7. Simulare di una coda

L'obiettivo della presente analisi è descrivere l'evoluzione di una coda attraverso un modello matematico appropriato. Una coda può essere concettualizzata nel seguente modo:

Si consideri un sistema in cui viene erogato un servizio. In ogni istante di tempo, una sola persona può essere servita, mentre le altre in attesa vengono organizzate in una coda.

Ogniqualvolta un individuo tenta di accedere al servizio, è necessario verificare:

- Il numero di persone già in coda.
- Se un altro individuo è attualmente in fase di servizio.

Sulla base di queste condizioni, si decide se servire immediatamente la nuova persona oppure inserirla in fondo alla coda.

Nota. Il modello presentato non è generale. Sono possibili estensioni, come la gestione di code concorrenti o la presenza di più sportelli di servizio.

Categorizzazione della coda

Per caratterizzare il comportamento della coda, occorre definire i seguenti parametri:

1. La probabilità con cui si verificano nuovi arrivi.
2. La distribuzione probabilistica del tempo di servizio.

Assumendo che gli arrivi seguano un **processo di Poisson**, la probabilità che si verifichino k arrivi in un intervallo di tempo T è data da:

$$P[\text{arrivi in tempo } T = k] = e^{-\lambda T} \frac{(\lambda T)^k}{k!}$$

Inoltre, si assume che il tempo di servizio sia una variabile aleatoria, con distribuzione:

$$P(S_n \leq x) = B(x)$$

dove S_n rappresenta il tempo di servizio del cliente n-esimo.

A questo punto, è possibile modellare il fenomeno attraverso un **processo di Markov**, registrando i tempi in cui i clienti lasciano la coda (ovvero, quando vengono serviti). Si definisce Q_n come la lunghezza della coda quando l'n-esimo cliente viene servito e libera il posto per il cliente successivo.

Strategia di simulazione

Una possibile strategia per la simulazione prevede la gestione esplicita degli **eventi** significativi, ovvero:

- t_a: tempo del prossimo **arrivo**
- t_s: tempo del prossimo completamento del **servizio**
- T: tempo in cui la simulazione si interrompe

Il tempo del **prossimo evento** è determinato come segue:

$$t_e = \min(t_s, t_a, T)$$

La procedura della simulazione si articola nei seguenti passi principali:

MAIN STEP

1. Avanzare nel tempo, ponendo $t = t_e$.
2. Se $t_e = t_a$, eseguire il **ARRIVAL STEP**.
3. Se $t_e = t_s$, eseguire il **SERVICE STEP**.
4. Se $t_e = T$, terminare la simulazione.

ARRIVAL STEP

1. Generare il tempo fino al prossimo arrivo, a.
2. Impostare $t_a = t + a$.
3. Se la coda è **vuota** e **nessuno** è attualmente servito, il nuovo arrivato viene immediatamente servito:
   - Generare il tempo di servizio s.
   - Impostare $t_s = t + s$.
4. Altrimenti, aggiungere il nuovo cliente alla coda.
5. Ritornare al **MAIN STEP**.

SERVICE STEP

1. Se non vi sono persone in coda, impostare $t_s = \infty$.
2. Se vi è almeno una persona in coda:
   - Ridurre la lunghezza della coda di 1.
   - Generare il tempo di servizio s.
   - Impostare $t_s = t + s$.
3. Ritornare al **MAIN STEP**.

Generazione di dati secondo una distribuzione probabilistica

Data una densità di probabilità $f(x)$ che soddisfa le seguenti proprietà:

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$ *(normalizzazione)*

Nota. Gli estremi saranno al più "ridondanti". Strettamente parlando, questo include $\int_a^b f(x)\, dx = 1$ se X assume valori su $[a, b]$ e $f(x) = 0$ per $x \notin [a, b]$

La funzione di distribuzione associata è

$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy$

Poiché $F(x)$ è una funzione crescente, essa è invertibile e consente la generazione di dati nel modo seguente:

1. Generare un numero casuale uniforme $\varepsilon \in [0, 1]$.
2. Calcolare $X = F^{-1}(\varepsilon)$.

In questo modo, X risulta distribuito secondo $f(x)$.

Nota. Se $F^{-1}(\varepsilon)$ non è determinabile analiticamente, è possibile adottare un approccio numerico.

Distribuzione dei tempi tra arrivi successivi in coda

Un'alternativa alla distribuzione Poissoniana per modellare gli arrivi può essere ottenuta risolvendo un'equazione differenziale.

La probabilità di **nessun arrivo** entro il tempo t è data da

$P_0(t) = e^{-\lambda t}$

Questo risultato deriva dalla valutazione della distribuzione di Poisson per zero arrivi. Da esso si deduce che:

$P(a > t) = e^{-\lambda t} = P_0(t)$

$P(a \leq t) = 1 - e^{-\lambda t}$

Di conseguenza, la **densità di probabilità** dei tempi di intercorrenza è:

$f(x) = \lambda e^{-\lambda x}$

Analogamente, per i tempi di servizio si può assumere:

$P(S_n \leq x) = B(x) = 1 - e^{-\mu x}$

Dove $B(x)$ rappresenta la probabilità che il tempo di servizio del cliente n-esimo sia inferiore a x.

Infine, si noti che i **tempi di intercorrenza** Z_n, definiti come:

$Z_n = t_n - t_{n-1}$

dove t_n è il tempo dell'n-esimo evento, seguono la distribuzione:

$P(Z_n \leq x) = 1 - e^{-\lambda x}$

La matrice stocastica per la coda

Per descrivere l'evoluzione della lunghezza della coda, abbiamo formulato un'**equazione di evoluzione** da cui è stato possibile derivare la **matrice stocastica** che descrive il sistema. La struttura della matrice dipende dalla presenza o meno di una **sala di attesa**:

- **Con sala di attesa**: La matrice stocastica è finita, garantendo l'esistenza della **distribuzione stazionaria**.
- **Senza sala di attesa**: La matrice è infinitamente dimensionale e l'esistenza della distribuzione stazionaria non è garantita.

Abbiamo verificato numericamente che, nel caso con sala di attesa e tempo di servizio costante, il sistema converge a una distribuzione stazionaria, confermando i risultati della teoria dei processi stocastici. In particolare, questa distribuzione è l'autovettore della matrice stocastica associato all'autovalore 1, normalizzato opportunamente.

Una volta definito il modello generale, abbiamo esteso le simulazioni anche al caso **senza sala di attesa**, semplicemente rimuovendo una condizione dal codice. Questo ha permesso di osservare che:

- Quando il numero medio di arrivi è inferiore alla capacità di servizio, si ottiene una distribuzione stazionaria.
- Quando il numero medio di arrivi supera la capacità di servizio, la coda cresce indefinitamente, impedendo il raggiungimento di uno stato stazionario.
- Esiste un **valore di soglia** che distingue i due regimi: quando il tasso di arrivi e il tasso di servizio sono bilanciati, il sistema è al limite tra stabilità e crescita indefinita.

Definizione della configurazione della coda

Abbiamo adottato come **parametro di configurazione**

$Q_n = \text{lunghezza della coda al completamento del servizio } n-\text{esimo}$

La lunghezza della coda è una scelta naturale per descrivere lo stato del sistema, ma diventa significativa solo se si specifica **quando** viene misurata. Abbiamo adottato il **punto di vista dello sportello**, in cui la lunghezza viene registrata immediatamente prima dell'inizio di un nuovo servizio.

A seconda della presenza o meno di una sala di attesa, lo spazio delle configurazioni assume caratteristiche differenti:

Caso (a): con sala di attesa

- Se la sala di attesa ha N posti, lo spazio delle configurazioni è finito e comprende $N+1$ stati: $Q_n = 0, 1, \ldots, N$
- La distribuzione stazionaria esiste sicuramente

Caso (b): senza sala di attesa

- Lo spazio delle configurazioni è infinito: $Q_n = 0, 1, 2, \ldots$
- La distribuzione stazionaria potrebbe non esistere

Evoluzione della lunghezza della coda

Possiamo descrivere l'evoluzione di Q_n con l'equazione:

$$Q_{n+1} = Q_n - H(Q_n) + X_{n+1}$$

dove:

- $H(Q_n)$ è la **funzione di Heaviside**, che assicura che la coda si accorci solo se è maggiore di zero.
- X_n è il numero di arrivi durante il servizio n-esimo.

La probabilità che si verifichino j arrivi durante un servizio dipende dalla distribuzione della durata del servizio S_n. Sia $B(x)$ la **funzione di distribuzione**:

$$B(x) = P(S_n < x)$$

con densità di probabilità

$$B'(x) = \frac{d}{dx} B(x)$$

La probabilità di osservare esattamente j arrivi è data da:

$$K_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t)$$

Dove $e^{-\lambda t} \frac{(\lambda t)^j}{j!}$ è la probabilità di avere j arrivi nel tempo t e $dB(t)$ è la probabilità che il servizio duri un tempo t.

Matrice stocastica W

Gli elementi della matrice stocastica W sono definiti come:

$$W_{ij} = P(i \leftarrow j)$$

dove i e j rappresentano la lunghezza della coda prima e dopo un evento di servizio.

- Per $j = 0$ (coda inizialmente vuota): $W_{i,0} = P(Q_{n+1} = i | Q_n = 0) = K_i$
- Per $j > 0$: $W_{ij} = P(Q_{n+1} = i | Q_n = j) = K_{i-j+1} \quad \text{se } i - j + 1 \geq 0$

La matrice risultante ha la forma

$$\begin{pmatrix} K_0 & K_0 & 0 & 0 & 0 & \dots \\ K_1 & K_1 & K_0 & 0 & 0 & \dots \\ K_2 & K_2 & K_1 & K_0 & 0 & \dots \\ K_3 & K_3 & K_2 & K_1 & K_0 & \dots \\ K_4 & K_4 & K_3 & K_2 & K_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Il problema diventa allora determinare se esiste un vettore Π tale che:

$W\Pi = \Pi$

dove Π rappresenta la distribuzione stazionaria, ovvero l'autovettore di W corrispondente all'autovalore 1.

Nel caso in cui ci sia una sala d'attesa con N posti, il massimo valore di Q_m è N, quindi la matrice stocastica W ha dimensione finita $(N+1) \times (N+1)$ e assume la forma:

$$W = \begin{pmatrix} K_0 & K_0 & 0 & 0 & \dots \\ K_1 & K_1 & K_0 & 0 & \dots \\ K_2 & K_2 & K_1 & K_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ K_{N-1} & K_{N-1} & K_{N-2} & K_{N-3} & \dots \\ 1 - \sum_{i=0}^{N-1} K_i & 1 - \sum_{i=0}^{N-1} K_i & 1 - \sum_{i=0}^{N-2} K_i & 1 - \sum_{i=0}^{N-3} K_i & \dots \end{pmatrix}$$

Dove l'ultima riga è fissata dalla condizione $\sum_i W_{ij} = 1, \quad \forall j$

Caso particolare: tempo di servizio costante

Se il tempo di servizio è costante, $S_n = s$, allora:

$K_j = P(X_n = j) = e^{-\lambda s} \frac{(\lambda s)^j}{j!}$

cioè segue una distribuzione di Poisson. In tal caso, la matrice W può essere calcolata in forma chiusa e utilizzata direttamente per determinare la distribuzione stazionaria.