P(X(t_{k-1}) = x_{k-1} \cap \dots \cap X(t_1) = x_1) \end{aligned}$$

Procedendo in questo modo per tutte le probabilità assolute, otteniamo il seguente risultato finale:

$$\begin{aligned} P(X(t_1)=x_1 \cap X(t_2)=x_2 \cap \dots \cap X(t_k)=x_k \cap X(t_{k+1})=x_{k+1}) = \\ = P(X(t_1)=x_1) \prod_{n=1}^k P(X(t_{n+1})=x_{n+1} | X(t_n)=x_n) \end{aligned}$$

In perfetta analogia a quanto abbiamo fatto per le catene di Markov tempo-discrete, noi diremo che una catena di Markov tempo-continua è **omogenea** quando il termine $P(X(t_{n+1})=x_{n+1} | X(t_n)=x_n)$ dipende solo dalla quantità $t_{n+1}-t_n$.

La generica quantità $P(X(t_{n+1})=x_{n+1} | X(t_n)=x_n)$ prende sempre il nome di **probabilità di transizione ad un passo** e rappresenta la probabilità che il sistema passi dallo stato x_n allo stato x_{n+1} nell'intervallo di tempo $[t_n, t_{n+1}]$. Ovviamente, per **passo** si intende l'intervallo di tempo che intercorre tra due istanti successivi di osservazione.

Vediamo adesso di semplificare un po' le nostre notazioni: in primo luogo, anziché indicare con x_{n+1} e x_n gli stati in cui il sistema si trova negli istanti t_{n+1} e t_n , possiamo più comodamente indicarli con j e con i , per cui la nostra generica probabilità di transizione ad un passo diventa

$$P(X(t_{n+1})=j | X(t_n)=i)$$

Inoltre, se d'ora in poi consideriamo solo catene di Markov (tempo-continue) omogenee, abbiamo prima detto che quella probabilità condizionata dipende solo dall'ampiezza dell'intervallo $[t_n, t_{n+1}]$ e non dall'istante iniziale e da quello finale: possiamo allora porre genericamente $s=t_n$ e $t+s=t_{n+1}$, per cui abbiamo

$$P(X(s+t)=j | X(s)=i)$$

Detto questo, poniamo infine per semplicità

$$p_{ij}(t) = P(X(s+t)=j | X(s)=i)$$

Quindi, $p_{ij}(t)$ è la probabilità che il sistema, in un qualsiasi intervallo di tempo di ampiezza t , passi dallo stato i allo stato j . Vediamo due intuitive proprietà di $p_{ij}(t)$ che sfrutteremo nel seguito:

- la prima proprietà è che

$$p_{ij}(0) = 0$$

Infatti, $p_{ij}(0)$ rappresenta la probabilità che il sistema, in un intervallo di tempo di ampiezza 0, passi dallo stato i allo stato j ; è chiaro che ciò è impossibile, per cui quella probabilità è nulla;

- la seconda proprietà è invece che

$$p_{jj}(0) = 1$$

Infatti, $p_{ij}(0)$ rappresenta la probabilità che il sistema, in un intervallo di tempo di ampiezza 0, passi dallo stato j allo stato j , ossia rimanga nello stato j ; è chiaro che questo è certo, in quanto in un intervallo di ampiezza 0 non ci possono essere cambiamenti di stato, per cui quella probabilità è unitaria.

Tempo di permanenza in uno stato

Si definisce tempo di permanenza in uno stato il tempo che intercorre tra due successivi cambiamenti di stato, ossia, in altre parole, il tempo che intercorre dal momento in cui il sistema passa in un generico stato al momento in cui esso cambia nuovamente stato.

Si tratta, in pratica, di quello che nel processo telegrafico casuale abbiamo definito *tempo di interarrivo*. Anzi, proprio quanto sappiamo sul processo telegrafico casuale ci è molto utile adesso: infatti, sappiamo che il tempo di interarrivo, nel processo telegrafico casuale, è una variabile aleatoria con distribuzione esponenziale (quindi senza memoria); questo è un risultato di carattere assolutamente generale: *data una qualsiasi catena di Markov omogenea e tempo-continua, il tempo di permanenza in uno stato è una variabile aleatoria con distribuzione esponenziale*.

Vediamo di verificare questa proprietà.

Indichiamo con T_i il tempo di permanenza del sistema nel generico stato i . Trattandosi di una variabile aleatoria, la sua distribuzione cumulativa è definita come

$$F_{T_i}(t) = P(T_i \leq t)$$

Fissato un certo stato i e un certo tempo t , essa rappresenta la probabilità che il sistema rimanga nello stato i per un tempo non superiore a t . Vogliamo valutare quanto vale quella funzione.

La prima cosa che possiamo dire è evidentemente che

$$F_{T_i} = P(T_i \leq t) = 1 - P(T_i > t)$$

Al fine di valutare $P(T_i > t)$ facciamo il discorso seguente: la quantità $P(T_i > t + s | T_i > s)$ rappresenta la probabilità che il sistema si mantenga nello stato i durante un tempo maggiore di $t+s$, sapendo che è già rimasto in tale stato per un tempo pari ad s ; dire che il sistema è nello stato i già da un tempo pari ad s equivale a dire che

$$X(\tau) = i \quad \forall \tau \in [0, s]$$

per cui possiamo scrivere che

$$P(T_i > t + s | T_i > s) = P(T_i > t + s | X(\tau) = i \quad \forall \tau \in [0, s])$$

Ora, se scegliamo l'istante s come istante di osservazione, possiamo sfruttare la *proprietà di Markov* per dire che l'unica informazione che per noi conta è il fatto che, all'istante $t=s$, il sistema si trovasse nello stato i (mentre non ci interessa quello che è accaduto prima dell'istante s): quindi

$$P(T_i > t+s | T_i > s) = P(T_i > t+s | X(s) = i)$$

Inoltre, è evidente che l'evento per cui $T_i > t+s$ equivale all'evento per cui $X(t+s)=i$, ossia all'evento per cui, all'istante $t+s$, il sistema si trovi ancora nello stato i : quindi

$$P(T_i > t+s | T_i > s) = P(X(t+s) = i | X(s) = i)$$

e ancora, data l'omogeneità, possiamo scrivere che

$$P(T_i > t+s | T_i > s) = P(X(t) = i | X(0) = i)$$

Ma $P(X(t) = i | X(0) = i) = P(T_i > t)$, per cui possiamo concludere che

$$P(T_i > t+s | T_i > s) = P(T_i > t)$$

A questo punto, l'unica variabile aleatoria che soddisfa quest'ultima condizione risulta essere quella con distribuzione esponenziale, il che significa che

$$F_{T_i} = P(T_i \leq t) = 1 - e^{-v_i t}$$

N.B. La presenza del pedice "i" nella espressione di quella funzione serve ad indicare la dipendenza del parametro v dallo stato i considerato

In precedenza, abbiamo già studiato quali sono le caratteristiche statistiche di una variabile aleatoria con distribuzione esponenziale: ricordiamo in particolare che il suo valore medio $E[T_i] = \frac{1}{v_i}$. Il valor medio del tempo di permanenza nello stato i è dunque $1/v_i$: essendo questo un tempo, è chiaro allora che v_i rappresenta una *frequenza*; si tratta perciò del numero di volte in cui il sistema passa nello stato i nell'unità di tempo.

Le frequenze di transizione di stato

A questo punto, usando quello che abbiamo detto fino ad ora e usando una serie di relazioni che ci accingiamo a ricavare, arriveremo a valutare le probabilità asintotiche.

Consideriamo intanto un intervallo di tempo di ampiezza generica δ : abbiamo trovato prima che la funzione di distribuzione del tempo di permanenza è data da

$$F_{T_i} = P(T_i \leq t) = 1 - e^{-v_i t}$$

che poi equivale anche a

$$F_{T_i} = 1 - P(T_i > t) = 1 - e^{-v_i t}$$

Allora, possiamo scrivere che

$$P(T_i > \delta) = e^{-v_i \delta}$$

Se supponiamo che δ sia piccolo, possiamo sviluppare in serie quell'esponenziale a scrivere che

$$P(T_i > \delta) = 1 - v_i \delta + o(\delta)$$

dove il termine $o(\delta)$ indica degli *infinitesimi di ordine superiore*.

Noi vogliamo calcolare le probabilità di transizione, la generica delle quali è

$$p_{ij}(\delta) = P(X(t+\delta) = j | X(t) = i)$$

In base a quanto visto nel precedente paragrafo, scriviamo che

$$p_{ii}(\delta) = P(T_i > \delta) + o(\delta) = 1 - v_i \delta + o(\delta)$$

Si può inoltre dimostrare, ancora più in generale, che

$$p_{ij}(\delta) = (1 - p_{ii}(\delta)) \tilde{q}_{ij}$$

dove abbiamo indicato con \tilde{q}_{ij} la probabilità che il sistema, una volta uscito dallo stato i , passi nello stato j ; facciamo subito osservare come si tratti di qualcosa di diverso dalle probabilità di transizione ad un passo: infatti, mentre quelle rappresentano la probabilità che il sistema passi direttamente da i a j , il termine \tilde{q}_{ij} prevede anche la possibilità che il sistema, uscendo da i , passi per altri stati prima di arrivare a j .

Quella espressione di $p_{ij}(\delta)$ può essere ulteriormente modificata tenendo conto che

$$p_{ii}(\delta) = 1 - v_i \delta + o(\delta)$$

Infatti, da quest'ultima relazione si deduce che

$$1 - p_{ii}(\delta) = v_i \delta - o(\delta)$$

per cui

$$p_{ij}(\delta) = (v_i \delta - o(\delta)) \tilde{q}_{ij} = v_i \delta \tilde{q}_{ij} - o(\delta)$$

Generalmente si pone $v_i \tilde{q}_{ij} = \gamma_{ij}$, per cui possiamo concludere che

$$p_{ij}(\delta) = P(X(t+\delta) = j | X(t) = i) = \delta \gamma_{ij} - o(\delta)$$

Si pone inoltre $\gamma_{ii} = -v_i$. Il significato concreto dei parametri γ_{ii} e γ_{ij} si deduce dalle seguenti relazioni:

$$\lim_{\delta \rightarrow 0} \frac{p_{ij}(\delta)}{\delta} = \gamma_{ij}$$

$$\lim_{\delta \rightarrow 0} \frac{p_{ii}(\delta) - 1}{\delta} = \gamma_{ii}$$

Il termine γ_{ij} prende il nome di **frequenza di transizione dallo stato i allo stato j**: esso rappresenta infatti il numero medio di volte in cui, nell'unità di tempo, il sistema passa dallo stato i allo stato j.

Determinazione delle probabilità asintotiche

Premesso tutto questo, andiamo a valutare le cosiddette **probabilità di stato**, la generica delle quali è $P(X(t) = j)$: note queste probabilità, facendo il limite per $t \rightarrow \infty$, troveremo le probabilità asintotiche.

Intanto, poniamo per comodità

$$p_j(t) = P(X(t) = j)$$

$$p_j(t + \delta) = P(X(t + \delta) = j)$$

Usando il *teorema delle probabilità totali*, possiamo scrivere che

$$p_j(t + \delta) = P(X(t + \delta) = j) = \sum_i P(X(t + \delta) = j | X(t) = i) P(X(t) = i)$$

Il termine $P(X(t + \delta) = j | X(t) = i)$ è la generica probabilità di transizione ad un passo $p_{ij}(\delta)$, mentre il termine $P(X(t) = i)$ è la generica probabilità di stato $p_i(t)$: quindi

$$p_j(t + \delta) = \sum_i p_{ij}(\delta) p_i(t)$$

N.B. Facciamo osservare come quella sommatoria sia estesa a tutti i possibili stati del sistema, incluso lo stato j

A questo punto, calcoliamo il seguente limite:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t + \delta) - p_j(t)}{\delta}$$

Sostituiamo l'espressione di $p_j(t + \delta)$ trovata prima:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t + \delta) - p_j(t)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\sum_i p_{ij}(\delta) p_i(t) - p_j(t)}{\delta}$$

Separiamo in due parti quel limite:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\sum_i p_{ij}(\delta) p_i(t)}{\delta} - \lim_{\delta \rightarrow 0} \frac{p_j(t)}{\delta}$$

Adesso, tiriamo fuori dalla sommatoria il termine che si ottiene per $i=j$:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\sum_{i \neq j} p_{ij}(\delta) p_i(t) + p_{jj}(\delta) p_j(t)}{\delta} - \lim_{\delta \rightarrow 0} \frac{p_j(t)}{\delta}$$

Mettiamo insieme il termine per $i=j$ con l'altro termine di cui si compone quel limite:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\sum_{i \neq j} p_{ij}(\delta) p_i(t)}{\delta} + \lim_{\delta \rightarrow 0} \frac{(p_{jj}(\delta) - 1) p_j(t)}{\delta}$$

Sempre nel secondo termine, il termine $p_j(t)$ non dipende da δ , per cui possiamo portarlo fuori dal limite:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \lim_{\delta \rightarrow 0} \frac{\sum_{i \neq j} p_{ij}(\delta) p_i(t)}{\delta} + p_j(t) \lim_{\delta \rightarrow 0} \frac{(p_{jj}(\delta) - 1)}{\delta}$$

Aggiustiamo adesso anche il primo termine del limite:

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \sum_{i \neq j} p_i(t) \lim_{\delta \rightarrow 0} \frac{p_{ij}(\delta)}{\delta} + p_j(t) \lim_{\delta \rightarrow 0} \frac{(p_{jj}(\delta) - 1)}{\delta}$$

A questo punto, è evidente che

$$\lim_{\delta \rightarrow 0} \frac{p_j(t+\delta) - p_j(t)}{\delta} = \frac{dp_j(t)}{dt}$$

per cui possiamo scrivere che

$$\frac{dp_j(t)}{dt} = \sum_{i \neq j} p_i(t) \lim_{\delta \rightarrow 0} \frac{p_{ij}(\delta)}{\delta} + p_j(t) \lim_{\delta \rightarrow 0} \frac{(p_{jj}(\delta) - 1)}{\delta}$$

Questa è una equazione differenziale nella incognita $p_j(t)$. Vogliamo provare a risolverla.

Per prima cosa, ricordandoci che

$$\lim_{\delta \rightarrow 0} \frac{p_{ij}(\delta)}{\delta} = \gamma_{ij}$$

$$\lim_{\delta \rightarrow 0} \frac{p_{jj}(\delta) - 1}{\delta} = \gamma_{jj}$$

possiamo riscriverla nella forma

$$\frac{dp_j(t)}{dt} = \sum_{i \neq j} \gamma_{ij} p_i(t) + \gamma_{jj} p_j(t)$$

e quindi anche nella forma

$$\frac{dp_j(t)}{dt} = \sum_i \gamma_{ij} p_i(t)$$

La risoluzione analitica di questa equazione non sempre è agevole. Tuttavia, possiamo fare un ragionamento di questo tipo: nel caso in cui il processo evolva verso una condizione finale di equilibrio, è chiaro che, per $t \rightarrow \infty$, non ci saranno variazioni delle probabilità di stato, le quali risulteranno essere costanti. Dire che la generica probabilità di stato $p_j(t)$ è costante nel tempo equivale a dire che $\frac{dp_j(t)}{dt} = 0$. Imponendo allora questa condizione all'equazione differenziale di prima, quest'ultima diventa

$$\sum_i \gamma_{ij} p_i(t) = 0$$

ossia diventa una equazione lineare. Risolvendola, al variare di j , otteniamo le varie probabilità di stato all'equilibrio. Anzi, proprio per indicare che ci riferiamo ad una condizione in cui è scomparsa la dipendenza dal tempo, possiamo riscrivere quella equazione nella forma

$$\sum_i \gamma_{ij} p_i = 0$$

Su questa possiamo fare delle ulteriori manipolazioni che rendano più facile la sua risoluzione. Intanto, noi possiamo separare il termine che si ottiene per $i=j$, ottenendo che

$$\sum_i \gamma_{ij} p_i = -\gamma_{jj} p_j$$

Sappiamo che $-\gamma_{jj} = v_j$, per cui quella diventa

$$\sum_{i \neq j} \gamma_{ij} p_i = v_j p_j$$

A questo punto, ci ricordiamo che $\gamma_{ji} = v_j \tilde{q}_{ji}$, ossia anche che

$$\gamma_{ji} = v_j \tilde{q}_{ji}$$

Facendo la sommatoria su i (diverso da j) di ambo i membri di questa relazione, otteniamo che

$$\sum_{i \neq j} \gamma_{ji} = \sum_{i \neq j} v_j \tilde{q}_{ji} = v_j \sum_{i \neq j} \tilde{q}_{ji} = v_j$$

Abbiamo dunque trovato che $v_j = \sum_{i \neq j} \gamma_{ji}$, per cui, andando a sostituire nella equazione che dobbiamo risolvere, abbiamo che

$$\boxed{\sum_{i \neq j} \gamma_{ij} p_i = p_j \sum_{i \neq j} \gamma_{ji} \quad \forall j}$$

Questo⁵ è dunque un sistema di equazioni lineari in cui il numero di equazioni è esattamente pari al numero degli stati possibili per il sistema. Per trovare le incognite di questo sistema, che sono le probabilità di stato a regime (ossia le *probabilità asintotiche*), dobbiamo anche considerare la cosiddetta **condizione di normalizzazione**, secondo la quale

$$\boxed{\sum_i p_i = 1}$$

In conclusione, quindi, risolvendo il sistema

$$\begin{cases} \sum_{i \neq j} \gamma_{ij} p_i = \sum_{i \neq j} \gamma_{ji} p_j & \forall j \\ \sum_i p_i = 1 \end{cases}$$

troviamo, nell'ipotesi preliminare che esistano, le probabilità asintotiche p_j della catena di Markov tempo-continua omogenea considerata.

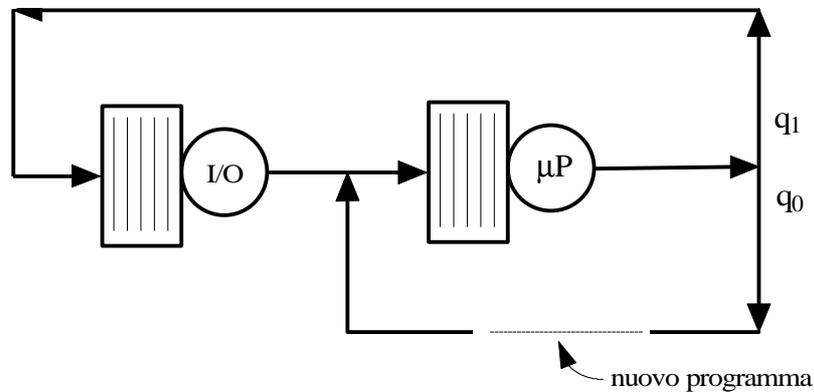
Esempio: calcolatore con multiprogrammazione

Consideriamo un esempio di catena di Markov tempo continua. In particolare consideriamo un calcolatore costituito da un unico microprocessore (μP) ed un'unica unità di input/output (I/O). Supponiamo anche che il calcolatore possa lavorare in **multiprogrammazione**: questo significa che, in un certo istante, possono essere presenti nella CPU più programmi, di cui però uno solo è in esecuzione (da

⁵ L'espressione appena ricavata prende il nome di **equazione di bilancio globale del flusso**, in quanto rappresenta l'uguaglianza tra il flusso entrante nello stato j (primo membro della relazione) ed il flusso uscente dallo stato j (secondo membro della relazione), al variare di j , cioè per tutti gli stati del sistema.

parte del μP) e tutti gli altri in attesa. Il numero n di programmi che possono trovarsi contemporaneamente nella CPU prende il nome di **livello di multiprogrammazione**.

Un modello schematico di questo calcolatore potrebbe essere il seguente:



Il processore è rappresentato con una notazione che diventerà poi frequente nei *sistemi a coda*: c'è infatti un *servente*, che rappresenta tutte le risorse usate per l'esecuzione del programma, e poi c'è una *coda*, nella quale sono allocati tutti i programmi che sono in attesa di essere eseguiti⁶. Analoga schematizzazione si usa per la periferica di I/O, la quale può servire un unico programma per volta, mentre gli altri sono messi in attesa. Indichiamo con μ il tempo medio di servizio del μP , ossia il tempo medio che il μP impiega per dedicarsi al singolo programma, e con λ il tempo medio di servizio della periferica, con lo stesso significato di μ .

Nel diagramma è anche evidenziato il "percorso" che un programma può seguire: ad esempio, dopo che è stato servito dal μP , il programma può richiedere servizio alla periferica con probabilità q_1 oppure può uscire dal sistema, con probabilità q_0 . L'uscita dal sistema può essere eventualmente compensata da un *nuovo programma* che entra nel sistema richiedendo i servizi sicuramente del μP , mentre successivamente ha la possibilità di passare per la periferica di I/O.

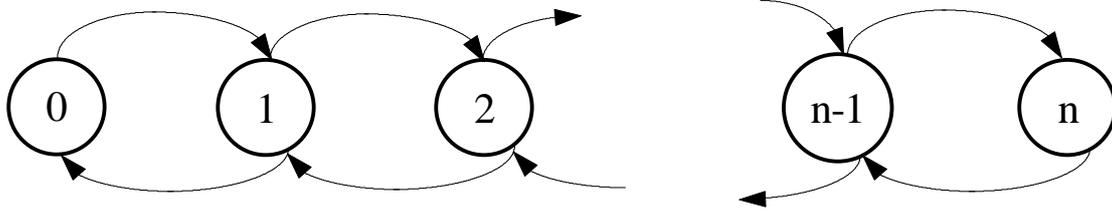
Questo è dunque il sistema reale che dobbiamo analizzare. Ci chiediamo se lo possiamo modellare tramite una catena di Markov tempo-continua.

Come al solito, il primo problema da porsi è il modo con cui *definire* lo stato del sistema. Abbiamo, in proposito, varie possibilità. Una di queste è quella di definire stato del sistema il numero di utenti che complessivamente si trovano nella CPU. Da questa posizione scaturiscono una serie di conseguenze. Per esempio, il numero di stati è $n+1$:

- potremmo avere la CPU vuota (stato 0), il che significa che il μP è inattivo e che tutti i programmi stanno richiedendo i servizi della periferica di I/O;
- potremmo avere la CPU con 1 solo programma (stato 1), il che significa necessariamente che il μP sta lavorando e non ci sono programmi in attesa;
- potremmo avere 2 programmi nella CPU (stato 2), di cui uno sicuramente in esecuzione e l'altro sicuramente in attesa;
- potremmo avere 3 programmi nella CPU (stato 3) e così via fino ad avere n programmi nella CPU (stato n), dei quali uno in esecuzione e i rimanenti $n-1$ in attesa.

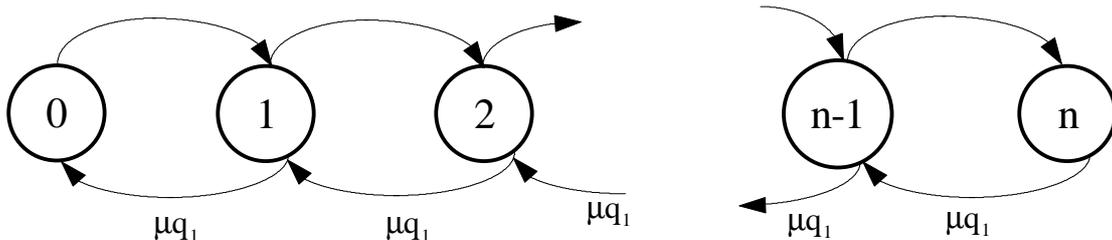
⁶ Si suppone che la gestione dei programmi in coda, sia nel μP sia nella periferica, sia di tipo **FIFO**, cioè *First In First Out*: significa che i programmi vengono serviti in ordine di arrivo, dal primo arrivato all'ultimo arrivato.

Individuato il numero di stati, possiamo costruire un diagramma di transizione di stato, ossia riportare graficamente tutte le possibili transizioni di stato. Si può verificare facilmente che tale diagramma è fatto nel modo seguente:

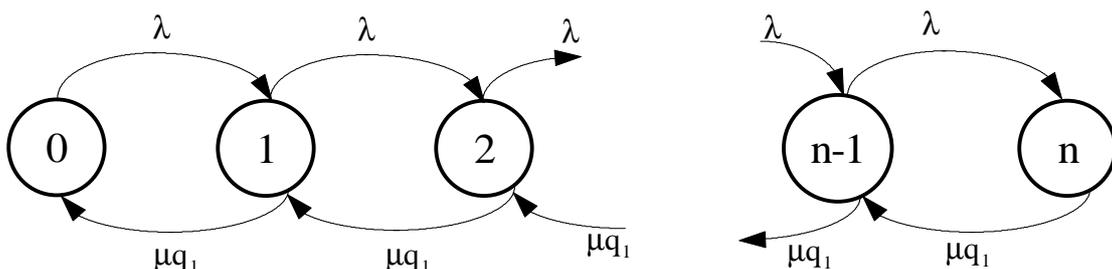


Ciò che risalta subito in questo diagramma è il fatto che sono possibili solo passaggi tra stati consecutivi ed è ovvio che sia così: se ci sono k programmi nella CPU, uno di essi è in esecuzione e gli altri sono in attesa; non appena il programma in esecuzione termina le proprie elaborazioni, o esce dal sistema (con probabilità q_0) oppure va verso la periferica (con probabilità q_1); contemporaneamente, la periferica di I/O può liberare un programma per volta. Di conseguenza possiamo avere variazioni, nel numero di programmi presenti nella CPU, di 1 sola unità per volta, in aumento o in diminuzione.

Così come vedremo nei sistemi a coda, è utile indicare, sul diagramma prima riportato, non tanto le probabilità di transizione, quanto le frequenze di transizione. Ad esempio, consideriamo lo stato 1, nel quale cioè c'è 1 solo programma nella CPU, che quindi sarà sicuramente sotto esecuzione: una volta terminata l'esecuzione, c'è una probabilità q_1 che il programma richieda i servizi della periferica, nel qual caso il sistema passa nello stato 0 (nessun utente nella CPU); dato che μ è il tempo medio di servizio del μP , deduciamo che la quantità μq_1 è la frequenza di transizione dallo stato 1 allo stato 0. D'altra parte, lo stesso identico ragionamento vale per qualunque numero di programmi presenti nella CPU, il che significa che le **frequenze di transizione all'indietro** (cioè dallo stato k allo stato k-1) valgono tutte μq_1 .



Per quanto riguarda, invece, le **frequenze di transizione in avanti**, sono tutte uguali a λ , dato che il tempo medio di servizio della periferica di I/O è pari a λ ed un programma in uscita dalla periferica ha una probabilità unitaria di passare nella CPU:



A questo punto, è facile rendersi conto del fatto che il sistema può essere modellato tramite una catena di Markov tempo-continua: se il sistema si trova nel generico stato k , l'evoluzione successiva dipende solo dal tempo di servizio del programma sotto esecuzione; se supponiamo che questo tempo di servizio sia di tipo esponenziale, cioè senza memoria, è chiaro che l'evoluzione futura dipende solo dall'istante di osservazione e non dagli istanti precedenti.

Accertato questo, possiamo applicare i risultati visti per le catene di Markov tempo-continue⁷. In particolare, possiamo applicare l'equazione di bilancio del flusso globale, che in questo caso assume l'espressione

$$\sum_{i \neq j} \lambda_{ij} p_i = p_j \sum_{i \neq j} \lambda_{ji} \quad j=0,1,\dots,n-1,n$$

Le frequenze di transizione in avanti sono tutte pari a λ , mentre quelle di transizione all'indietro sono pari μq_1 . Al variare di j tra 0 ed n , abbiamo allora quanto segue:

$$\begin{aligned} j=0 &\longrightarrow \lambda \cdot p_0 = \mu q_1 \cdot p_1 \\ j=1 &\longrightarrow (\lambda + \mu q_1) \cdot p_1 = \lambda \cdot p_0 + \mu q_1 \cdot p_2 \\ &\dots\dots\dots \end{aligned}$$

Dalla prima equazione si ricava evidentemente che $p_1 = \frac{\lambda}{\mu q_1} p_0$; sostituendo questa espressione nella seconda equazione, si ottiene

$$p_2 = \frac{\lambda}{\mu q_1} p_1 = \frac{\lambda}{\mu q_1} \frac{\lambda}{\mu q_1} p_0 = \left(\frac{\lambda}{\mu q_1} \right)^2 p_0$$

La procedura è chiaramente iterativa: per il generico stato j -simo, si ottiene

$$p_j = \left(\frac{\lambda}{\mu q_1} \right)^j p_0 \quad j=0,1,\dots,n-1,n$$

Abbiamo dunque $n+1$ probabilità di stato, tutte dipendenti da p_0 . Per ricavare quest'ultima probabilità, ci basta allora imporre la condizione di normalizzazione

$\sum_{j=0}^n p_j = 1$: abbiamo infatti che

$$\sum_{j=0}^n p_j = 1 \longrightarrow \sum_{j=0}^n \left(\frac{\lambda}{\mu q_1} \right)^j p_0 = 1 \longrightarrow p_0 \sum_{j=1}^n \left(\frac{\lambda}{\mu q_1} \right)^j = 1 \longrightarrow p_0 = \frac{1}{\sum_{j=1}^n \left(\frac{\lambda}{\mu q_1} \right)^j}$$

Nota p_0 , è dunque possibile conoscere tutte le probabilità di stato, per cui la catena di Markov può dirsi risolta.

Naturalmente, i risultati ottenuti possono poi essere sfruttati per calcolare alcuni indici di prestazione del sistema. Ad esempio calcoliamo il traffico in uscita

⁷ Il procedimento che ci accingiamo a seguire sarà ripetute numerose volte nel capitolo sui **Sistemi a Coda**.

(**throughput**) della CPU, ossia il numero medio S di programmi serviti dalla CPU nell'unità di tempo: dato che la CPU può servire un programma per volta e lo fa con un tempo medio di servizio μ e dato che l'unico caso in cui la CPU non serve nessuno è quello in cui nessun programma è nella CPU, deduciamo che

$$S = \mu \cdot (1 - p_0) + 0 \cdot p_0 = \mu \cdot (1 - p_0)$$

In questo senso, la quantità $1 - p_0$ prende il nome di **probabilità di utilizzazione** della CPU.

Autore: **SANDRO PETRIZZELLI**
e-mail: sandry@iol.it
sito personale: <http://users.iol.it/sandry>