Esercitazione del corso di Calcolo delle probabilità e Statistica

REGOLE GENERALI

- Ogni studente deve svolgere una esercitazione: può essere scelta fra quelle proposte o può essere concordata con il docente.
- L'esame orale verterà principalmente sulla discussione dell'esercitazione scelta.
- Deve essere presentata o una breve relazione scritta o una presentazione in PowerPoint (o simili).
  - Ogni studente avrà 5 minuti per presentare il proprio lavoro;
  - Lo studente dovrà essere in grado di giustificare il modello probabilistico adottato, le distribuzioni scelte, le assunzioni fatte...
- Deve essere fornito il codice sorgente
  - Non ci sono vincoli sul linguaggio di programmazione (si suggerisce C++ o Python);
  - È ammesso avvalersi di strumenti per la generazione di codice;
  - In ogni caso, gli studenti dovranno capire ed essere in grado di spiegare struttura e funzione del codice, spiegando come ogni sezione contribuisce a raggiungere l'obiettivo fissato.
- Per la generazione dei numeri casuali possono essere utilizzate solo le funzioni di default che generano numeri casuali secondo una **distribuzione uniforme fra 0 e 1** o una **distribuzione normale standard**.
  - Le altre distribuzioni vanno generate coi metodi visti in aula.
  - Le distribuzioni normali non standard vengano generate a partire dalla standard.
- Il lavoro può essere svolto in gruppi di due o (massimo) tre studenti.
  - In tal caso, gli studenti sono invitati a presentarsi allo stesso appello orale;
  - La presentazione sarà unica, di durata proporzionale al numero di studenti; la durata dovrà essere equamente suddivisa fra i membri del gruppo;
  - Ciascuno studente dovrà conoscere ogni aspetto del lavoro, inclusi quelli presentati dagli altri.

1 SISTEMA DI CODE

Implementare una simulazione Monte Carlo per analizzare e ottimizzare la gestione di un sistema di code in un centro di elaborazione dati (ad esempio, un sistema di server che gestisce richieste da parte degli utenti). L'obiettivo è valutare le performance del sistema e svolgere una analisi statistica dei risultati ottenuti.

1.1 SCENARIO

Considerare un centro di elaborazione dati che riceve richieste da utenti esterni secondo un **processo di arrivo con un tasso medio di arrivo** λ (richieste al secondo). Le richieste vengono elaborate da N server indipendenti, ciascuno con un **tempo di servizio esponenziale** con parametro μ (richieste elaborate al secondo).

Quando un utente invia una richiesta:

- Verificare se c'è un server libero: in tal caso, il server prende in carico la richiesta ed è occupato fino al completamento dell'elaborazione;
- Se tutti i server sono occupati, mettere la richiesta in coda in attesa di essere processata dal primo server che si libera.

1.2 OBIETTIVO

1. **Simulare il sistema di code** usando il Metodo Monte Carlo: andranno simulati l'arrivo delle richieste e il tempo di servizio.
2. **Raccogliere statistiche rilevanti** in funzione del tempo T:
    a. Stimare il tempo medio di attesa di una richiesta in coda;
    b. Stimare la probabilità di servizio immediato (percentuale di richieste elaborate senza attesa);
    c. Stimare la percentuale di tempo di occupazione di ciascun server.
3. **Analizzare le prestazioni**
    a. Confrontare i risultati ottenuti per diversi valori di N;
    b. Valutare come il tempo medio di attesa e la probabilità di servizio immediato variano al variare di λ e μ.
4. **Ottimizzare il sistema**, cioè trovare il numero ottimale di server N per mantenere il tempo medio di attesa al di sotto di una soglia prefissata, con determinati λ e μ; mostrare che se N scende sotto una certa soglia, la coda tende a crescere indefinitamente
5. **Effettuare una analisi statistica** del tempo medio di attesa a regime (stabilire un tempo simulato T lungo abbastanza e ripetere la simulazione K volte).
    a. Stimare attesa e varianza del tempo di attesa di una richiesta;
    b. Per K=50, 100, 200, 2000, fornire un intervallo di confidenza del 95% per il tempo di attesa;
    c. Calcolare 50 media campionarie usando campioni di 10, 30, 50, 100 prove e tracciarne le rispettive distribuzioni: mostrare che, specie se il campione è grande, valgono i risultati del Teorema Limite Centrale.
    d. Verificare se il tempo di attesa dipenda linearmente da N, da λ o da μ.

1.3 ESTENSIONI

Le seguenti estensioni sono facoltative: l'introduzione di una o più di queste sarà valutata positivamente nell'assegnazione del voto finale.

- Ripetere il processo considerando un tempo di servizio normale, invece che esponenziale; si ricordi che il tempo di servizio non può essere negativo.
- Considerare l'arrivo di due flussi paralleli di richieste (con λ diversi), il cui tempo di servizio è esponenziale, ma con tempo di servizio diversi.
- Si consideri che ogni richiesta abbia una certa probabilità di essere inoltrata ad un singolo server "superiore" (con il suo tempo di servizio): il server inoltrante sarà occupato finché il superiore non avrà elaborato la richiesta.

2 PROCESSI DI DIRAMAZIONE

Implementare un programma per simulare un processo di diramazione utilizzando il metodo Monte Carlo. Lo scopo è studiare il comportamento stocastico di una popolazione e calcolare statistiche rilevanti come la probabilità estinzione della popolazione e la sua distribuzione in un determinato istante.

2.1 SCENARIO

Una funzione esegue le seguenti istruzioni (tutte senza alcun ritardo):

1. Attendere 1 minuto;
2. Lanciare X copie di sé stessa;
3. Terminarsi.

X è una **variabile aleatoria distribuita secondo:** $X \sim Bin(n = 3, p = 0.4)$.

All'istante iniziale lanciate una copia della funzione. Il processo si estingue se, in qualsiasi istante di tempo, non restano più funzioni attive.

2.2 OBIETTIVO

1. **Simulare il numero di funzioni avviate a ogni minuto** usando il Metodo Monte Carlo: andrà simulato il numero di copie che ciascuna funzione attiva avvierà al minuto successivo. Si simuli l'andamento per almeno due giorni (2880 minuti).
2. **Raccogliere statistiche rilevanti** in funzione del tempo T:
   a. Stimare la probabilità di estinzione;
   b. Stimare il numero di processi attivi;
3. **Analizzare le prestazioni**
   a. Valutare se i risultati ottenuti cambiano significativamente riducendo il tempo T a 3, 6, 12, 24, 72 h;
   b. Considerando i soli casi in cui il processo di estingue, calcolare il tempo atteso necessario per l'estinzione e la sua varianza.
4. **Verificare la soluzione teorica**
   a. Ripetendo la simulazione più volte, verificare che la probabilità di estinzione converge al valore *w* dato dall'unica soluzione interpretabile come una probabilità della seguente equazione

$$27 - 71w + 24w^2 + 8w^3 = 0$$

5. **Effettuare una analisi statistica** della probabilità di estinzione (ripetere la simulazione K volte).
   a. Stimare attesa e varianza della probabilità di estinzione;
   b. Per K=50, 100, 200, 2000, fornire un intervallo di confidenza del 95% per la probabilità di estinzione;
   c. Calcolare 50 media campionarie usando campioni di 10, 30, 50, 100 prove e tracciarne le rispettive distribuzioni: mostrare che, specie se il campione è grande, valgono i risultati del Teorema Limite Centrale.
   d. Tracciare un grafico della probabilità di estinzione al variare della probabilità *p* della distribuzione considerata per *X*. Tracciare una retta di regressione e valutarne la significatività.

3 PROBLEMA DELLA ROVINA DEL GIOCATORE

Implementare un programma per simulare il problema della rovina del giocatore d'azzardo. L'obiettivo è stimare sperimentalmente la probabilità di rovina in diversi scenari e verificare la coerenza con le previsioni teoriche.

3.1 SCENARIO

Un giocatore d'azzardo parte con un capitale iniziale C_0 e gioca una serie di partite indipendenti con esito dicotomico. In ogni partita:

- Il giocatore può vincere o perdere una somma fissa δ (ad esempio, 1 unità).
- La probabilità di vincere una singola partita è p (con $0<p<1$).

Il giocatore continua a giocare fino a quando:

- Raggiunge un capitale massimo prefissato *C_max* (successo).
- Perde tutto il capitale (rovina).

3.2 OBIETTIVO

1. **Simulare il capitale del giocatore** nel tempo usando un approccio Monte Carlo: ad ogni passo andrà simulato il risultato della singola partita e aggiornato il capitale. Si imposti un numero massimo di partite (per evitare cicli infiniti) di almeno 50.000 partite.
2. **Raccogliere statistiche rilevanti** in funzione del tempo T:
   a. Stimare le probabilità di rovina e di successo;
   b. Si traccino gli andamenti del capitale del giocatore in funzione del tempo;
3. **Analizzare le prestazioni**
   a. Valutare se i risultati ottenuti cambiano significativamente riducendo il numero di partite massimo;
   b. Considerando i soli casi in cui il giocatore si rovina, calcolare il tempo atteso necessario per la rovina.
   c. Considerando i soli casi in cui il giocatore ha successo, calcolare il tempo atteso necessario per il successo.
4. **Verificare la soluzione teorica**
   a. Ripetendo la simulazione più volte, verificare che la probabilità di rovina converge alla soluzione teorica

$$P_{\text{rovin}} = \begin{cases} \dfrac{1-\left(\frac{p}{1-p}\right)^{C_0}}{1-\left(\frac{p}{1-p}\right)^{C_{\text{max}}}} & \text{se } p \neq 0.5, \\ 1 - \dfrac{C_0}{C_{\text{max}}} & \text{se } p = 0.5. \end{cases}$$

   b. Ripetere l'esercizio nel caso in cui il giocatore non può raggiungere il successo (cioè C_max è infinito): provare tutti i casi con p=0.05, 0.15, 0.25, ..., 0.85, 0.95.
   c. Verificare che se p>1/2 la rovina non è certa.

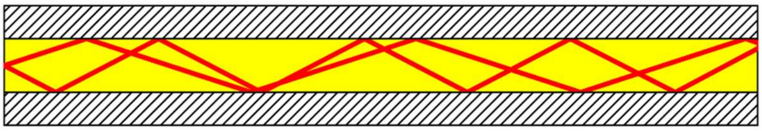
5. **Effettuare una analisi statistica** della probabilità di estinzione (ripetere la simulazione K volte).
    a. Stimare attesa e varianza della probabilità di rovina;
    b. Per K=50, 100, 200, 2000, fornire un intervallo di confidenza del 95% per la probabilità di rovina;
    c. Calcolare 50 media campionarie usando campioni di 10, 30, 50, 100 prove e tracciarne le rispettive distribuzioni: mostrare che, specie se il campione è grande, valgono i risultati del Teorema Limite Centrale.
    d. Tracciare un grafico della probabilità di rovina al variare della probabilità *p* della distribuzione considerata per *X*. Tracciare una retta di regressione e valutarne la significatività.

4 TRASMISSIONE DELLA LUCE

Implementare un programma per simulare la trasmissione della luce lungo una fibra ottica. L'obiettivo è stimare sperimentalmente la frazione di luce che raggiunge la destinazione a seconda delle differenti condizioni della fibra.

4.1 SCENARIO

Si consideri un modello semplificato (2D) di una fibra ottica diritta, caratterizzata da una lunghezza L e un diametro D, come in figura. Si ignori la terza dimensione (profondità)



Si modellizzi la luce come particelle che:

- Sono generate all'ingresso della fibra con un angolo casuale
- Viaggiano in linea retta finché
  - Interagiscono con l'aria nella fibra: il libero cammino medio LCM è un dato di ingresso; ad ogni interazione le particelle hanno probabilità a di essere assorbite (e quindi terminare la loro corsa;
  - Incontrano la parete: in tal caso vengono di norma riflesse, ma hanno probabilità di d di venire diffuse (cioè l'angolo di uscita è casuale con distribuzione uniforme);
  - Escono dal lato opposto del tubo.

4.2 OBIETTIVO

1. **Simulare il tragitto delle particelle** usando un approccio Monte Carlo. Ci si limiti a tracciare il tragitto delle particelle, senza considerare la variabile temporale.
2. **Raccogliere statistiche rilevanti**:
   a. Stimare il valore di T, frazione di particelle che vengono trasmesse;
   b. Stimare la distribuzione di particelle in uscita lungo il diametro;
   c. Simulare un numero di particelle sufficienti a che l'intervallo T±0.01 contenga il valore reale con probabilità del 95%
3. **Analizzare le prestazioni**
   a. Valutare come cambiano i risultati al variare di a e d;
   b. Valutare come cambiano i risultati se la luce in ingresso è polarizzata (ossia entra con direzione parallela all'asse della fibra, invece che con angolo casuale)
4. **Effettuare una analisi statistica** della frazione di particelle trasmesse.
   a. Stimare attesa e varianza della frazione di particelle trasmesse;
   b. Tracciare un grafico della frazione di particelle trasmesse al variare del valore LCM (si considerino valori di LCM di 0.05, 0.1, 0.2, 0.5, 0.75, 1, 1.2, 1.5, 2, 5, 10 volte la lunghezza L). Tracciare una retta di regressione e valutarne la significatività.

5 INCROCIO STRADALE

Implementare un programma per simulare un incrocio fra due strade a senso unico, regolato da un semaforo. L'obiettivo è ottimizzare la sequenza semaforica che minimizza il tempo di attesa in coda.

5.1 SCENARIO

Si consideri un quadrivio formato dall'intersezione di due strade (una Nord-Sud, l'altra Est-Ovest) a senso unico e singola corsia. Le auto arrivano all'incrocio secondo un **processo di arrivo con un tasso medio di arrivo** λ_NS o λ_EW (auto al minuto).

Quando un'auto arriva:

- Se il semaforo in quella direzione è verde e non ci sono auto in coda, l'auto passa;
- In caso contrario si aggiunge alla coda.

Nel momento in cui il semaforo diventa verde:

- Si perde un tempo di reazione μ (distribuzione uniforme fra 0 e 8 secondi);
- Le auto in coda passano, nell'ordine di arrivo, impiegando l'incrocio per un tempo α (distribuzione uniforme fra 2 e 5 secondi), fino all'arrivo del rosso (si ignori la fase di giallo).

Si assuma che il ciclo semaforico sia di 2 minuti: il tempo di verde dell'asse Nord-Sud t_NS sia inizialmente di 1 minuto.

5.2 OBIETTIVO

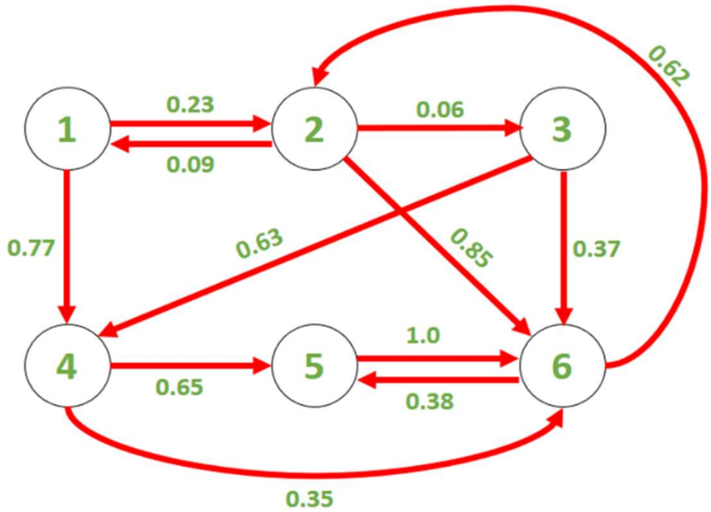
1. **Simulare il sistema di code** usando il Metodo Monte Carlo: andranno simulati l'arrivo delle auto e i tempi di attraversamento.
2. **Raccogliere statistiche rilevanti** in funzione di λ_NS e λ_EW:
    a. Stimare il tempo medio di attesa di un'auto in ciascuna strada;
    b. Stimare la probabilità di attraversamento immediato;
    c. Stimare la percentuale di tempo in cui ci sono auto in attesa con il rosso e la strada col verde non ha auto.
3. **Ottimizzare il sistema**
    a. Trovare il t_NS ottimale per mantenere il tempo medio di attesa, con determinati λ_NS e λ_EW, equo fra le due strade;
    b. Mostrare che se λ_NS+λ_EW supera una certa soglia, la coda tende a crescere indefinitamente.
4. **Effettuare una analisi statistica** del tempo medio di attesa a regime (stabilire un tempo simulato T lungo abbastanza e ripetere la simulazione K volte).
    a. Stimare attesa e varianza del tempo di attesa;
    b. Per K=50, 100, 200, 2000, fornire un intervallo di confidenza del 95% per il tempo di attesa;
    c. Calcolare 50 media campionarie usando campioni di 10, 30, 50, 100 prove e tracciarne le rispettive distribuzioni: mostrare che, specie se il campione è grande, valgono i risultati del Teorema Limite Centrale.
    d. Verificare se il tempo di attesa dipenda linearmente da N, da λ_NS+λ_EW.

6 ALGORITMO PAGERANK

Implementare una catena di Markov che simuli un algoritmo PageRank su un grafo che rappresenta un ipertesto. L'obiettivo è verificare sperimentalmente che la distribuzione di utenti, assumendo che questi si spostino seguendo un link a caso della pagina in cui si trovano, converge a una distribuzione invariante.

6.1 SCENARIO

Si consideri un ipertesto costituito da 6 pagine, collegate da link come in figura; i numeri sulle frecce indicano la probabilità che un utente clicchi su quello specifico link.



6.2 OBIETTIVO

1. **Simulare la catena di Markov** che descrive la visita di un utente su una pagina:
    a. L'utente arriva su una pagina a caso, scelta uniformemente fra le 6;
    b. Ad ogni passo si sposta ad un'altra pagina collegata, con probabilità proporzionale alla frequenza indicata sulla freccia;
2. **Raccogliere statistiche rilevanti**:
    a. Stimare la frequenza di visite per ciascuna pagina
    b. Verificare che la distribuzione delle frequenze, dopo alcuni passi, diventa indipendente dalla pagina di partenza.
3. **Considerare la possibilità che la passeggiata aleatoria termini**:
    a. Implementare il meccanismo per cui dopo ogni passo (incluso il passo numero 0), l'utente ha una probabilità p di concludere la passeggiata.
    b. Simulare numerosi utenti e tracciare le frequenze delle pagine al variare di p: osservare come cambia la distribuzione delle frequenze.
4. **Effettuare una analisi statistica** della frequenza con cui viene visitata la pagina 2; si consideri il caso p=0.01.
    a. Stimare attesa e varianza;
    b. Ripetere la passeggiata aleatoria di un utente K=50, 100, 200, 2000 volte (troncare ogni utente dopo 100 passi): fornire un intervallo di confidenza del 95% per la frequenza;
    c. Calcolare 50 frequenze usando campioni di 10, 30, 50, 100 prove e tracciarne le rispettive distribuzioni: mostrare che, specie se il campione è grande, valgono i risultati del Teorema Limite Centrale.