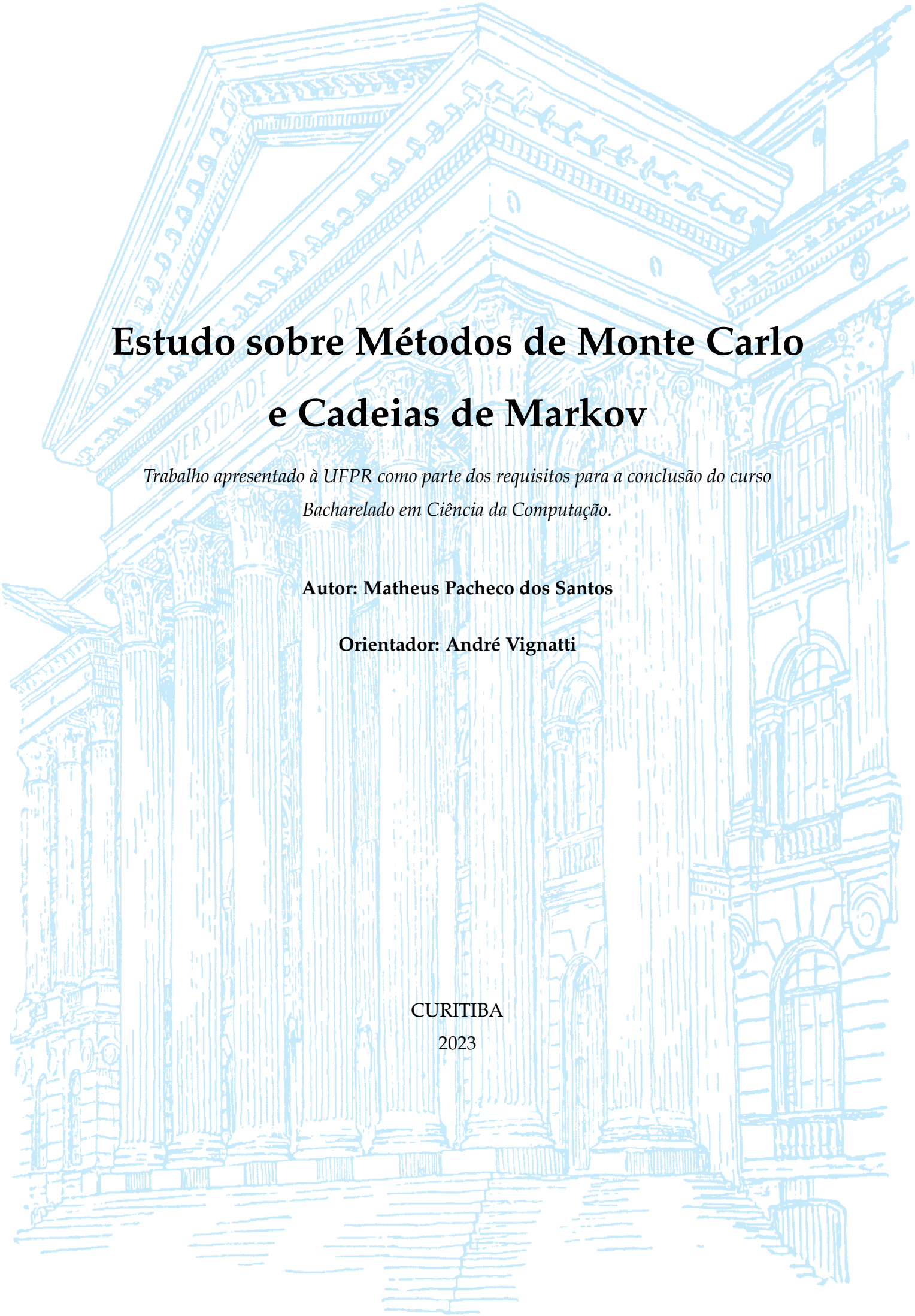


Estudo sobre Métodos de Monte Carlo e Cadeias de Markov

Trabalho apresentado à UFPR como parte dos requisitos para a conclusão do curso Bacharelado em Ciência da Computação.

Autor: Matheus Pacheco dos Santos

Orientador: André Vignatti

CURITIBA

2023

Resumo

Este trabalho explora os Métodos de Monte Carlo em conjunto com Cadeias de Markov, com foco nos algoritmos Metrópolis-Hastings e Gibbs Sampling. Essas técnicas são aplicadas para amostragem de distribuições complexas, sendo especialmente relevantes na inferência bayesiana. A análise aborda a questão crucial do tempo de mistura, destacando sua importância na geração de amostras precisas. Apesar de oferecer uma visão geral abrangente, este trabalho reconhece a evolução constante do campo e aponta para desafios futuros, como a determinação precisa do tempo de mistura.

Palavras-chave: Métodos de Monte Carlo, Cadeias de Markov, Metrópolis-Hastings, Gibbs Sampling, Inferência Bayesiana, Tempo de Mistura.

Abstract

This work explores Monte Carlo Methods in conjunction with Markov Chains, focusing on the Metropolis-Hastings and Gibbs Sampling algorithms. These techniques are applied for sampling complex distributions, with particular relevance in Bayesian inference. The analysis addresses the crucial issue of mixing time, emphasizing its importance in generating accurate samples. Despite providing a comprehensive overview, this work acknowledges the constant evolution of the field and points to future challenges, such as the precise determination of mixing time.

Keywords: Monte Carlo Methods, Markov Chains, Metropolis-Hastings, Gibbs Sampling, Bayesian Inference, Mixing Time.

Conteúdo

Lista de Figuras

1 Introdução

A presente pesquisa tem como foco de investigação as Cadeias de Markov e os Métodos de Monte Carlo, com o propósito de compreender o renomado algoritmo Monte Carlo por Cadeias de Markov (MCMC).

Cadeias de Markov, batizadas em homenagem ao matemático russo Andrei Markov, constituem modelos estocásticos que descrevem uma sequência de eventos, onde a probabilidade de transição entre estados depende exclusivamente do estado atual, desconsiderando a sequência de eventos precedentes. Ampla e diversamente aplicadas em campos como ciência da computação, estatística, economia e modelagem de fenômenos naturais, essas cadeias representam estados possíveis, regidos por probabilidades de transição. Sua propriedade notável é a independência condicional, onde a transição para um próximo estado depende unicamente do estado atual. Essa característica fundamental torna as Cadeias de Markov ferramentas essenciais para modelar sistemas dinâmicos e prever comportamentos futuros, sendo cruciais em análise de dados e teoria de probabilidade.

O Método de Monte Carlo, uma abordagem estatística e computacional amplamente empregada para resolver problemas complexos envolvendo incerteza e aleatoriedade, utiliza técnicas de amostragem aleatória para estimar soluções numéricas. Nomeado em referência ao famoso cassino de Monte Carlo, esse método é aplicado em diversas áreas, como física, finanças, engenharia e ciências da computação. Sua eficácia reside na habilidade de lidar com problemas multidimensionais complexos e integrar soluções para equações sem soluções analíticas diretas. Ao simular eventos aleatórios repetidamente, o método produz estimativas numéricas confiáveis, tornando-se uma ferramenta valiosa para análise e tomada de decisões em situações onde a modelagem analítica tradicional pode ser desafiadora.

O Método de Monte Carlo por Cadeias de Markov (MCMC) é uma técnica estatística que utiliza o algoritmo Metropolis-Hastings para amostrar dados de distribuições complexas. Esse algoritmo realiza uma "caminho aleatório" por diferentes valores possíveis, favorecendo os mais prováveis. Essa abordagem é particularmente útil na estimativa de parâmetros desconhecidos em modelos estatísticos complexos, especialmente na estatística bayesiana. A relevância do MCMC destaca-se em sua aplicação generalizada em áreas como aprendizado de máquina e bioinformática, oferecendo uma maneira eficiente de explorar e amostrar distribuições de probabilidade difíceis de serem analisadas diretamente.

Este trabalho foi fundamentado em um curso da UFRJ, intitulado "Algoritmos de Monte Carlo e Cadeias de Markov", ministrado por Daniel Ratton[1]. Apresentamos o Método de Monte Carlo, que utiliza a aleatoriedade para obter soluções aproximadas. Discutimos as Cadeias de Markov e seus conceitos fundamentais, como aperiodicidade, irredutibilidade, reversibilidade, distribuição estacionária, tempo de mistura e spectral gap, demonstrando também como simular Cadeias de Markov de forma eficiente. Por fim, alcançamos nosso objetivo ao apresentar o algoritmo MCMC, destacando o Metropolis-Hastings e o Gibs Sampling.

[1] https://www.cos.ufrj.br/~daniel/

2 Método de Monte Carlo

Nesta seção, abordaremos o **Método de Monte Carlo**, uma técnica estatística e computacional poderosa para estimativa e simulação. Vamos explorar desde a estimativa de somatórios até aplicações práticas, como a Integração de Monte Carlo.

2.1 Visão Geral

O **Método de Monte Carlo** é uma classe de algoritmos baseada em amostragem aleatória repetida, com o objetivo de obter uma solução aproximada para problemas determinísticos. A ideia central é que um grande número de amostras repetidas acaba por revelar a solução. A base teórica é a **Lei dos Grandes Números** (veja o Teorema D.5).

A técnica de Monte Carlo tem origens notáveis na Segunda Guerra Mundial, durante o Projeto Manhattan, quando cientistas como Stanislaw Ulam e John von Neumann trabalhavam em problemas relacionados à difusão de nêutrons. Ulam teve a ideia enquanto se recuperava de uma doença e jogava paciência (solitário) para passar o tempo [1].

A abordagem de Monte Carlo recebeu esse nome em referência ao famoso cassino Monte Carlo em Mônaco, conhecido por jogos de azar e aleatoriedade. Ulam e von Neumann perceberam que a simulação de eventos aleatórios usando métodos estocásticos poderia ser aplicada para resolver uma variedade de problemas complexos. Uma leitura seminal sobre o assunto é fornecida por Ulam e von Neumann em seu trabalho pioneiro [2].

Atualmente, o método de Monte Carlo é amplamente utilizado para abordar diversos problemas, incluindo pesquisa em árvores [3], localização de robôs [4], entre outros desafios.

2.2 Estimando Somatório

Podemos estimar um valor de um somatório quando o número de parcelas é muito grande e o calcular o valor de cada parcela é fácil. Vamos estimar o somatório abaixo

$$G_N = \sum_{i=i}^{N} g(i)$$

utilizaremos a aleatoriedade com o artifício do valor esperado para aproximar o somatório.

Seja X uma v.a uniforme em $[1..N]$ ou seja

$$\Pr[X = i] = \frac{1}{N}, \forall i \in [1..N]$$

podemos calcular a esperança como:

$$\mathbb{E}[g(X)] = \sum_{i=1}^{N} \Pr[X = i]\, g(i) = \frac{1}{N} \sum_{i=1}^{N} g(i) = \frac{G_N}{N}$$

logo, temos que $G_N = N\mathbb{E}[g(x)]$. Podemos então gerar amostras e fazer a média para estimar $\mathbb{E}[g(x)]$.

Seja X_i uma sequência de v.a uniforme em $[1..N]$, escolhemos um valor $n < N$ para o número de amostras e fazemos a média amostral

$$\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

observe que

$$\begin{aligned}\mathbb{E}\big[\bar{M}_n\big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g(X_i)] \\ &= \frac{1}{n}n\mathbb{E}[g(X)] \\ &= \mathbb{E}[g(X)]\end{aligned}$$

logo $\mathbb{E}\big[\bar{M}_n\big] = \mathbb{E}[g(X)]$ para qualquer valor de n, ou seja pela lei dos grandes números temos que

$$\lim_{n\to\infty}\bar{M}_n = \mathbb{E}[g(X)]$$

portando concluímos que $N\bar{M}_n$ é um estimador para G_N.

2.2.1 Vantagens e Desvantagens

Quando substituir $G_N = \sum_{i=1}^{N} g(i)$ por $\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ é uma escolha vantajosa? Vários fatores devem ser considerados:

- número de parcelas (N);
- número de amostras (n);
- complexidade da função $g(i)$

Esta abordagem é apropriada

- Se computar o valor de $g(i)$ é eficiente, então é benéfico o uso quando N é grande.
- Se $g(i)$envolver cálculos computacionais intensivos é vantajoso usar mesmo com um valor menor de N.

Quando a função $g(i)$ é altamente "errática", o uso do método pode não ser recomendado. Por exemplo, se um determinado valor de $g(i)$ for significativamente maior que o restante da soma, o estimador pode se tornar impreciso, especialmente quando o número de amostras não é grande o suficiente.

2.3 Estimando PI

Podemos estimar o valor de π utilizando Monte Carlo[2]. A ideia central consiste em expressar π como uma relação entre áreas e, em seguida, utilizar o método de Monte Carlo para estimar essa relação.

Considere que

- $A_q = 1 :=$ Área do quadrado
- $A_c = \pi r^2 = \frac{\pi}{4} :=$ Área do círculo

expressamos então π como $\pi = 4\frac{A_c}{A_q}$.

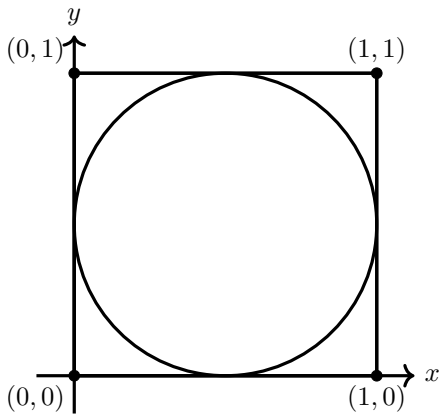


Figura 1: Exemplo do Círculo

Utilizando o exemplo ilustrado na Figura 1, podemos estimar $\frac{A_c}{A_q}$ gerando n pontos uniformemente distribuídos no quadrado e medindo a fração de pontos que se encontram dentro do círculo.

Sejam X e Y duas v.a contíguas uniformes em $[0..1]$ e $g(x, y)$ a função indicadora do ponto (x, y) estar dentro do círculo, podemos definir $g(x, y)$ como

$$g(x,y) = \begin{cases} 1 & \text{se } (x-0.5)^2 + (y-0.5)^2 \leq \frac{1}{4}, \\ 0 & \text{caso contrário} \end{cases}$$

Considerando sequências i.i.d. X_i, Y_i uniformemente distribuídas em $[0, 1]$, podemos obter n amostras e calcular a média amostral conforme apresentado abaixo:

$$\lim_{n\to\infty} \bar{M}_n = \frac{\pi}{4}$$

Dado que $\frac{A_c}{A_q} = \frac{\pi}{4}$ pela Lei dos Grandes Números, temos que

$$\lim_{n\to\infty} \bar{M}_n = \frac{\pi}{4}.$$

Assim, podemos estimar π por $4\bar{M}_n$.

[2]podemos estimar qualquer valor que tenha relação geométrica

2.3.1 Defininindo Precisão e Confiança

Vamos utilizar Chebyshev para para determinar o número de amostras necessário para atingir uma precisão $\epsilon = 10^{-4}$ com uma confiança de $\beta = 0.99$ (veja Subseção D.3). Para isso, precisamos calcular a variância σ^2, que é dada por:

$$\sigma^2 = \text{Var}[g(X, y)] = \left(\frac{\pi}{4}\right)\left(1 - \frac{\pi}{4}\right)$$

embora o valor exato de π seja desconhecido, sabemos que a variância de uma variável indicadora é máxima quando a probabilidade p é $\frac{1}{2}$, ou seja $\sigma^2 = \frac{1}{4}$.

Substituindo essa variância na desigualdade de Chebyshev, obtemos:

$$\begin{aligned}
1 - \frac{0.25}{(10^{-4})^2\, n} &= 0.99 \\
\frac{0.25}{10^{-8} n} &= 0.01 \\
\frac{1}{10^{-8} n} &= 0.04 \\
10^{-8} n &= \frac{1}{0.04} \\
n &= \frac{1}{0.04 \cdot 10^{-8}} \\
n &= \frac{1}{4 \cdot 10^{-10}} \\
n &= 2.5 \cdot 10^{9}
\end{aligned}$$

Portanto, concluímos que o número de amostras necessário para estimar π com uma precisão de 10^{-4} e uma confiança de 0.99 é $2.5 \cdot 10^9$ amostras.

2.4 Integração Numérica

Considere o problema de calcular a integral definida de uma função:

$$\int_{x=a}^{x=b} h(x)\, \mathrm{d}x$$

Podemos encontrar um **problema** quando a função não é integrável analiticamente, como ilustrado no exemplo da Figura 2.

Podemos abordar esse problema usando o método de Monte Carlo. A classe que lida com esse tipo de problema é conhecida como **Integração de Monte Carlo**.

2.4.1 Integração de Monte Carlo

A integração de Monte Carlo é uma generalização do método usado para estimar o valor de π (veja Subseção 2.3). O método consiste em estimar a razão entre as áreas abaixo da curva (integral) e a área de um quadrado usando amostras uniformemente distribuídas.

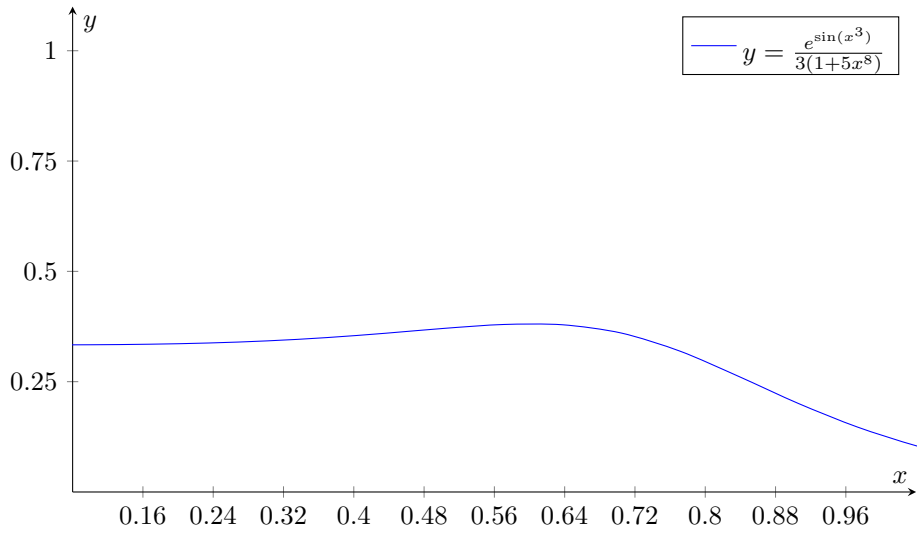


Figura 2: Exemplo de função não integrável analiticamente

Suponha que $0 \leq h(x) \leq 1$ para todo $x \in [0..1]$, onde $h(x)$ é definido como:

$$I = \int_{x=0}^{x=1} h(x)\,\mathrm{d}x$$

Podemos definir uma função indicadora para pontos abaixo da curva:

$$g(x,y) = \begin{cases} 1 & \text{se } h(x) \leq y, \\ 0 & \text{caso contrário} \end{cases}$$

Sejam X, Y variáveis aleatórias contínuas distribuídas uniformemente em $[0..1]$. Podemos definir o valor esperado:

$$\mathbb{E}[g(X,Y)] = \int_{y=0}^{y=1} \int_{x=0}^{y=0} f_{XY}(x,y) g(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

onde $f_{XY}(x,y) = f_X(x) f_Y(y) = 1$, pois é a densidade conjunta de duas variáveis aleatórias contínuas independentes uniformemente distribuídas em $[0..1]$.

Então, $\mathbb{E}[g(X,Y)]$ é a fração entre a área abaixo da curva e a área do quadrado $[0..1]$ (ou seja, $\int_{x=0}^{x=1} h(x)$).

Sejam X_i, Y_i sequências i.i.d uniformes em $[0..1]$. Temos que:

$$\bar{M}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i)$$

e, pela Lei dos Grandes Números, temos que:

$$\lim_{n \to \infty} \bar{M}_n = h(x)$$

3 Introdução a Cadeias de Markov

Nesta seção, exploraremos conceitos fundamentais relacionados às cadeias de Markov, uma ferramenta robusta na modelagem matemática de sistemas dinâmicos que incorporam elementos de aleatoriedade. Abordaremos temas cruciais, como as próprias cadeias de Markov, o modelo on-off, a distribuição ao longo do tempo, a irredutibilidade e a aperiodicidade. Para uma compreensão mais aprofundada desses conceitos, recomendamos a leitura de [5] como uma referência sólida.

3.1 Visão Geral

As cadeias de Markov constituem uma representação matemática dinâmica de sistemas que envolvem elementos aleatórios. Essa teoria, desenvolvida por Andrey Markov, tem suas raízes em um trabalho seminal intitulado "Sobre a distribuição das grandes quantidades dependentes umas das outras" publicado em 1906 [6]. A abordagem inovadora de Markov proporcionou uma base fundamental para a modelagem de processos estocásticos, onde a evolução futura do sistema depende apenas do estado presente, marcando assim uma contribuição significativa para a teoria das probabilidades e a compreensão de fenômenos aleatórios.

Uma cadeia de Markov pode ser definida em termos de:

- **Espaço de estados**: conjunto de todos os possíveis estados que o sistema pode assumir, representando os valores possíveis das variáveis aleatórias (podendo ser finito ou infinito);
- **Matriz de transição (ou matriz estocástica)**: descreve as possíveis transições de estado que o sistema pode realizar;
- **Tempo discreto**: implica uma transição a cada intervalo de tempo, embora a representação em tempo contínuo também seja possível;
- **Estado inicial**: indica o estado no qual o sistema inicia sua evolução.

Uma representação visual comum dessas cadeias é usando um grafo direcionado ponderado, onde os vértices representam os estados do sistema, as arestas representam as possíveis transições, e os pesos associados indicam as probabilidades das transições (a soma dos pesos de saída de um estado deve ser igual a 1).

Definição 3.1 (**Cadeia de Markov**). Uma Cadeia de Markov CM é definida por:

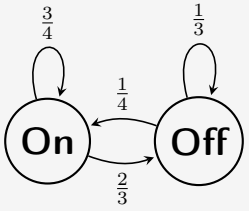
$$
\begin{aligned}
S &:= \text{ espaço de estados} \\
P &:= \text{ matriz quadrática de transição de estados com dimensão } |S| \\
p_{i,j} &:= P[i,j] \\
X_t &:= \text{ v.a. que determina o estado do sistema instante de tempo } t \text{ para todo } t = 0, 1, 2 \ldots
\end{aligned}
$$

onde cada t, X_t possui uma distribuição diferente e **obrigatoriamente** temos que

$$\sum_{j=1}^{|S|} p_{i,j} = 1$$

para todo $i \in [1..|S|]$.

Exemplo 3.1 (Modelo On-Off)**.** O modelo on-off é a CM (cadeia de Markov) mais simples, vamos exemplificá-la aqui. Veja a representação da CM



a sua matriz de transição de probabilidade é

$$P = \begin{pmatrix} p_{1,1} = 3/4 & p_{1,2} = 2/3 \\ p_{2,1} = 1/4 & p_{2,2} = 1/3 \end{pmatrix}$$

Podemos dizer que o sistema começa no estado On e então calcular as probabilidades:

$$\Pr[X_0 = 1] = 1, \Pr[X_1 = 1] = \frac{3}{4}$$
$$\Pr[X_1 = 2] = \frac{1}{4}, \Pr[X_2 = 1] = \frac{35}{48}$$
$$\Pr[X_2 = 2] = \frac{13}{48}, \ldots$$

e assim seguir infinitamente.

3.2 Propridade Sem Memória

As cadeias de Markov não tem memória[3], isso refere-se à independência dos eventos ou, mais especificamente, à independência de v.a.s no tempo t entre eventos: o tempo até o próximo evento $(t + 1)$ não pode ser previsto a partir do tempo até o evento anterior. Isso quer dizer que o próximo estado só depende do estado atual, e não de como chegamos ao estado atual.

Considere uma trajetória T de estados sobre o espaço de estado S, com uma matriz de transição de estado P definida abaixo:

$$T = (X_0 = s_0, X_1 = s_1, \ldots, X_{t-1} = s_{t-1}, X_t = s_t)$$

[3]memoryless property

onde $s_i \in S$. Temos que

$$\begin{aligned}\Pr[X_{t+1} = s|T] &= \Pr[X_{t+1} = s|X_t = s_t] \\ &= \Pr[st, s]\end{aligned}$$

ou seja, podemos concluir que **toda evolução na CM é em função da matriz** P.

3.3 Distribuição no Tempo t

Vamos representar o estado da CM através do vetor de distribuição de probabilidade, veja Definição 3.2

Definição 3.2. Dado uma CM qualquer definimos o **vetor de distribuição de** X_t como:

$$\pi(t)$$

denotamos

$$\Pr[X_t = s] = \pi_s(t) := \text{ probabilidade do sistema estar no estado } s \text{ no tempo } t$$

No modelo On-Off em Exemplo 3.1 temos que $\pi(t)$ é

$$\begin{aligned}\pi_1(t) &= \frac{3}{4}\pi_1(t-1) + \frac{2}{3}\pi_2(t-1) \\ \pi_2(t) &= \frac{1}{4}\pi_1(t-1) + \frac{1}{3}\pi_2(t-1)\end{aligned}$$

intuitivamente $\pi_i(t+1)$ é a probabilidade de a CM estar no estado i no tempo $t+1$, isto é, a soma das probabilidades de transição de j para i dado o estado no tempo t.

Por exemplo vamos calcular $\pi(1)$ dado que $\pi(0) = (1, 0)$, veja o resultado abaixo:

$$\begin{aligned}\pi_1(t) &= \frac{3}{4}1 + \frac{2}{3}0 = \frac{3}{4} \\ \pi_2(t) &= \frac{1}{4}1 + \frac{1}{3}0 = \frac{1}{4}\end{aligned}$$

3.3.1 Generalizando

Vamos generalizar a distribuição no tempo t, veja em

Teorema 3.1 (Distribuição no tempo t)**.** Dado uma CM qualquer, um espaço de estados S, um tempo t, temos que

$$x_i(t) = \Pr[X_t = i] = \sum_{j=1}^{|S|} p_{i,j}\pi_j(t-1)$$

Demonstração. Vamos demonstrar que $x_i(t) = \Pr[X_t = i] = \sum_{j \in S} P_{i,j}\pi_j(t-1)$, para isto basta usarmos a **lei de probabilidade total** (veja em Teorema A.1), observe a manipulação abaixo

$$\begin{aligned} x_i(t) = \Pr[X_t = i] &= \Pr[X_t = i | X_{t-1} = j] \Pr[X_{t-1} = j] \\ &= \sum_{j \in S} p_{i,j}\pi_j(t-1) \end{aligned}$$

□

Podemos escrever de forma matricial (π é um vetor linha), veja abaixo

$$\pi(t) = \pi(t-1)P = \pi(t-2)PP = \ldots = \pi(0)P^t$$

A **vantagem** é que para encontrar a distribuição $\pi(t)$ basta fazer multiplicações pela matriz P, mas a má notícia é que multiplicar matriz tem complexidade $O(n^3)$[4]

3.4 Irredutibilidade e Periodicidade

Nesta subseção vamos definir os termos de **irredutibilidade** e **periodicidade** de uma CM. Mas antes precisamos definir a comunicação entre estados. Vamos às definições

Definição 3.3. Dado uma CM, considere dois estados s_i e s_j, dizemos que s_i se comunica com s_j se e somente se s_j existe algum $t > 0$ tal que $\Pr[X_{t_0+t} = s_j | X_{t_0} = s_i]$, denotamos

$$s_i \to s_j$$

Note que esta propriedade independe de t_0. Neste caso a probabilidade depende apenas de $(P^t)[i,j]$[a]. Basta haver caminho de s_i para s_j no grafo direcionado da CM.

Se $s_i \to s_j$ e $s_j \to s_j$ dizemos que s_i e s_j se intercomunicam, denotamos:

$$s_i \leftrightarrow s_j$$

[a] Valor nos índices i, j da matriz P multiplicada t vezes

Definição 3.4 (Irredutibilidade de CM). Uma CM é dita **irredutível** se para par de estados (s_i, s_j) temos que $s_i \leftrightarrow s_j$, caso contrário dizemos que ela é redutível.

Olhando para o grafo direcionado induzido pela matriz de transição P dizemos que a cadeia é irredutível se há caminho entre qualquer par de vértices, ou seja, um **grafo fortemente conexo**

[4] Existem algoritmos de múltiplicação de matrizes com complexidades melhores que cúbica, mas nem sempre são usados na prática.

Observação. No modelo On-Off em Exemplo 3.1 a CM é irredutível!

Exemplo 3.2. Vamos dar um exemplo de uma CM **redutível**, veja matriz de transição da cadeia abaixo:

$$P = \begin{pmatrix} 0 & 0.2 & 0.8 & 0 \\ 0.6 & 0 & 0 & 0.4 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

note que não existe nenhum valor de t tal que $(P^t)[3,2] > 0$ e $(P^t)[3,1] > 0$.

Veja que intuitivamente uma CM redutível significa que podemos "reduzir" a cadeia a um certo vértice. Neste exemplo a partir do estado 3 não conseguimos alcançar o vértice 2 e 1.

Se uma CM é irredutível, isso implica que, ao longo do tempo, é possível alcançar qualquer estado a partir de qualquer outro estado, tornando a cadeia globalmente conectada. Essa propriedade é essencial para garantir que a cadeia alcance o equilíbrio, ou seja, atinja uma distribuição estacionária (veremos em Definição 4.1) em que as probabilidades de estar em qualquer estado se estabilizem.

Definição 3.5. Seja s_i um estado de uma CM qualquer, e A_i o conjunto dos **comprimentos de caminho** que iniciam e terminam em s_i, ou seja

$$A_i = \{t : (P^t)[i,j] > 0\}$$

definimos que o período de s_i é um máximo divisor comum de A_i, denotamos

$$d(s_i) = \text{mdc}(A_i)$$

Definição 3.6. Seja s_i um estado de um CM qualquer, dizemos que o estado s_i é aperiódico se e somente se $d(s_i) = 1$

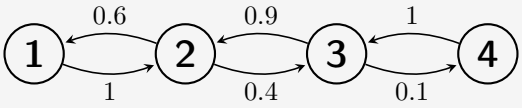
Veja que a aperiodicidade de um estado se relaciona à ideia de "caminhos de volta", pois intuitivamente a aperiodicidade implica que você pode encontrar caminhos de retorno a s_i de comprimentos diferentes. Se s_i não fosse aperiódico, todos os caminhos de retorno teriam comprimentos múltiplos inteiros de um mesmo número, e o estado seria periódico.

Definição 3.7 (Cadeia de Markov Aperiódica). Dizemos que uma CM é aperiódica se todos os seus estados são aperiódicos.

Em termos práticos, a aperiodicidade é desejável porque garante que a cadeia de Markov alcance um equilíbrio mais rapidamente e não exiba padrões periódicos previsíveis. A cadeia de Markov aperiódica tende a ter uma convergência mais rápida para sua distribuição estacionária.

Observação. No modelo On-Off em Exemplo 3.1 a CM é aperiódica!

Exemplo 3.3. Seja uma CM representada pelo grafo abaixo:



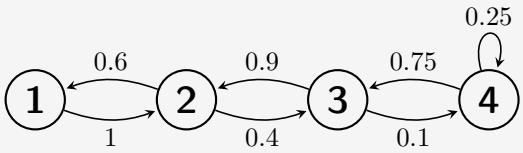
veja que a CM é periódica com período 2 pois não há como sair do estado 1 e voltar para ele em tempo ímpar.

Teorema 3.2. Seja um CM de irredutível tal que existe um estado $s_i \in S$ tal que $p_{i,i} > 0$, ou seja s_i tem uma aresta em laço, então temos que a cadeia é **aperiódica**.

Demonstração. Basta ver que se a CM é irredutível e possui um estado com laço então podemos gerar qualquer caminho de retorno grande o suficiente para qualquer estado. □

Lema 3.1. Em uma CM irredutível, todos os estados são aperiódicos ou todos são periódicos com o mesmo período

Exemplo 3.4. Se adicionarmos um laço a um estado da CM exemplificada em Exemplo 3.3, como abaixo:



tornamos a cadeia aperiódica!

4 Explorando Conceitos Fundamentais em Cadeias de Markov

Nesta seção vamos explorar conceitos importantíssimos em Cadeias de Markov: distribuição estacionária, tempo de chegada, distância de variação total, convergência e a reversibilidade. A distribuição estacionária destaca-se pela estabilidade probabilística ao longo do tempo, enquanto o tempo de chegada é crucial para compreender intervalos entre eventos sucessivos. A distância de variação total quantifica a dispersão em dados, e a convergência é essencial para a estabilidade de modelos estatísticos. A reversibilidade em processos estocásticos sugere uma simetria temporal única.

4.1 Convergência - Distribuição Estacionária

Vamos olhar agora para a convergência da CM. Sabemos que para um certo estado inicial temos que

$$\pi(t) = \pi(t-1)P = \pi(0)P^t,$$

mas para onde vai $\lim_{t\to\infty} \pi(t)$?

É possível que uma variável aleatória X_t não convirja em uma cadeia de Markov, continuando a "passear" indefinidamente entre diferentes estados. No entanto, é igualmente possível que a probabilidade de X_t atingir um estado específico s_i convirja, ou seja, a fração de vezes que X_t assume o valor s_i pode convergir para um valor constante ao longo do tempo.

Portanto estamos interessados na distribuição π tal que

$$\lim_{t\to\infty} \pi(t) = \pi(0)P^t$$

Definição 4.1 (Distribuição Estacionária). Dada uma cadeia de Markov (CM) com matriz de transição P e espaço de estados S, dizemos que π é uma distribuição estacionária se e somente se satisfizer as seguintes condições:

$$\begin{aligned} \pi_s &\geq 0 \quad \text{para todo } s \in S, \\ \sum_{s\in S} \pi_s &= 1, \\ \pi P &= \pi, \\ \pi_i &= \sum_{j\in S} \pi_j p_{j,i} \quad \text{para todo } i \in S. \end{aligned}$$

Ou seja, ao multiplicar π por P temos p de volta. A CM "estacionou" (convergimos), não temos mais evolução.

Exemplo 4.1. Considere a CM em Exemplo 3.1, neste exemplo temos que a distribuição estacioná-

ria é

$$\pi = \begin{pmatrix} 8/11 & 3/11 \end{pmatrix}$$

observe o resultado de πP abaixo

$$\pi_1 = \frac{8}{11}\frac{3}{4} + \frac{3}{11}\frac{2}{3} = \frac{8}{11}$$
$$\pi_2 = \frac{8}{11}\frac{1}{4} + \frac{3}{11}\frac{1}{3} = \frac{3}{11}$$

Definição 4.2 (**Tempo de Chegada**)**.** Dada uma CM e dois estados s_i e s_j, definimos o **tempo de chegad**a (em número de transições) necessário para sair do estado s_i e chegar ao estado s_j como

$$T_{i,j} = \min\{t \mid X_t = s_j \wedge X_0 = s_i\}.$$

Observa-se que o número o tempo de chegada $T_{i,j}$ é **aleatório**. A esperança é denotada como $\tau_{i,j} = \mathbb{E}[T_{i,j}]$.

Quando $i = j$, temos $\tau_{i,i}$, chamado de **tempo médio de retorno** ao estado s_i.

Observe que condicionamos estarmos em s_i, o tempo de chegada $T_{i,j}$ é determinado pelas probabilidades de transição entre os estados da cadeia de Markov, que são características fixas da cadeia e não dependem de $\pi(t)$.

Teorema 4.1. Para qualquer CM irredutível e aperiódica, para quaisquer dois estados si e s_j, temos o seguinte:

$$\Pr[T_{i,j} < \infty] = 1 \text{ e } \mathbb{E}[T_{i,j}] = \tau_{i,j} < \infty$$

isto reflete que a probabilidade de $T_{i,j}$ ser infinito é 0 e que o valor esperado do tempo médio de retorno a s_i **é finito**.

Demonstração. Seja um CM irredutível e aperiódica, e dois estados s_i, s_j. Se ela é irredutível estão existe um caminho de comprimento "finito" de s_i e s_j, logo $\Pr[T_{i,j} < \infty] = 1$; e como ela é aperiódica então existem diversos caminhos de s_i a s_j maiores que um determinado valor, ou seja $\tau_{i,j} < \infty$ □

Teorema 4.2. Para qualquer CM irredutível e aperiódica, para qualquer estado si, temos a seguinte relação:

$$\pi_i = \frac{1}{\tau_{i,i}}$$

A prova do Teorema 4.2 é um pouco complexa, não veremos aqui, mas a intuição é que na média visitamos s_i a cada $\tau_{i,i}$ passos.

Exemplo 4.2. O tempo médio de retorno do estados on e pff visto no Exemplo 3.1 é

$$\tau_{1,1} = \frac{1}{\pi_1} = \frac{11}{8} = 1.375$$
$$\tau_{2,2} = \frac{1}{\pi_2} = \frac{11}{3} = 3.666$$

4.2 Distância de Variação Total

Agora que já sabemos a re11lação entre o tempo de retorno e a distribuição estacionária queremos saber o quão longe um vetor de distribuição $\pi(0)$ está da distribuição estacionária.

Definição 4.3 (Distância de Variação Total). Sejam α e β dois vetores de que representam uma distribuição de probabilidade (discreta) em S, a distância de variação total entre eles é dada por

$$d_{\mathrm{TV}}(\alpha, \beta) = \frac{1}{2} \sum_k |\alpha_k - \beta_k|$$

onde k é o número de elementos em S.[a].

[a] Normalizamos por $\frac{1}{2}$ para mantermos a distância entre $[0..1]$

Exemplo 4.3. Seja α e β dois vetores de que representam uma distribuição de probabilidade (discreta) de dimensão 5, descritos abaixo

$$\alpha = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.3 & 0.1 \end{pmatrix}$$
$$\beta = \begin{pmatrix} 0.2 & 0.1 & 0.1 & 0.5 & 0.1 \end{pmatrix}$$

temos que

$$\begin{aligned} d_{\mathrm{TV}}(\alpha, \beta) &= \frac{1}{2} \sum_k |\alpha_k - \beta_k| \\ &= \frac{1}{2}(0.1 + 0.1 + 0.2 + 0.2 + 0) \\ &= 0.3 \end{aligned}$$

Teorema 4.3. Para qualquer CM irredutível e aperiódica, e qualquer distribuição inicial $\pi(0)$, temos que

$$\lim_{t \to \infty} d_{\mathrm{TV}}(\pi(t), \pi) = 0$$

ou seja sempre converge para a distribuição estacionária[a], **independente da condição inicial**!

[a] Além disso π é única, veremos em Teorema 5.2

4.3 Encontrando a distribuição Estacionária

Em síntese, a distribuição estacionária em cadeias de Markov é uma peça-chave para compreender a estabilidade e o comportamento de sistemas dinâmicos ao longo do tempo. Sua aplicação abrange desde a previsibilidade a longo prazo até a modelagem eficaz de processos estocásticos. Vamos enumerar três maneiras de se encontrar a distribuição estacionária:

1. Método iterativo: fazemos a iteração até algum critério de convergência desejado:

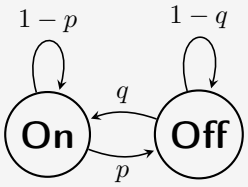
$$\pi(t) = \pi(t-1)P = \pi(0)P^t$$

2. Sistema de equações: resolvemos o sistema de equações abaixo

$$\pi = \pi P \text{ e } \sum_{i \in S} \pi_i = 1$$

3. **Monte Carlo**: usamos a própria cadeia para gerar amostras para estimar π_i ou estimar $\tau_{i,i}$ para todo s_i

Exemplo 4.4. Dado o modelo de CM On-Off genérico representado abaixo



vamos resolver o sistema de equações:

$$\begin{cases} \pi_1 & = (1-p)\pi_1 + q\pi_2 \\ \pi_2 & = p\pi_1 + (1-q)\pi_2 \\ \pi_1 + \pi_2 & = 1 \end{cases}$$

substituindo π_2 por $1 - \pi_1$ na primeira equação obtemos:

$$\begin{aligned}
\pi_1 &= (1-p)\pi_1 + q(1-\pi_1) \\
1 &= (1-p) + \frac{q(1-\pi_1)}{\pi_1} \\
0 &= \frac{q - p\pi_1 - q\pi_1}{\pi_1} \\
0 &= \frac{q}{\pi_1} - p - q \\
\pi_1 &= \frac{q}{p+q}
\end{aligned}$$

logo

$$\pi_2 = 1 - \frac{q}{p+q}$$

4.4 Reversibilidade

A reversibilidade em cadeias de Markov refere-se à propriedade em que as transições entre estados são igualmente prováveis, independentemente da direção do tempo. Essa propriedade é crucial porque implica a existência de uma distribuição estacionária.

Definição 4.4 (Cadeia de Markov Reversível). Uma CM é dita reversível para uma distribuição π se e somente se

$$\pi_i p_{i,j} = \pi_j p_{j,i}$$

Intuitivamente se uma CM é reversível então para cada par de estados i, j, a taxa de transição a longo prazo da cadeia de um estado i para um estado j é igual à taxa de transição a longo prazo da cadeia de um estado j para um estado i.

4.4.1 Passeios Aleatórios

Seja $G = (V, E)$ um grafo não direcionado, considere um andarilho que passeia pelo grafo de forma aleatória, sem preferência e nem memória. Ele apenas escolhe uniformemente o próximo vértices entre os vizinhos do vértice atual. Veja uma representação em Figura 3. Nessa representação as probabilidades são:

$$\begin{aligned}
p_{1,2} &= p_{1,3} = \frac{1}{2} \\
p_{2,1} &= p_{2,3} = p_{2,4} = \frac{1}{3} \\
&\dots \\
p_{8,4} &= p_{8,7} = \frac{1}{2}
\end{aligned}$$

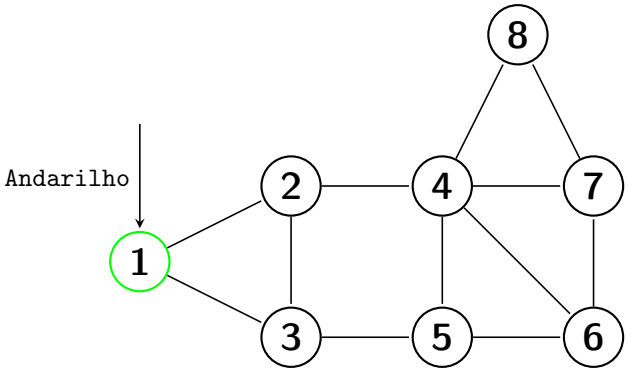


Figura 3: Exemplo de Passeio Aleatório

Veja que o andarilho induz uma CM sobre o grafo G. Definimos

$$X_t := \text{ vértice onde o andarilho se encontra no tempo } t$$
$$p_{i,j} = \frac{1}{\deg(i)}, \text{ onde } \deg(i) \text{ é o grau do vértice } i$$

É possível verificar que a distribuição estacionária do andarilho é

$$\pi_i = \frac{\deg(i)}{K}$$

onde K (fator de normalização) é

$$K = \sum_i \deg(i) = 2m, \text{ onde } m \text{ é o número de arestas.}$$

note que a CM é reversível, pois

$$\pi_i p_{i,j} = \pi_j p_{j,i}$$

tanto faz o andarilho andar para "frente" ou para "atrás". No exemplo da Figura 3 temos a seguinte distribuição estacionária:

$$\pi = \left(\tfrac{2}{24} \quad \tfrac{3}{24} \quad \tfrac{3}{24} \quad \tfrac{5}{24} \quad \tfrac{3}{24} \quad \tfrac{3}{24} \quad \tfrac{3}{24} \quad \tfrac{2}{24}\right)$$

4.5 Processo de Nascimento e Morte (Birth-Death)

O processo de nascimento e morte é um modelo estocástico utilizado para descrever a evolução temporal de sistemas nos quais entidades individuais podem surgir (nascimento) e desaparecer (morte) ao longo do tempo. Ela á uma generalização do modelo On-Off - funciona como uma fila.

Seja uma CM com estados $S = \{1, \ldots, k\}$, as suas transições são apenas entre estados vizinhos e possíveis laços, ou seja a matriz de transição P é tridiagonal[5], veja sua representação na Figura 4.

[5]Geralmente $p_{i,i+1}$ e $p_{i+1,i}$ possuem uma forma de serem definidas.

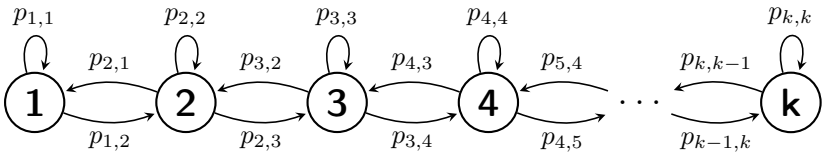


Figura 4: Modelo Estocástico Processo de Nascimento e Morte

Vamos assumir que a cadeia é reversível, logo sua distribuição estacionária respeita:

$$\pi_i p_{i,j} = \pi_j p_{j,i}$$

Para chegarmos na distribuição estacionária descobrirmos a distribuição estacionária π, vamos inicialmente assumir um valor para π_1, temos então que

$$\pi_2 = \frac{\pi_1 p_{1,2}}{p_{2,1}} = \pi_3 = \frac{\pi_2 p_{2,3}}{p_{3,2}} = \frac{\pi_1 p_{1,2} p_{2,3}}{p_{2,1} p_{3,2}} = \ldots = \pi_i = \frac{\pi_1 \prod_{j=1}^{i-1} p_{j,j+1}}{\prod_{j=1}^{i-1} p_{j+1,j}}$$

Agora vamos calcular π_1

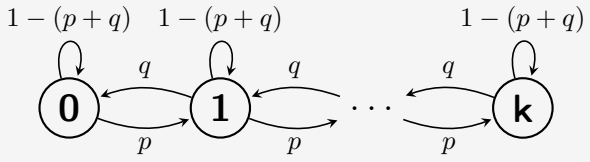
$$\begin{aligned} 1 = \sum_{i=1}^{k} \pi_i &= \sum_{i=1}^{k} \left(\frac{\pi_1 \prod_{j=1}^{i-1} p_{j,j+1}}{\prod_{j=1}^{i-1} p_{j+1,j}} \right) \\ &= \pi_1 \sum_{i=1}^{k} \left(\frac{\prod_{j=1}^{i-1} p_{j,j+1}}{\prod_{j=1}^{i-1} p_{j+1,j}} \right) \\ \pi_1 &= \left(\sum_{i=1}^{k} \left(\frac{\prod_{j=1}^{i-1} p_{j,j+1}}{\prod_{j=1}^{i-1} p_{j+1,j}} \right) \right)^{-1} \end{aligned}$$

Portando concluímos que BD (Birth Death) é reversível e π é a sua distribuição estacionária.

Exemplo 4.5. Considere uma fila que segue um processo de nascimento e morte, onde a cada ponto no tempo, um elemento pode chegar, sair ou permanecer no mesmo estado. A probabilidade de um elemento "chegar" é denotada por p, enquanto a probabilidade de "sair" é representada por q, e então $(p + q) < 1$. A cadeia de Markov que representa a fila possui $k + 1$ estados, de 0 a k, e cada estado i representa o fato da fila possuir i elementos. Assim, as transições de estado na fila podem ser descritas pelas seguintes equações:

$$\begin{aligned} p_{i,i+1} &= p, \quad \forall i \in [0..k-1], \\ p_{i+1,i} &= q, \quad \forall i \in [1..k], \\ p_{i,i} &= 1 - (p+q), \quad \forall i \in [0..k]. \end{aligned}$$

Veja a representação da CM a seguir



Temos que seu estado estacionário é

$$\begin{aligned}\pi_i &= \pi_0 \frac{\prod_{j=0}^{i-1} p_{j,j+1}}{\prod_{j=0}^{i-1} p_{j+1,q}} \\ &= \pi_0 \left(\frac{p}{q}\right)^i\end{aligned}$$

e pela fórmula de progressão geométrica finita

$$\pi_0 = \left(\sum_{i=0}^{k} \left(\frac{p}{q}\right)^i\right) = \frac{1 - \frac{p}{q}}{1 - \left(\frac{p}{q}\right)^{k+1}}, \text{ para } p < q.$$

Seja $k = 10$, $p = 0.3$ e $q = 0.4$, aplicando as fórmulas temos:

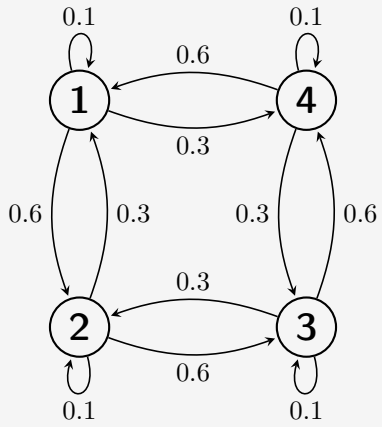
$$\pi_0 = \frac{1 - \frac{0.3}{0.4}}{1 - \left(\frac{0.3}{0.4}\right)^9} = 0.261 \text{ e } \pi_i = \pi_0 \left(\frac{0.3}{0.4}\right)^i$$

e a partir disso obtemos $\pi_0 = 0.261$ que é probabilidade da fila estar vazia. e $\pi_k = 0.015$ que é a probabilidade da fila estar cheia. O que faz sentido pois a probabilidade de sair um elemento é maior que a probabilidade de chegar um elemento.

Seja X_t o estado da fila no instante t, temos que o valor esperado do **tamanho da fila** é

$$\begin{aligned}\mathbb{E}[X_t] &= \sum_{i=0}^{10} i\pi_i \\ &= \sum_{i=0}^{10} i\pi_0 \left(\frac{0.3}{0.4}\right)^i \\ &= 0.261 \cdot 9.63 \\ &= 2.51343\end{aligned}$$

Exemplo 4.6. Seja a CM representado pelo grafo abaixo



a sua distribuição estacionária é $\pi_i = \frac{1}{4}$, mas não é reversível, veja que

$$\pi_1 p_{1,2} = \frac{1}{4}\frac{6}{10} = \frac{3}{20} \neq \pi_2 p_{2,1} = \frac{1}{4} = \frac{3}{10} = \frac{3}{40}$$

A visão intuitiva é que o processo **"gira mais" em uma direção** (anti-horário).

5 Convergência para Estacionaridade e Tempo de Mistura em Cadeias de Markov

Nesta seção, abordaremos conceitos fundamentais relacionados a processos estocásticos e teoria de Markov. Discutiremos autovalores, autovetores e decomposição associada, bem como convergência para estacionaridade. Avaliaremos o tempo de mistura, o Lacuna Espectral e exploraremos o tempo de mistura de passeios aleatórios. Estes tópicos são essenciais para entender as propriedades dos processos estocásticos em estudo.

5.1 Autovalores e Autovetores

Definição 5.1 (Autovalores e Autovetores). Seja P uma matriz quadrada. Um vetor não nulo v é considerado um autovetor de P associado ao autovalor λ se a multiplicação da matriz P pelo vetor v resultar na mesma direção de v, escalada pelo valor λ. Matematicamente, isso é expresso pela equação:

$$Pv = \lambda v$$

Em outras palavras, a aplicação da matriz P sobre o vetor v apenas amplia ou reduz o vetor sem alterar sua direção, sendo a escala dada pelo autovalor λ.

Definição 5.2. Dado uma matriz P e um vetor não nulo u, se $uP = \lambda u$ então chamamos u de autovetor a esquerda de P.

Teorema 5.1. Dado uma matriz P e um autovetor a esquerda u, então existe um autovetor v tal que

$$P^{\mathrm{T}}v = \lambda u$$

5.1.1 Autovetor Estacionário

Dadas as definições, podemos prontamente observar que, para uma distribuição estacionária π de uma Cadeia de Markov com matriz de transição P, temos $\pi P = \pi$, indicando que π é o **autovetor à esquerda de P associado ao autovalor** 1. Contudo, é necessário assegurar ainda que

$$\sum_{s \in S} \pi_s = 1$$

A abordagem para resolver isso é normalizar o autovetor, garantindo que ele seja um vetor de probabilidade, ou seja, que a soma dos elementos seja igual a 1.

Teorema 5.2. Se P é uma matriz estocástica, verifica-se que para cada autovalor λ_i temos que $|\lambda_i| \leq 1$. Além disso, a matriz estocástica P possui um único autovetor associado ao autovalor $\lambda = 1$.

O Teorema 5.2 nos revela que existe apenas uma distribuição estacionária, e ainda que os outros autovalores da matriz P convergem a 0, veremos em mais detalhes futuramente que isso nos garante a convergência para distribuição estacionária.

5.1.2 Decomposição em Autovetores

Uma matriz quadrada P pode ser escrita através de seus autovetores e autovalores da seguinte forma:

$$P = QLQ^{-1}$$

onde Q é a matriz com autovetores de P como colunas e L é matriz diagonal onde $L_{i,i}$ é o autovalor associado ao autovetor i (i-ésima coluna de Q).

Exemplo 5.1. Dado uma matriz estocástica P, onde

$$P = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.5 & 0.1 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}$$

e os autovalores $\lambda_1 = 1$, $\lambda_2 = 0.3$, $\lambda_3 = -0.4$ respectivamente associados aos autovetores:

$$\begin{aligned} v_1 &= \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ v_2 &= \begin{pmatrix} 2 & -5 & 2 \end{pmatrix} \\ v_3 &= \begin{pmatrix} -43 & 13 & 55 \end{pmatrix} \end{aligned}$$

temos que $P = QLQ^{-1}$ onde

$$Q = \begin{pmatrix} 1 & 2 & -43 \\ 1 & -5 & 13 \\ 1 & 2 & 55 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & -0.4 \end{pmatrix}, \quad Q^{-1} = \begin{pmatrix} 43/98 & 2/7 & 27/98 \\ 3/49 & -1/7 & 4/49 \\ -1/98 & 0 & 1/98 \end{pmatrix}$$

5.2 Convergência

Sabemos que uma Cadeia de Markov aperiódica e irredutível **sempre converge** para a sua distribuição estacionária, alcançando o equilíbrio de forma garantida (veja Teorema 4.2). Além disso, a **distribuição estacionária é única** (veja Teorema 5.2). Contudo, a questão que se coloca é: quão rápida é essa convergência? Como $d_{\mathrm{TV}}(\pi(t), \pi)$ converge para 0?

É possível que a taxa de convergência $d_{\mathrm{TV}}(\pi(t), \pi)$ siga um comportamento específico, tal como:

$$d_{\mathrm{TV}}(\pi(t), \pi) = \begin{cases} \Theta(e^{-at}) & \text{se for rápido} \\ \Theta(t^{-b}) & \text{se for lento (lei da potência)} \\ \Theta\left((\log t)^{-c}\right) & \text{se for muito lento} \end{cases}$$

Além disso, a velocidade de convergência pode depender dos valores iniciais π_0 e/ou da matriz de transição P. Esses questionamentos serão explorados ao longo dessa seção.

5.2.1 Distribuição no Tempo t

Vamos agora considerar a distribuição no tempo t e decompor a matriz de transição em termos de seus autovetores e autovalores.

Seja $\pi(0)$ a distribuição inicial da cadeia de Markov CM. A distribuição no tempo t é expressa por

$$\pi(t) = \pi(t-1)P = \pi(0)P^t$$

Ao decompor a matriz estocástica, obtemos

$$\begin{aligned} P^t &= PP \dots P \\ &= (QLQ^{-1})(QLQ^{-1})(QLQ^{-1}) \dots (QLQ^{-1}) \\ &= QLILI \dots LQ^{-1} \\ &= QLL \dots LQ^{-1} \\ &= QL^tQ^{-1} \end{aligned}$$

Portanto, podemos afirmar que

$$\pi(t) = \pi(0)QL^tQ^{-1}$$

Neste ponto, surge a questão: para onde converge L^t conforme t aumenta? Ou seja, qual é o valor para

$$\lim_{t \to \infty} L^t$$

Concluímos que, à medida que $t \to \infty$, o autovalor $\lambda = 1$ permanece inalterado, enquanto os outros autovalores λ_i convergem para zero, pois $\lambda_i < 1$.

Considerando o modelo On-Off visto em Exemplo 3.1, com a matriz estocástica

$$P = \begin{pmatrix} 0.75 & 0.25 \\ 0.67 & 0.33 \end{pmatrix}$$

e seus autovalores $\lambda_1 = 1$ e $\lambda_2 = \frac{1}{12}$, associados aos autovetores

$$v_1 = \begin{pmatrix} 8/3 & 1 \end{pmatrix}$$
$$v_2 = \begin{pmatrix} 1 & -1 \end{pmatrix}$$

normalizaremos então v_1 para obtermos nossa distribuição estacionária

$$v_1 = \begin{pmatrix} \frac{8/3}{11/3} & \frac{1}{11/3} \end{pmatrix} = \begin{pmatrix} 8/11 & 3/11 \end{pmatrix}$$

Suponhamos agora

$$\pi = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

podemos então escrever $\pi(0)$ em termos dos autovetores de P, ou seja

$$\pi = \begin{pmatrix} 1 & 0 \end{pmatrix} = \pi + \frac{3}{11} v_2$$

agora vamos escrever a distribuição no tempo t

$$\begin{aligned} \pi(t) &= \pi(0)P^t \\ &= (\pi + \frac{3}{11} v_2)P^t \\ &= \pi P^t + \frac{3}{11} v_2 P^t \\ &= \pi + \frac{3}{11} \pi_2^t v_2 \\ &= \begin{pmatrix} 8/11 + 3/11\,(1/12)^t & 3/11 - 3/11\,(1/12)^t \end{pmatrix} \end{aligned}$$

Agora, medindo a distância de variação total

$$\begin{aligned} d_{\mathrm{TV}}(\pi(t), \pi) &= \frac{1}{2}\frac{3}{11}\left(\left(\frac{1}{12}\right)^t + \left(\frac{1}{12}\right)^t\right) \\ &= \left(\frac{3}{11}\right)\left(\frac{1}{12}\right)^t \\ &= \theta(12^{-t}) \end{aligned}$$

Ou seja, **converge exponencialmente rápido!** As constantes dependem de P e $\pi(0)$. Este resultado é válido para **qualquer Cadeia de Markov**, conforme o Teorema 5.3 apresenta.

Teorema 5.3 (**Teorema da Convergência**). Considere uma Cadeia de Markov aperiódica e irredutível representada por sua matriz estocástica P e com uma distribuição estacionária π.

Existem constantes $\alpha \in [0, 1]$ e $C > 0$ tais que

$$\max \left\{ \pi(0) \mid d_{\mathrm{TV}}(\pi(t), \pi) \leq C\alpha^t \right\}$$

Veja que o teorema implica que a distribuição transiente $\pi(t)$ converge exponencialmente rápido em t para a distribuição estacionária π, independentemente das características de P e $\pi(0)$.

Quantos passos são necessários para atingir a convergência? Podemos definir $\epsilon > 0$ como critério de convergência e calcular t de forma que

$$d_{\mathrm{TV}}(\pi(t), \pi) = \epsilon$$

Considere o exemplo anterior onde

$$d_{\mathrm{TV}}(\pi(t), \pi) = \frac{3}{11} \cdot 12^{-t} \Rightarrow t = \frac{\log \epsilon + \log(11/3)}{-\log 12}$$

Se definirmos $\epsilon = 10^{-6}$, então $t = 5$. Ou seja, poucas transições são necessárias para se aproximar do equilíbrio.

Em geral, podemos afirmar que

$$t = \Theta\left(\log \frac{1}{\epsilon}\right)$$

e a constante em Θ depende de P e $\pi(0)$.

5.3 Tempo de Mistura

Vamos definir formalmente o tempo até o equilíbrio desejado.

Definição 5.3. O Tempo de Mistura é definido como o menor valor de t para o qual, independentemente da distribuição inicial $\pi(0)$, a distribuição $\pi(t)$ está a uma distância inferior ou igual a ϵ da distribuição estacionária. Denotamos isso como

$$\tau_\epsilon = \min\{t : \max\{\pi(0) \mid d_{\mathrm{TV}}(\pi(t), \pi) \leq \epsilon\}\}.$$

Teorema 5.4. Para qualquer cadeia de Markov aperiódica, irredutível temos que

$$\tau_\epsilon \leq \tau_{1/4} \log \frac{1}{\epsilon}$$

O teorema enunciado nos revela um resultado interessante, ao estar a $1/4$ da distribuição estacionária, minimizar ainda mais o erro é mais fácil, um fator $\log^{-\epsilon}$

5.4 Lacuna Espectral

A Lacuna Espectral está relacionado à rapidez com que a cadeia converge para o equilíbrio, refletindo a lacuna entre os autovalores da matriz de transição. Esses dois elementos são cruciais para entender a convergência e o desempenho de algoritmos MCMC na amostragem de distribuições complexas.

5.4.1 Velocidade de Convergência

A velocidade de convergência em uma cadeia de Markov é intrinsecamente ligada à relação entre os autovalores da matriz estocástica P, onde o segundo maior autovalor (em módulo) desempenha um papel crucial na determinação dessa convergência.

Definição 5.4. Denominamos *Lacuna Espectral* δ à distância entre os dois maiores autovalores (em módulo) de uma matriz estocástica P, expressa como

$$\delta = 1 - \max\{k > 1 : |\lambda_k|\}$$

onde o maior autovalor vai ser sempre 1.

Quanto maior for δ, mais rápida será a convergência. A base exponencial que governa essa convergência é determinada pelo segundo maior autovalor λ_2, enquanto todos os outros convergem para zero mais rapidamente.

Exemplo 5.2. Considere um fila com capacidade K onde

$$\begin{cases} p & := \text{ probabilidade de chegada} \\ q & := \text{ probabilidade de saída} \\ r = 1 - (p+q) & := \text{ probabilidade de permanecer no mesmo estado} \end{cases}$$

podemos definir duas filas:

1. $K = 4$, $p = 0.3$, $q = 0.4$

   Nesses parâmetros temos que

   $$\lambda_2 \approx 0.86 \quad \text{e} \quad \delta = 1 - 0.86 = 0.14$$

2. $K = 4$, $p = 0.1$, $q = 0.8$

   Nesses parâmetros temos que

   $$\lambda_2 \approx 0.56 \quad \text{e} \quad \delta = 1 - 0.56 = 0.44$$

No caso da fila o *Lacuna Espectral* depende da simetria de p e q, quanto maios parecidos, **menor**

a lacuna, mais lento será a convergência.

5.4.2 Limitante Inferior e Superior

Vamos expressar agora a relação de δ e τ_ϵ.

Considere uma Cadeia de Markov reversível, irredutível e aperiódica com Lacuna Espectral δ e $\pi_o = \min\{\pi_i : i \in S\}$ (menor valor na distribuição estacionária), temos a seguinte relação

$$\left(\frac{1}{\delta} - 1\right) \log \frac{1}{2\epsilon} \leq \tau_\epsilon \leq \frac{\log 1/\pi_o\epsilon}{\delta}$$

note que usar o limitante superior na prática não é fácil pois precisamos conhecer π_o e δ.

Exemplo 5.3. Considere uma das filas apresentadas em Exemplo 5.2, onde $K = 4$, $p = 0.3$, $q = 0.4$. Temos que

$$\delta = 0.14 \quad \text{e} \quad \pi_K = 0.015 \Rightarrow \pi_o = 0.015$$

Assumindo $\epsilon = 10^{-6}$, temos

$$\left(\frac{1}{0.14} - 1\right) \log \frac{1}{2 \cdot 10^{-6}} \leq \tau_\epsilon \leq \frac{\log 1/0.015 \cdot 10^{-6}}{0.14}$$

$$80.6 \leq \tau_\epsilon \leq 128.7$$

portanto com $t = 129$ estaremos 10^{-6} próximo de π, **independente da distribuição inicial**.

5.4.3 Tempo de Mistura em Função de n

Os resultados apresentados anteriormente referem-se a Cadeias de Markov de tamanho fixo. No entanto, existem situações em que o espaço de estados da Cadeia de Markov, denotado por n, pode aumentar consideravelmente de acordo com a complexidade do problema.

Defina $n = f(k)$ como o número de estados da Cadeia de Markov, onde k é um parâmetro do modelo. Por exemplo:

- $n = 2^k$: número de colorações de k vértices com duas cores.
- $n = k!$: permutações de k cartas.

Portanto, a convergência τ_ϵ é em função de n, veremos mais informações a seguir.

5.4.4 Passeios Aleatórios

Considere passeios aleatórios onde:

$$p_{i,i} = \frac{1}{2} \quad \text{e} \quad p_{i,j} = \frac{1}{2} - \frac{1}{\deg(i)}$$

ou seja, a cadeia permanece no mesmo estado com probabilidade $\frac{1}{2}$, caso contrário, escolhe um vizinho uniformemente. Esse tipo de passeio implica uma Cadeia de Markov aperiódica e irredutível.

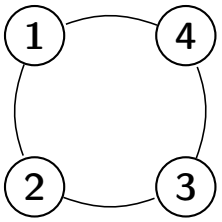
Em tais passeios, como a estrutura do grafo influencia no tempo de mistura? Estruturas como grafos completos, em anel, hipercubo, etc.

Seja τ_n o tempo de mistura com n vértices para um ϵ constante. Vamos apresentar a seguir alguns exemplos de estruturas de grafos juntamente com o seu tempo de mistura. Não iremos, no entanto, provar esses resultados.

- O tempo de mistura em **anéis** (ciclos) como n vértices é

  $$cn^2 \leq \tau_n \leq n^2$$

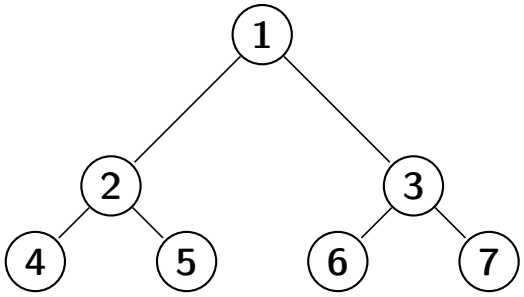
  ou seja é quadrático no tamanho do anel. Veja uma anel de 4 vértices abaixo

- Em **árvores binárias** completas com n vértices é

  $$\tau_n \leq 16n$$

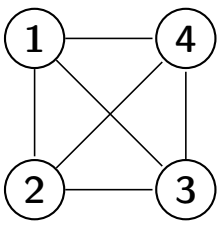
  ou seja é linear no tamanho da aŕvore. Veja uma árvore binária completa de 7 vértices abaixo:

- Em grafo completos de n vértices temos que

  $$\tau_n = 1$$

  ou seja **é constante!** Veja um grafo completo de 4 vértices abaixo:

6 Teorema Ergódico e Simulação de Cadeia de Markov

Nesta seção, vamos explorar conceitos-chave relacionados a cadeias de Markov. Iniciaremos com o caminho amostral, que destaca a sequência de estados percorrida pela cadeia ao longo do tempo. Em seguida, abordaremos o Teorema Ergódico e suas implicações para a estabilidade da cadeia. A simulação eficiente de cadeia será discutida, incluindo estratégias para gerar amostras representativas. Além disso, exploraremos desafios específicos ao lidar com cadeias de Markov de grande porte, oferecendo soluções práticas para simulação em larga escala. Esta seção proporciona uma visão prática e acessível da análise e simulação de cadeia, desde conceitos fundamentais até técnicas eficientes.

6.1 Caminho Amostral

Definição 6.1. Um caminho amostral é uma realização de uma sequência de variáveis aleatórias na evolução do tempo t, ou seja

$$X_t, \text{ para } t = 0, 1, \ldots$$

definimos um caminho amostral ω como

$$\omega = (\omega_0, \omega_1, \ldots)$$

A probabilidade de uma Cadeia de Markov realizar exatamente ω é

$$\begin{aligned} \Pr[\omega] &= \Pr[X_0 = \omega_0 \wedge X_1 = \omega_1 \wedge \ldots] \\ &= \pi_{\omega_0}(0) p_{\omega_0,\omega_1} p_{\omega_1,\omega_2} \cdots \end{aligned}$$

Observação. Veja que para todo caminho amostral ω tem uma probabilidade converge a 0 com o comprimento do caminho.

Exemplo 6.1. Dado o modelo On-Off visto em Exemplo 3.1 e os seguintes caminhos amostrais e suas probabilidades:

$$\begin{aligned} \omega &= (1,1,1,1,1,1,1,1,1,1) \Rightarrow \Pr[\omega] = 0.8\,(3/4)^9 \\ \omega &= (0,1,0,1,0,1,0,1,0,1) \Rightarrow \Pr[\omega] = 0.2 \prod_{i=1}^{9} \frac{1 + (i \mod 2)}{3} \end{aligned}$$

nos dois casos temos que $\Pr[\omega]$ é um valor pequeno.

6.2 Convergência em Termos do Caminho Amostral

Como todo caminho amostral tem probabilidade que converge para 0 conforme t aumenta, o que podemos dizer sobre a sequência X_t, para $t = 0, 1, \ldots$? Podemos utilizar a média sobre os valores da

sequência, para ver a sua convergência quando k é grande, usando:

- A média amostral dos valores observados

$$S_k = \frac{1}{k}\sum_{t=0}^{k-1} X_t \Rightarrow \lim_{k\to\infty} S_k = \mathbb{E}_\pi[X] = \sum_{s\in S} s\pi_s$$

- A fração de vezes que um certo estado s é visitado:

$$F_k(s) = \frac{1}{k}\sum_{t=0}^{k-1} I(X_t = S) \Rightarrow \lim_{k\to\infty} F_k(s) = \pi_s$$

Teorema 6.1 (**Teorema Ergódico**). Considere uma cadeia de Markov com um espaço de estados S e

$$f : S \to \mathbb{R}$$

se a cadeia é irredutível e aperiódica com distribuição estacionária π, temos que

$$\Pr\left[\lim_{k\to\infty} \frac{1}{k}\sum_{t=0}^{k-1} f(X_t) = \mathbb{E}_\pi[f(X)]\right] = 1$$

O **Teorema Ergódico** é um **teorema fundamental**, ele nos diz que a *média no tempo* converge para *média no espaço*. Os exemplos anteriores são apenas casos especiais.

6.2.1 Estimando a Distribuição Estacionária

O Teorema 6.1 garante que o Método de Monte Carlo funciona em Cadeias de Markov, ou seja é a **conexão com a teoria e prática**, análogo à lei dos grandesnúmeros!

Podemos utilizar a própria Cadeia de Markov para estimar sua distribuição estacionária π, a ideia é gerar um caminho amostral ω bem longo e calcular a fração de visitas a cada estado, denotamos:

$$\hat{\pi}_s(k) = \frac{1}{k}\sum_{t=0}^{k-1} I(\omega_t = s)$$

6.3 Simulando uma Cadeia de Markov

Em suma, simular uma cadeia de Markov é gerar um caminho amostral, podemos fazer isto de diversas formas, iremos explorar elas a seguir:

6.3.1 Força Bruta

Podemos utilizar diretamente os caminhos amostrais:

1. Enumerar todos os caminhos amostrais de tamanho k;
2. Determinar a probabilidade de cada caminho;

3. Gerar amostras deste conjunto.

A intuição é que cada caminho amostral pode ser uma face de um dado enviesado. Sabemos como gerar amostras de um dado enviesado, veja na Subseção E.2.

Exemplo 6.2. Vamos simular o modelo On-Off visto em Exemplo 3.1, podemos definir

$$\pi(0) = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix}$$

Considerando todos os caminhos de tamanho k, ou seja:

$$\omega_0 = (0,0,0,0,0), \omega_1 = (0,0,0,0,1) \ldots, \omega_{31} = (1,1,1,1,1)$$

temos então que

$$\Pr[\omega_0] = 0.2 \cdot \left(\frac{1}{3}\right)^4, \Pr[\omega_1] = 0.2 \left(\frac{1}{3}\right)^3 \frac{2}{3}, \ldots$$

ou seja temos um dado enviesado de 32 faces.

Podemos gerar amostras de um dado enviesado em **tempo constante** utilizando o Método Alias (visto em Subseção E.2). Mas temos um **grande problema**, o número de caminhos é exponencial em k, ou seja levamos tempo exponencial para construir o dado.

6.3.2 Método Iterativo

Podemos amostrar a sequência de variáveis aleatórias X_t de forma iterativa, para $t = 0, 1, \ldots$. Dessa maneira, construímos dinamicamente o caminho amostral:

- Utilizamos $\pi(0)$ para gerar X_0.
- Dado X_0, empregamos a matriz estocástica P para gerar X_1, ou seja,

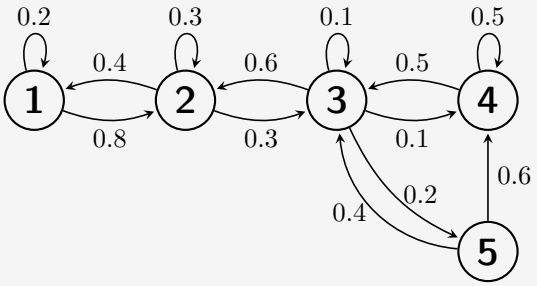
$$\Pr[X_1 = s_1 \mid X_0 = s_0] = p_{s_0,s_1}$$

- Dado X_1, utilizamos novamente a matriz estocástica P para gerar X_2, ou seja,

$$\Pr[X_2 = s_2 \mid X_1 = s_1] = p_{s_1,s_2}$$

- De forma geral, dado X_k, empregamos P para gerar X_{k+1}.

Exemplo 6.3. Seja uma Cadeia de Markov representada pelo grafo abaixo:



Dado $X_4 = 3$, podemos gerar X_5 a partir dos vizinhos de saída do estado 3. Como há 4 estados possíveis $\{2, 3, 4, 5\}$ com a probabilidades de transições: $p_{3,2}, p_{3,3}, p_{3,4}, p_{3,5}$. Ou seja, podemos utilizar um dado enviesado com 4 faces de acordo com as respectivas probabilidades.

Uma maneira intuitiva de implementar o método é construir um dado enviesado para cada linha da matriz P, onde cada face corresponde a um estado da Cadeia de Markov, utilizando diretamente a linha de matriz P, veja a ideia em Exemplo 6.3. Por exemplo a linha i da matriz P:

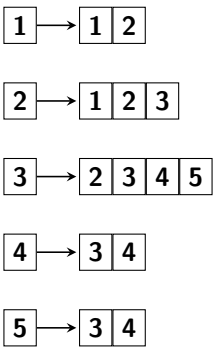
$$p_{i,j}, \forall j \in S$$

Utilizamos um algoritmo básico para gerar amostra de um dado aleatório, um tempo médio $n/2$, mas um dos problemas é quando a matriz P é esparsa[6].

6.3.3 Simulação Eficiente da Cadeia de Markov

Vamos ilustrar como simular eficientemente uma Cadeia de Markov.

Podemos representar a matriz de transições de estado P como um vetor de adjacência, de modo que apenas as entradas **não nulas** são representadas. Dado o grafo em Exemplo 6.3, criamos os vetores de adjacência:



Dessa forma, podemos criar dois vetores para cada estado j:

$y_j[i] :=$ estado destino da i-ésima transição não nula de j

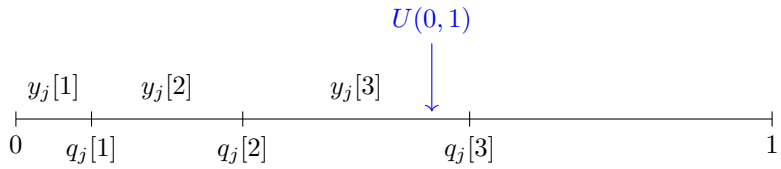
$q_j[i] :=$ probabilidade de transição acumulada pelas primeiras i-ésimas transições de saída de j

Por exemplo, nesta cadeia, temos que:

[6] Há muitas transições nulas na matriz.

$$y_2[2] = 2 \quad \text{e} \quad q_2[2] = 0.7,$$
$$y_5[2] = 4 \quad \text{e} \quad q_5[2] = 1$$

Portanto, para gerar o próximo estado a partir de um estado atual j, podemos utilizar os vetores y e q para simular um dado enviesado. Veja a ilustração da amostragem de uma variável uniforme em $[0..1]$ para gerar o próximo estado:



Ou seja, a geração do próximo estado tem complexidade $O(d)$, onde d é o grau de saída do estado.

Poderíamos utilizar o Método Alias (visto em Subseção E.2) para gerar em tempo constante o próximo estado, mas pagaríamos o custo inicial de construir o dado enviesado no método. Podemos optar por usar o Método Alias levando em consideração se o tempo médio de retorno dos estados são pequenos e o tamanho do caminho amostral seja grande. Se voltarmos muito ao estado (o tempo de retorno τ é pequeno), vale a pena pagar o custo inicial do método, pois iremos gerar muitas amostras no caso em que o caminho amostral seja longo.

6.3.4 Gerando Amostras

Em alguns casos, o caminho amostral não nos interessa e desejamos apenas uma amostra no tempo t. Podemos seguir as seguintes abordagens:

1. **Abordagem Matricial**
   (a) Calcular $\pi(t)$, a distribuição no tempo t, onde $\pi(t) = \pi(0)P^t$
   (b) Gerar amostras dessa distribuição
2. **Abordagem Iterativa**
   (a) Gerar um caminho amostral até X_t
   (b) Retornar a amostra gerada para X_t

Vamos avaliar a eficiência de cada uma. Considere que queremos gerar r amostras. Temos então que:

- Calcular $\pi(t)$ tem complexidade $O(tn^2)$, pois envolve t multiplicações de vetor por matriz. A geração de uma amostra tem complexidade $O(n)$. Se r for grande, podemos utilizar o Método Alias em $\pi(t)$ para gerar amostras de forma constante, mas ainda assim será necessário realizar as multiplicações de vetor por matriz.

- Gerar X_t tem complexidade $O(td)$ para cada amostra. Se r for grande, podemos usar o Método Alias nas transições da cadeia de Markov, com um custo amortizado de $O(t)$ por amostra.

Para gerar amostras da distribuição estacionária usamos a mesma abordagem com a diferença que na abordagem matricial temos que calcular a distribuição estacionária e na segunda abordagem tempos que simular a Cadeia de Markov para t grande o suficiente, ou seja, para $t \geq \tau_\epsilon$.

6.4 Simulando Cadeias de Markov Enormes

No caso de simular Cadeias de Markov cujo o espaço de estados é muito grande, ou até mesmo infinito, não conseguimos representar a cadeia na memória.

Para tais cadeias podemos gerar apenas os possíveis próximos estados a partir do estado atual, a dinâmica é:

1. Determinar as transições de saída do estado atual;
2. Construir o dado enviesado
3. Escolhe o próximo estado de acordo

Portanto temos a restrição de gerar transições de saída a partir do estado atual sem conhecer a Cadeia de Markov por inteiro, ou seja, no caso em que há regras de definem possíveis transições.

Exemplo 6.4. Seja um passeio aleatório em hipercubo de 100 dimensões, a cadeia possui 2^{100} estados, impossível de representar em qualquer computador.

Podemos representar um estado da cadeia como um vetor binário de 100 bits, cujo os estados vizinhos são os vetores que diferem em 1 bit[a].

Para gerarmos o próximo estado basta escolher uniformemente um bit do estado atual e inverter-lo

Portanto a complexidade de transição é $O(1)$.

[a] Está é definição de hipercubo.

7 Markov Chain Monte Carlo e Metropolis-Hastings

Nesta seção, finalmente alcançamos nosso objetivo: introduzir o Método de Monte Carlo via Cadeia de Markov (MCMC). Abordaremos a amostragem de espaços complexos, apresentaremos o MCMC e discutiremos um caso específico e crucial do MCMC, o **Metropolis-Hastings**. Em seguida, exploraremos o Gibs Sampling.

O Método de Monte Carlo via Cadeia de Markov (MCMC) representa uma abordagem poderosa para a amostragem de espaços complexos, sendo fundamental em diversas aplicações, especialmente na estatística bayesiana e em problemas computacionalmente desafiadores. Utilizando métodos de Monte Carlo e cadeias de Markov, os quais discutimos anteriormente.

O desenvolvimento do MCMC ganhou impulso significativo na década de 1950 com a colaboração entre Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller e Edward Teller. A contribuição mais notável desse grupo foi a introdução do famoso algoritmo de Metropolis-Hastings [7], um dos pilares do MCMC, que se tornou uma ferramenta essencial para a geração de amostras representativas de distribuições de probabilidade complexas. Considerado um dos 10 mais influentes algoritmos do Século XX pela revista IEEE Comput in Science & Engineering [8].

Ao longo das décadas seguintes, o MCMC evoluiu e se tornou uma técnica amplamente utilizada em estatística, aprendizado de máquina [9], física computacional e outras disciplinas como o processamento de imagens [10]. Sua capacidade de explorar eficientemente espaços de parâmetros complexos e multidimensionais o tornou uma ferramenta indispensável para a inferência estatística e a modelagem probabilística em diversas áreas do conhecimento.

7.1 Geração de Amostras Uniformes de Maneira Iterativa

Como podemos gerar uma amostra uniforme sobre o espaço amostral S de todos os grafos de n vértices? Podemos gerar de forma iterativa.

Seja S o conjunto de todos os grafos com n vértices, onde $S = 2^{n(n-1)/2}$[7]. Podemos então gerar uma amostra para cada aresta sobre uma distribuição Bernouli($\frac{1}{2}$), para a inclusão ou não da aresta no grafo. Feito isso para todas as arestas possíveis temos então um grafo escolhido uniformemente.

7.2 MCMC

Veja que se adicionarmos a restrição de grafo conexo no exemplo da seção anterior, ou seja, gerar uniformemente um grafo conexo de n vértices, o método usado já não funciona. Portanto há espaços complicados que a maneira iterativa não funciona ou funciona de maneira ineficiente.

Vamos utilizar então **Método de Monte Carlo via Cadeia de Markov!** Tal método é basedo em Cadeia de Markov para gerar amostras de espaços arbitrários com qualquer **distribuição de probabilidade** (não precisa ser uniforme).

Nos casos onde o espaço amostral S é muito grande construímos a Cadeia de Markov iterativamente. Para determinar as transições precisamos de uma descrição fácil dos estados vizinhos a partir

[7] O número de arestas que um grafo de n vértices pode ter é $n(n-1)/2$.

do estado atual. Para uma melhor convergência precisamos de poucas transições de saída de cada estado e um baixo tempo de mistura.

Observação. O Markov Chain Monte Carlo é tema atual de muita pesquisa!

7.2.1 Algoritmo para MCMC

Dado uma descrição do espaço amostral S e a distribuição de probabilidade π sobre os estados, retornamos uma amostra aleatória de S de acordo com a distribuição π. Os passos são os seguintes:

1. Construir uma cadeia de Markov (CM) irredutível, onde cada estado corresponde a um elemento do espaço amostral, formando assim a cadeia base.
2. Transformar a cadeia base em outra CM que seja reversível e possua distribuição estacionária π.
3. Simular um caminho amostral suficientemente longo e, ao final, retornar o estado alcançado.

7.3 MCMC - Caso Simétrico

Considere o caso onde a cadeia base tem matriz de transição P simétrica, ou seja, $p_{i,j} = p_{j,i}$ para todo estado $i, j \in S$.

Vamos modificar P para que a CM seja reversível com distribuição estacionária π, onde π é a entrada do problema. A ideia é que a nova cadeia **não aceita todas as transições da cadeia base**, criando laços para induzir a distribuição estacionária.

Seja $a(i, j)$ a probabilidade de aceitar a transição $i \to j$. Podemos definir então a nova matriz de transição P' como

$$p'_{i,j} = \begin{cases} p_{i,j}a(i,j) & \text{se } i \neq j \\ 1 - \sum_{k:k\neq i} p_{i,k}a(i,k) & \text{se } i = j \end{cases}$$

portanto temos que escolher $a(i, j)$ tal que P' a nova CM seja reversível, onde a distribuição estacionária π seja dada por

$$\pi_i p_{i,j} a(i,j) = \pi_j p_{j,i} a(j,i)$$

Como $p_{i,j} = p_{i,j}$ temos que

$$\begin{aligned} \pi_i a(i,j) &= \pi_j a(j,i) \\ a(i,j) &= \frac{\pi_j}{\pi_i} a(j,i) \end{aligned}$$

veja que temos infinitas soluções para o par $a(i, j)$ e $a(j, i)$, mas nós queremos maximizar a probabili-

dade de aceite, por questões de eficiência[8], logo

$$a(i,j) = \begin{cases} 1 & \text{se } \pi_i \leq \pi_j \\ \pi_j/\pi_i & \text{caso contrário} \end{cases}$$

ou seja

$$a(i,j) = \min\left\{1, \frac{\pi_j}{\pi_i}\right\} \quad \text{e} \quad a(j,i) = \min\left\{1, \frac{\pi_i}{\pi_j}\right\}$$

Exemplo 7.1. Vamos gerar amostras de pares ordenados (x, y) em $[1..n] \times [1..n]$, a distribuição de probabilidade é:

$$\Pr[(x,y)] = \frac{(x+y)^2}{Z}$$

onde Z é uma constante de normalização definida como:

$$Z = \sum_{(x,y)\in[1..n]^2} (x+y)^2$$

A Cadeia de Markov base tem como espaço de estados os pares $(x, y) \in [1..n]^2$, onde cada estado tem 4 transições: norte, sul, leste e oeste com possíveis laços.

Veja que dessa forma cadeia base é simétrica, $p_{i,j} = p_{j,i}$.

Agora precisamos alterar a cadeia base com matriz de transição P para uma matriz de transição P' de tal forma que

$$\pi_{(x,y)} = \frac{(x+y)^2}{Z}$$

Seja o estado $i = (x_i, y_i)$ e o estado $j = (x_j, y_j)$, definimos então $a(i, j)$ como

$$a(i,j) = \min\left\{1, \frac{\pi_j}{\pi_i}\right\} = \min\left\{1, \frac{(x_j+y_j)^2}{(x_i+y_i)^2}\right\}$$

portanto P' é definida como

$$p'_{i,j} = \begin{cases} p_{i,j}a(i,j) & \text{se } i \neq j \\ 1 - \sum_{k:k\neq i} p_{i,k}a(i,k) & \text{se } i = j \end{cases}$$

A boa notícia é que não precisamos calcular Z para definir a probabilidade de aceite $a(i, j)$.

Por fim para gerar uma amostra, basta simular a CM definida por P' por um tempo τ_ϵ e retornar o estado final.

[8] A ideia é que quanto maior a probabilidade de aceite, maior a probabilidade transição entre estados, transitando mais fácil na cadeia

7.4 Caso Geral

Vamos definir o caso geral do MCMC:

1. Dado uma descrição do espaço amostral S e a distribuição de probabilidade π sobre os estados, construímos uma Cadeia de Markov irredutível com matriz de transição P, onde cada estado corresponde a um elemento do espaço amostral, formando assim a cadeia base.

2. Construímos uma nova matriz de transição P', definida por

$$p'_{i,j} = \begin{cases} p_{i,j}a(i,j) & \text{se } i \neq j \\ 1 - \sum_{k:k\neq i} p_{i,k}a(i,k) & \text{se } i = j \end{cases}$$

   onde $a(i,j)$ é a probabilidade de aceite da transição $i \to j$. A nova cadeia tem que ser reversível, ou seja

$$\pi_i p_{i,j} a(i,j) = \pi_j p_{j,i} a(j,i)$$

3. Simular um caminho amostral suficientemente longo (τ_ϵ) com base em P' e, ao final, retornar o estado alcançado.

7.5 Metropolis–Hastings

No caso simétrico de MCMC existem infinitas soluções para o par $a(i,j)$ e $a(j,i)$. Mas por objetivos de eficiência queremos maximizar a probabilidade de aceite, logo

$$a(i,j) = \begin{cases} 1, & \text{se } \pi_i p_{i,j} \leq \pi_j p_{j,i} \\ \pi_j p_{j,i} / \pi_i p_{i,j} & \text{caso contrário} \end{cases}$$

ou seja

$$a(i,j) = \min\left\{1, \frac{\pi_j p_{j,i}}{\pi_i p_{i,j}}\right\}$$

A cadeia definida com a matriz de transição P' definida com essa probabilidade de aceite é chamada de Metropolis–Hastings, dando origem ao famoso algoritmo Metropolis–Hastings.

7.5.1 Amostrando Vértices de uma Rede

Dado o conjunto de vértices, podemos querer amostrar (de maneira simples) um vértice do conjunto. Podemos construir uma cadeia Metropolis–Hasting tal que π seja sua distribuição estacionária:

$$\pi_v = \frac{1}{Z}$$

onde Z é o número de vértices da rede. E sua matriz de transição é P, definida por

$$p_{i,j} = \frac{1}{\deg(i)}$$

ou seja, a cadeia base é um passeio aleatório simples.

Agora definimos a probabilidade aceite da nova cadeia como

$$\begin{aligned} a(i,j) &= \min\left\{1, \frac{\pi_j p_{j,i}}{\pi_i p_{i,j}}\right\} \\ &= \min\left\{1, \frac{\deg(i)}{\deg(j)}\right\} \end{aligned}$$

logo

$$p'_{i,j} = \begin{cases} \frac{1}{\deg(i)} \min\left\{1, \frac{\deg(i)}{\deg(j)}\right\} & \text{se } i \neq j \\ 1 - \sum_{k:k\neq i} p_{i,k} \min\left\{1, \frac{\deg(i)}{\deg(j)}\right\} & \text{se } i = j \end{cases}$$

Dessa maneira enviesamos o passeio contra vértices de grau alto, adicionando laços, tornando a distribuição uniforme.

Mas como não conhecemos o grafo, podemos simular o caminho amostral iterativamente, a cada passo:

1. Descobrimos os vizinhos do vértice atual
2. Descobrimos o grau de cada vizinho do vértice atual
3. Determinamos as probabilidade transição de cada vizinho
4. Fazemos a escolha aleatória, atualizamos o vértice atual e voltamos ao passo 1

7.6 Gibs Sampling

O Gibs Sampling ou Glauber Dynamics é um algoritmo para construir uma Cadeia de Markov com distribuição estacionária π sobre o espaço de estados S, onde os elementos do espaço de estados S é um vetor. Para mais detalhes sobre o Gibs Sampling consulte [11].

$$V = \Big(V_1, \ldots, V_n\Big)$$

Cada V_i assume valores de um certo conjunto K, onde

$$|S| \leq |K|^{|V|}$$

ou seja nem todas as possíveis combinações V precisam estar em S.

Esta Cadeia de Markov baseia-se na distribuição condicional de uma variável dado o valor de todas

as outras, ou seja

$$\Pr[V_k = a_k \mid V_1 = a_1, \dots V_n = a_n]$$

esta probabilidade precisar ser conhecida a priori ou encontrada para induzirmos π.

As probabilidades de transição são proporcionais as probabilidade condicionais. A probabilidade de transição é a probabilidade condicional dividida por n. Veja um exemlpo em Figura 5.

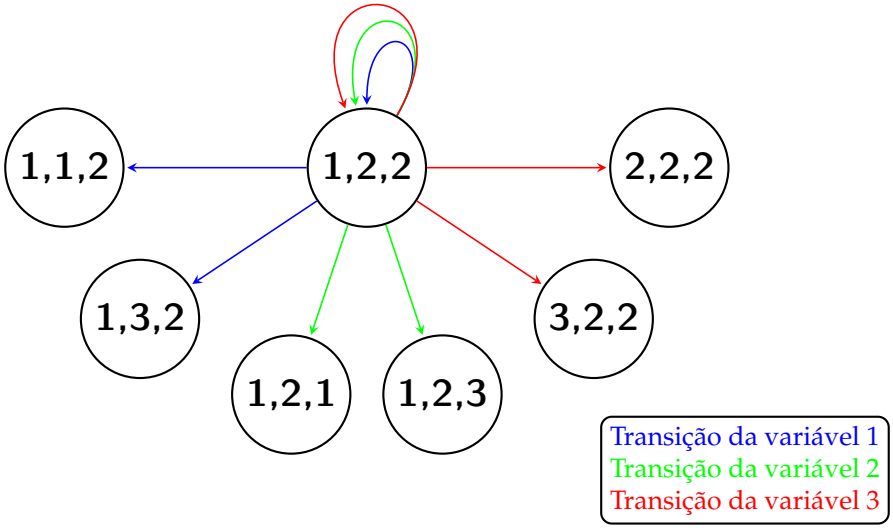


Figura 5: Exemplo Cadeia Gibs Sampling

Seja $s_i = (1, 2, 2)$ e $s_j = (1, 2, 3)$, temos que a probabilidade da transição $s_i \to s_j$ é dada por

$$\frac{1}{3} \Pr[V_3 = 2 \mid V_1 = 1, V_2 = 2]$$

A Cadeia de Markov construída pelo método Gibs Sampling é **reversível**.

7.6.1 Algoritmo

Vamos ilustrar brevemente o algoritmo para o método Gibs Sampling:

1. Dado um estado atual $X_t = (V_1(t), V_2(t), \dots, V_n(t))$;
2. Escolhemos um variável k de forma uniforme, ou seja $k \sim U(1, n)$;
3. Escolhemos então um valor a_k para V_k dado a probabilidade condicional e número de variáveis;
4. Transicionar para o estado X_{t+1} copiando os valores de $V_i(t)$ e atualizando apenas V_k com o valor a_k;
5. Repetir itens anteriores por τ_ϵ passos para que a distribuição de X_t esteja próxima da distribuição estacionária π.

7.6.2 Distribuição Conjunta

Considere as variáveis aleatórias (X, Y) descritas para seguinte distribuição conjunta

	X									
Y	1	2	3	4	5	6	7	8	9	10
1	0.02	0.03	0.05	0.07	0.1	0.1	0.1	0.1	0.1	0.1
2	0.01	0.03	0.05	0.07	0.1	0.1	0.1	0.1	0.1	0.1
3	0.01	0.02	0.04	0.06	0.08	0.1	0.1	0.1	0.1	0.1
4	0.01	0.02	0.03	0.05	0.07	0.1	0.1	0.1	0.1	0.1
5	0.01	0.02	0.03	0.05	0.07	0.1	0.1	0.1	0.1	0.1

Podemos obter a distribuição condicional a partir da distribuição conjunta e gerar amostras (X, Y) usando Gibs Sampling.

O método é o seguinte:

1. Escolhemos um $Z_0 = (X_0, Y_0)$ de forma arbitrária, por exemplo $Z_0 = (1, 1)$;
2. Escolhemos uniformemente entre as variáveis X e Y (lançamos uma moeda);
3. Escolhemos um novo valor a_t para X ou Y utilizando a distribuição condicional;
4. Atualizamos Z_{t+1} com (X_t, a_t) se eu escolhi Y no passo 2, ou (a_t, Y_t) caso contrário;
5. Repetimos τ_ϵ vezes e retornamos Z_{τ_ϵ} como sendo a amostra.

8 Conclusão

A pesquisa abordou inicialmente os Métodos de Monte Carlo por meio de Cadeias de Markov, destacando particularmente duas metodologias proeminentes: Metrópolis-Hastings e Gibbs Sampling. Essas abordagens revelaram-se ferramentas valiosas, especialmente ao enfrentarmos desafios complexos, como a inferência bayesiana ([12] e [13]). Ao longo da análise, dedicamos atenção à questão importante da convergência para estacionariedade em relação ao tempo de mistura.

O tempo de mistura desempenha um papel crucial na geração de amostras com precisão definida. No entanto, determinar esse tempo não é uma tarefa trivial, sendo um tema atualmente investigado nos campos da estatística e da computação ([14]). Nossa análise proporcionou uma visão geral desses conceitos, evidenciando sua relevância prática. Vale ressaltar que, apesar de fornecermos uma perspectiva geral neste trabalho, a compreensão precisa do tempo de mistura permanece como um desafio significativo.

Oferecemos um panorama abrangente do tema de Métodos de Monte Carlo nesta pesquisa. No entanto, é crucial reconhecer que o campo está em constante evolução, com pesquisas atuais e emergentes moldando nossa compreensão sobre o assunto. Este trabalho estabelece uma base para a compreensão da literatura existente sobre Métodos de Monte Carlo via Cadeias de Markov, estimulando futuros estudos.

Agradecimentos

Gostaria de expressar minha sincera gratidão a todas as pessoas que contribuíram para o sucesso deste trabalho. Em especial, quero agradecer:

- Ao meu orientador, André Vignatti, por sua orientação valiosa e apoio contínuo ao longo do processo de pesquisa.
- Aos meus amigos e familiares, por seu incentivo constante e compreensão.
- Ao sistema de ensino público brasileiro, em particular à UFPR, pela oportunidade de estudar, pelo acesso a recursos educacionais e pela contribuição significativa para o meu desenvolvimento acadêmico.

Referências

[1] Stanislaw Ulam. *Adventures of a Mathematician*. University of California Press, Berkeley, CA, 1976.

[2] Stanislaw Ulam and John von Neumann. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1950.

[3] Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

[4] Sebastian Thrun, Dieter Fox, Wolfram Burgard, and Frank Dellaert. Robust monte carlo localization for mobile robots. *Artificial Intelligence*, 128(1):99–141, 2001.

[5] Olle Haggstrom. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2001.

[6] Andrey Markov. On the distribution of large numbers dependent on each other. *Mathematische Annalen*, 62(1):1–42, 1906.

[7] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[8] J. Dongarra and F. Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science & Engineering*, 2(01):22–23, 2000.

[9] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

[10] Zhuowen Tu, Song-Chun Zhu, and Heung-Yeung Shum. Image segmentation by data driven markov chain monte carlo. 2:131–138 vol.2, 2001.

[11] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[12] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

[13] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Bayesian modeling and inference for nonlinear state space models: A flexible, efficient, and robust sampler. *Statistics and Computing*, 2010.

[14] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[15] Sheldon Ross. *A First Course in Probability*. Pearson, 10 edition, 2019.

[16] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

A Fundamentos da Teoria da Probabilidade

Nesta seção, abordamos conceitos fundamentais na teoria da probabilidade. Iniciamos com o conceito de "espaço amostral", que representa todos os resultados possíveis de um experimento aleatório. Em seguida, exploramos a "probabilidade" como a medida da chance de ocorrência de eventos específicos dentro desse espaço amostral. Discutimos também os "eventos", que são conjuntos de resultados, e introduzimos conceitos fundamentais como "independência" e "exclusão mútua". Destacamos a relevância da "probabilidade total" e apresentamos a "Regra de Bayes" como ferramenta crucial para análises probabilísticas mais avançadas. Para uma compreensão mais aprofundada desses conceitos, recomendamos a leitura de [15].

Definição A.1 (**Espaço Amostral**). O espaço amostral S é um conjunto de objetos enumerável.

Exemplo A.1. Dado o espaço amostral S abaixo

$$S = \{a, b, c, \ldots, z\}$$

onde $|S|$ é a sua cardinalidade

$$|S| = 26$$

Definição A.2 (**Probabilidade**). Função que associa a cada elemento de S um valor entre 0 e 1:

$$p : S \to [0, 1]$$

de forma que

$$\sum_{s \in S} p_s = 1$$

Definição A.3 (**Evento**). Um **evento** é um subconjunto do espaço amostral.

Exemplo A.2. Dado o espaço amostral

$$S = \{a, b, c, \ldots, z\}$$

temos alguns possíveis eventos:

- $A = \{a, b, c, d\}$
- $V =$ todas as vogais

- $C =$ todas as consoantes

Definição A.4 (Probabilidade de um Evento). A **probabilidade de um evento** é a soma das probabilidades dos elementos que compõem o evento:

$$\Pr[A] = \sum_{e \in A} p_e$$

Exemplo A.3. Dado o espaço amostral

$$S = \{a, b, c, \ldots, z\}$$

e os eventos

- $A = \{a, b, c, d\}$
- $V =$ todas as vogais
- $C =$ todas as consoantes

temos as probabilidades

- $\Pr[A] = \frac{4}{26} = \frac{2}{13}$
- $\Pr[V] = \frac{5}{26}$
- $\Pr[C] = \frac{21}{26}$

Notação. Para uma melhor notação e usabilidade vamos usar operadores lógicos no lugar de operadores de conjunto:

- $A \cup B = A \vee B$
- $A \cap B = A \wedge B$
- $A^c = \overline{A}$

Definição A.5 (Independência). Dois eventos A e B são independentes se e somente se

$$\Pr[A \wedge B] = \Pr[A] \cdot \Pr[B]$$

Exemplo A.4. Dado o espaço amostral $S = \{a, b, c, \ldots, z\}$ e os eventos:

- $A = \{a, b, c, d\}$
- $B =$ todas são consoantes
- $C =$ todas as letras antes de n)
- $D = \{a, z\}$

veja que

$$\begin{aligned}\Pr[A \wedge B] = \Pr[\{b, c, d\}] = \frac{3}{26} &\neq \Pr[A] \cdot \Pr[B] \\ &= \frac{4}{26} \cdot \frac{21}{26} \\ &= \frac{84}{676} \\ &= \frac{21}{169}\end{aligned}$$

e

$$\begin{aligned}\Pr[C \wedge D] = \Pr[\{a\}] = \frac{1}{26} &= \Pr[C] \cdot \Pr[D] \\ &= \frac{13}{26} \cdot \frac{2}{26} \\ &= \frac{1}{26}\end{aligned}$$

note que a intuição pode nos enganar pois A e B não são independentes mas C e D são!

Definição A.6 (Exclusão Mútua). Dois eventos A e B são mutuamente exclusivos se e somente se

$$\Pr[A \vee B] = \Pr[A] + \Pr[B]$$

Exemplo A.5. Dado o espaço amostral $S = \{a, b, c, \ldots, z\}$ e os eventos:

- $A = \{a, b, c, d\}$
- $B :=$ todas são consoantes
- $C :=$ todas as letras antes de n
- $D = \{a, z\}$

veja que

- $A \vee B :=$ todas as consoantes mais a letra a

- $C \vee D$:= todas as letras antes de n mais a letra z

logo

$$\begin{aligned}\Pr[A \vee B] = \frac{22}{26} &\neq \Pr[A] + \Pr[B] \\ &= \frac{4}{26} + \frac{21}{26} \\ &= \frac{25}{26}\end{aligned}$$

e

$$\begin{aligned}\Pr[C \vee B] = \frac{14}{26} &\neq \Pr[C] + \Pr[D]] \\ &= \frac{13}{26} + \frac{2}{26} \\ &= \frac{15}{26}\end{aligned}$$

portanto A e B não são mutuamente exclusivos e C e D apesar de serem independentes eles não são mutuamente exclusivos.

Definição A.7 (Probabilidade Condicional). A **probabilidade condicional** é a probabilidade de do evento A acontecer dado um evento B ocorreu. Desta maneira o novo espaço amostral para a ocorrência de A passa ser o conjunto B, que é dado por

$$\Pr[A|B] = \frac{\Pr[A \wedge B]}{\Pr[B]}$$

Exemplo A.6. Sejam

- o espaço amostral $S = \{a, b, c, \ldots, z\}$
- $A = \{a, b, c, d\}$
- $B :=$ todas as consoantes
- $C = \{x, y, z\}$

temos que

- $\Pr[A|B]$ é:

$$\begin{aligned}\Pr[A|B] &= \frac{\Pr[A \wedge B]}{\Pr[B]} \\ &= \frac{\frac{3}{26}}{\frac{21}{26}} \\ &= \frac{3}{21} \\ &= \frac{1}{7}\end{aligned}$$

- $\Pr[B|A]$ é:

$$\begin{aligned}\Pr[B|A] &= \frac{\Pr[B \wedge A]}{\Pr[A]} \\ &= \frac{\frac{3}{26}}{\frac{4}{26}} \\ &= \frac{3}{4}\end{aligned}$$

- $\Pr[C|B]$ é:

$$\begin{aligned}\Pr[C|B] &= \frac{\Pr[B \wedge C]}{P[B]} \\ &= \frac{\frac{3}{26}}{\frac{21}{26}} \\ &= \frac{1}{7}\end{aligned}$$

Definição A.8 (Particionamento). Um **particionamento** de um conjunto S é um conjunto de subconjuntos de S tal que todo elemento de S aparece exatamente em um subconjunto.

Exemplo A.7. Dado o espaço amostral $S = \{a, b, c, \ldots, z\}$ podemos ter os particionamentos:

- $O_1 :=$ todas as vogais
- $O_2 :=$ toas as consoantes

Teorema A.1 (Lei da Probabilidade Total). Dado um espaço amostral S, um evento A e B_i uma

partição qualquer de S, a lei de probabilidade total no diz:

$$\Pr[A] = \sum_i \Pr[A \wedge B_i] = \sum_i \Pr[A|B_i] \Pr[B_i]$$

Teorema A.2 (Regra de Bayes). A **Regra de Bayes** é uma relação fundamental entre probabilidades condicionais expressa por

$$\Pr[A|B] = \frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B]}$$

Observação. A **Regra de Bayes** é empregada quando desejamos calcular $\Pr[A|B]$, mas calcular essa probabilidade diretamente através da fórmula simples de probabilidade condicional pode ser desafiador. Em alguns casos, pode ser mais viável calcular $\Pr[B|A]$, e, nesses casos, a **Regra de Bayes** oferece uma alternativa eficaz.

B Variáveis Aleatórias e Distribuições

Nesta seção, exploraremos as variáveis aleatórias, focando em suas distribuições. As variáveis aleatórias fornecem uma maneira de atribuir valores numéricos aos resultados de experimentos aleatórios, sendo essenciais para entender padrões e comportamentos probabilísticos. Vamos definir diferentes distribuições e propriedades estatísticas associadas a essas variáveis.

B.1 Variável Aleatória

Definição B.1 (Variável Aleatória). Uma variável aleatória X é uma função que mapeia o espaço amostral nos inteiros:

$$X : S \to \mathbb{R}$$

Usaremos a variável aleatória (v.a) X para definir eventos em função de seus valores e elementos do espaço amostral.

Para uma melhor usabilidade denotaremos que dado uma v.a X, denotamos $X > 5$ de forma que

$$A = \{X > 5\} = \{e \in S \mid X(e) > 5\}$$

B.1.1 Probabilidade de V.A

A probabilidade de uma v.a é de dada pelo evento definido pela mesma. Sejam

- o espaço amostral $S := \{a, b, c, \dots z\}$
- uma v.a X tal que $X(a) = 1$, $X(b) = 2$, ..., $X(z) = 26$
- uma v.a Y tal que $Y(vogal) = 1$ e $Y(consoante) = 2$

veja agora algumas probabilidades:

- $\Pr[X > 13] = \Pr[\{n, o, \dots z\}]$
- $\Pr[Y = 1] = \Pr[\{a, e, i, o, u\}]$

B.1.2 Manipulando V.A

AS v.a podem ser manipuladas algebricamente. Por exemplo, uma multiplicação por escalar, dado uma v.a X podemos definir um v.a Y como:

$$Y = 2X$$

B.1.3 V.A Indicadora

Uma variável aleatória dita indicadora é aquela que assume apenas dois valores: 0 ou 1, ou seja $\Pr[X = 0 \vee X = 1] = 1$.

B.2 Função de Distribuição de Probabilidade

Definição B.2 (Funções de Distribuição de Probabilidade). Sejam X uma v.a e x um de seus possíveis valores, temos as seguintes funções

- a função de probabilidade $f_X(x) = \Pr[X = x]$
- a função cumulativa $F_X(x) = \Pr[X \leq x]$

de forma que satisfaça as restrições

- $0 \leq f_X(x) \leq 1, \forall x \in O_x$
- $\sum_{x \in O_x} f_X(x) = 1$

onde O_x é o conjunto de valores que v.a X pode assumir.

Definição B.3 (Distribuição de Bernoulli). Uma v.a X possui distribuição de Bernoulli se ela é uma v.a indicadora:

$$f_X(1) = \Pr[X = 1] = p$$
$$f_X(0) = \Pr[X = 0] = 1 - p$$

denotamos $X \sim Bernoulli(p)$ onde X é uma v.a que possui distribuição de Bernoulli com parâmetro p.

Definição B.4 (Distribuição Binomial). Considere uma sequência *i.i.d* de Bernoulli:

$$(X_1, X_2, \ldots, X_n)$$

onde

- $X_i \sim Bernoulli(p), \forall i \in [1..n]$
- Espaço amostral: todas as sequências binárias de tamanho n

Seja Z a soma destas *v.a.s*

$$Z = \sum_{i=1}^{n} X_i$$

note que $Z \in [0..n]$.

Dizemos que Z possui distribuição binomial com parâmetros n e p, denotamos

$$Z \sim Bin(n, p)$$

Desta maneira podemos definir a probabilidade da soma ser igual a i como

$$f_Z(i) = \Pr[Z = i] = \binom{n}{i} p^i (1-p)^{n-i}$$

Exemplo B.1. Podemos calcular a probabilidade de o número de caras ao jogar uma moeda 20 vezes, seja $X_i \in [1..n]$ uma *v.a* indicadora de uma sequência *i.i.d* de n *v.a.s*. Se $X_i = 1$ então a jogada i resultou em cara, caso contrário resultou em coroa. Podemos calcular:

$$Z \sim Bin(n, p)$$

onde $p = 0.5$; vamos calcular

$$\begin{aligned} f_Z(10) &= \binom{20}{10} 0.5^{10} (1-0.5)^{20-10} \\ &= \binom{20}{10} 0.5^{20} \\ &= \frac{20!}{10!(20-10)!} 0.5^{20} \end{aligned}$$

Definição B.5 (Distribuição Geométrica). Considere uma sequência de v.a *i.i.d* que possuem a distribuição de Bernoulli

$$X_1, X_2, \ldots$$

Seja Z o menor valor tal que $X_Z = 1$, ou seja a primeira ocorrência do valor 1 na sequência:

$$Z = \min\{i \mid X_i = 1\}$$

dessa forma Z possui distribuição geométrica com parâmetro p, denotamos

$$Z \sim Geo(p)$$

Desta maneira podemos definir a probabilidade da primeira ocorrência acontecer na v.a na posição i como

$$f_Z(i) = \Pr[Z = i] = p(1-p)^{i-1}$$

Por exemplo podemos calcular o número de vezes de moedas a serem jogadas até a primeira cara. Vamos calcular a probabilidade de a cara aparecer pela primeira vez na 2 jogada:

$$f_Z(2) = (1-0.5)^{2-1} 0.5 = 0.5^2 = 0.25$$

Definição B.6 (Independência). Seja X e Y duas v.a, elas são independentes se e somente se para todo par i, j do espaço amostral temos que

$$\Pr[X = i \wedge Y = j] = \Pr[X = i] \Pr[Y = j]$$

Definição B.7 (Sequência de V.A). Considere uma sequência S de v.a:

$$S = (X_1, X_2, \ldots, X_n)$$

Dizemos que uma sequência S de v.a é *i.i.d* (**independente e identicamente distribuída**) se:

- as v.a.s são independentes entre si
- as v.a.s possuem a mesma função de distribuição
- as v.a.s são distintas

Por exemplo uma sequência *i.i.d*: uma sequência n jogadas de um dado, onde X_i é o valor observado na i-ésima jogada.

Definição B.8 (Esperança). A esperança (ou valor esperado) é trata-se da média ponderada dos valores que X pode assumir:

$$\mu_X = \mathbb{E}[X] = \sum_{i \in O_X} i f_{X(i)}$$

onde O_X trata-se dos valores que X pode assumir.

Exemplo B.2. Veja algumas esperanças de acordo com a sua distribuição:

- $X \sim Bernoulli(p)$

$$\mathbb{E}[X] = \sum_{i \in O_X} i f_{X(i)} = 0(1-p) + 1p = p$$

- $X \sim Bin(n, p)$

$$\begin{aligned} \mathbb{E}[X] = \sum_{i \in O_X} i f_{X(i)} &= \sum_{i \in O_X} \binom{n}{i} p^i (1-p)^{n-i} i \\ &= \sum_{i \in n} \binom{n}{i} p^i (1-p)^{n-i} i \end{aligned}$$

e pelo teorema binomial temos que

$$\sum_{i \in n} \binom{n}{i} p^i (1-p)^{n-i} i = np$$

- $X \sim Geo(p)$

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i \in O_X} ip(1-p)^{i-1} \\ &= p \sum_{i \in n} i(1-p)^{i-1} \\ &= \frac{1}{p} \end{aligned}$$

Teorema B.1 (**Linearidade da Esperança**). Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias e $a_1, a_2, \ldots, a_n$ constantes. A linearidade da esperança estabelece que para qualquer combinação linear destas variáveis e constantes, temos:

$$\mathbb{E}[a_1 X_1 + a_2 X_2 + \ldots + a_n X_n] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2] + \ldots + a_n \mathbb{E}[X_n]$$

B.3 Função de V.A

Seja $g : \mathbb{Z} \to \mathbb{R}$ e X uma *v.a*, podemos aplicar: $g(X)$.

Podemos definir a esperança de uma função de *v.a* como:

$$\mathbb{E}[g(X)] = \sum_{i \in O_x} g(i) f_X(i)$$

Definição B.9 (**Variância**). A variância é a medida de dispersão ao redor da média, a dispersão é medida como:

$$g(X) = (X - \mu_X)^2$$

ou seja é o quadrado da diferença com o valor esperado.

Assim, podemos definir a variância como:

$$\mathrm{Var}[X] = \mathbb{E}[g(x)] = \mathbb{E}\big[(X - \mu_X)^2\big]$$

Exemplo B.3. Veja algumas variâncias de acordo com a sua distribuição:

- $X \sim Bernoulli(p)$

$$\begin{aligned}\mathrm{Var}[X] &= g(0)(1-p) + g(1)p \\ &= (0-p)^2(1-p) + (1-p)^2 p \\ &= p^2 - p^3 + (p^2 - 2p + 1)p \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

- $X \sim Bin(n, p)$

$$\mathrm{Var}[X] = np(1-p)$$

- $X \sim Geo(p)$

$$\mathrm{Var}[X] = \frac{1-p}{p^2}$$

Definição B.10 (desvio padrão). O desvio padrão é a raiz quadrada da variância:

$$\sigma_X = \sqrt{\mathrm{Var}[X]}$$

B.4 Espaço Amostral Contínuo

O **espaço amostral contínuo** é um espaço amostral não enumerável. A questão é *como associar a probabilidade* de cada ponto do espaço amostral? A solução é **dar probabilidade a subconjuntos** do espaço amostral: pedaços contíguos do espaço amostral tem uma "densidade" de probabilidade.

Definição B.11 (Função de Densidade). Seja A um evento qualquer de S (um evento é um subconjunto de S).

Dizemos que $f(x)$ é uma função de densidade se e somente se

$$\Pr[A] = \int_A f(x)\,dx$$

note que $\Pr[A]$ é a área da função de densidade dentro do espaço definido por A. Temos as mesmas restrições que antes:

$$0 \leq \int_A f(x)\,dx \leq 1 \quad \text{e} \quad \int_S f(x)\,dx = 1$$

Exemplo B.4. Podemos representar a distribuição uniforme em um espaço amostral contínuo $S =$

$[a, b]$ para a, b constantes e $A = [a', b']$ para $a' \geq a$ e $b' \leq b$, da seguinte forma

$$f(x) = \frac{1}{b-a}, \text{ para } x \in [a, b] \quad \text{e} \quad \Pr[A] = \int_{a'}^{b'} f(x)\, dx = \frac{b'-a'}{b-a}$$

.

Podemos definir uma v.a X contínua é como:

$$X : S \to \mathbb{R}$$

todos os conceitos sobre v.a.s discretas e contínuas são equivalentes basta trocar os somatórios por integrais, usando CDF[9] quando necessário.

[9] Função de Distribuição de Probabilidade Cumulativa

C Limitantes e Desigualdades Probabilísticas

Nesta seção vamos definir alguns limitantes e desigualdades probabilísticas,

C.1 Limitantes para Probabilidade

Calcular probabilidades de eventos pode ser difícil, analiticamente intratável e/ou computacionalmente intratável. Portanto usar limitantes, inferior ou superior pode ser mais fácil:

$$\Pr[A] \leq U_A := \text{Limitante superior;}$$
$$\Pr[A] \geq L_A := \text{Limitante inferior;}$$

Definição C.1 (Cauda e Cabeça). Seja X uma v.a aleatória com $\mu = \mathbb{E}[X]$, definimos:

- **Cauda**: valores de X bem maiores que μ;
- **Cabeça**: valores de X bem menores que μ.

veja o exemplo na Figura 6:

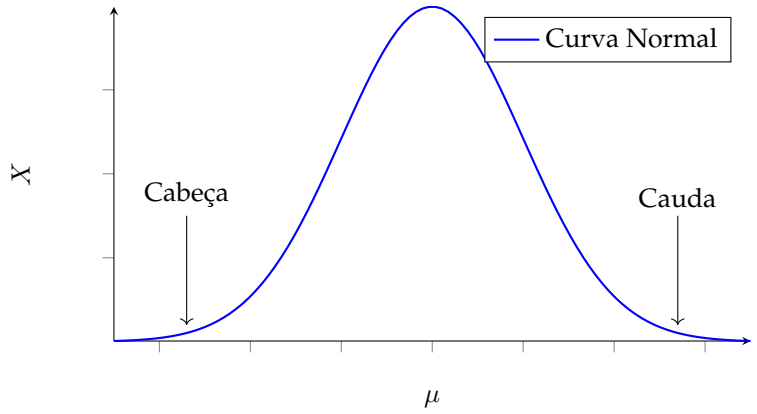


Figura 6: Cabeça e Cauda

Exemplo C.1. Vamos contar quantas vezes o resultado de uma jogada de dado é um número primo, dado um número específico de jogadas:

- Jogar 50 vezes um dado honesto com 10 faces ($[1..10]$);
- $X_i :=$ variável indicadora de que a i-ésima jogada foi um número primo;
- $Z = \sum_{i \in 50} X_i$

$$X_i = \begin{cases} 1 & \text{se o número é primo} \\ 0 & \text{caso contrário} \end{cases}$$

logo

$$Z \sim Bin\left(50, \frac{2}{5}\right)$$

A probabilidade de $Z \geq 40$ é

$$\Pr[Z \geq 40] = \sum_{i=40}^{50} \binom{50}{i} \left(\frac{2}{5}\right)^{i} \left(\frac{3}{5}\right)^{50-i}$$

portanto é muito difícil calcular!

Teorema C.1 (Desigualdade de Markov). A *Desigualdade de Markov* é um limitante que expressa a relação entre o valor esperado e a probabilidade.

Para qualquer v.a X não negativa e constante $a > 0$, temos que

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

note que a probabilidade acima só faz sentido quando a está na cauda da distribuição.

Demonstração. Seja $I(X \geq a) :=$ v.a indicadora do evento $X \geq A$, logo

$$aI(X \geq a) \leq X$$

e então

$$E\left[aI(X \geq a)\right] \leq \mathbb{E}[X]$$

e lembre-se que a esperança de uma variável indicadora é a própria função de distribuição de probabilidade quando o v.a vale 1, ou seja

$$E\left[aI(X \geq a)\right] \leq \mathbb{E}[X]$$
$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

□

Exemplo C.2. Dado $Z \sim Bin\left(50, \frac{2}{5}\right)$, podemos calcular o limitante para $\Pr[Z \geq 40]$. Sabemos que

$$\mathbb{E}[Z] = \frac{2}{5} \cdot 50 = 20$$

logo

$$\begin{aligned}\Pr[Z \geq 40] &\leq \frac{20}{40} \\ &= \frac{1}{2}\end{aligned}$$

Teorema C.2 (Desigualdade de Chebyshev). A *Desigualdade de Chebyshev* é um limitante que expressa a relação entre o valor esperado, variância e probabilidade. Ela é mais precisa que a Desigualdade de Markov.

Para qualquer v.a X com o valor esperado μ e variância σ^2 e qualquer $k > 0$ temos que

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Demonstração. Seja $Y = (X - \mu)^2$ e $a = (k\sigma)^2$, aplicando a Desigualdade de Markov temos que

$$\begin{aligned}\Pr[Y \geq a] &\leq \frac{\mathbb{E}[Y]}{a} \\ \Pr\left[(X - \mu)^2 \geq (k\sigma)^2\right] &\leq \frac{\mathbb{E}\left[(X - \mu)^2\right]}{(k\sigma)^2} \\ &\leq \frac{\mathbb{E}\left[\sigma^2\right]}{(k\sigma)^2} \\ &\leq \frac{1}{k^2}\end{aligned}$$

agora basta notar que

$$\Pr[|X - \mu| \geq k\sigma] = \Pr\left[(X - \mu)^2 \geq (k\sigma)^2\right]$$

□

Exemplo C.3. Um caso interessante da Desigualdade de Chebyshev é quando $k = \sqrt{2}$, onde temos:

$$\Pr[|X - \mu| \geq 1.41\sigma] \leq \frac{1}{2}$$

portanto temos que

$$P\left[X \notin [\mu - 1.41\sigma, \mu + 1.41\sigma]\right] \leq \frac{1}{2}$$

Exemplo C.4. Dado $Z \sim Bin\left(50, \frac{2}{5}\right)$, podemos calcular o limitante para $\Pr[Z \geq 40]$. Sabemos que

$$\mu = \mathbb{E}[X] = \frac{2}{5} \cdot 50 = 20 \quad \text{e} \quad \sigma^2 = 50\frac{2}{5}\frac{3}{5} = 12$$

e note que

$$\{Z \geq 40\} = \{|Z - \mu| \geq 20\}$$

portanto

$$\begin{aligned} k\sigma &= 20 \\ k &= \frac{20}{\sqrt{12}} \end{aligned}$$

logo

$$\begin{aligned} \Pr[Z \geq 40] \leq \Pr[|N - \mu| \geq 20] &\leq \frac{1}{\left(\frac{20}{\sqrt{12}}\right)^2} \\ &= \frac{12}{400} \\ &= \frac{3}{100} \\ &= 0.03 \end{aligned}$$

Teorema C.3 (Desigualdade de Chernoff). A **Desigualdade de Chernoff** é um limitante superior para soma da v.a.s independentes. Ela é muito mais precisa que a Desigualdade de Chebyshev.

Seja $Y_i \sim Bernoulli(p), X = \sum_{Y_i} \mu = \mathbb{E}[X] = np$ e qualquer $\delta > 0$, temos

- Probabilidade da cauda:

$$\Pr[X \geq (1+\delta)\mu] \geq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

- Probabilidade da cabeça:

$$\Pr[X \leq (1-\delta)\mu] \geq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

Exemplo C.5. Dado $Z \sim Bin\left(50, \frac{2}{5}\right)$, podemos calcular o limitante para $\Pr[Z \geq 40]$ aplicando Desigualdade de Chernoff. Sabemos que

$$\mu = \mathbb{E}[X] = \frac{2}{5} \cdot 50 = 20$$

e

$$\begin{aligned}(1+\delta)\mu &= 40\\ \delta &= 2-1\\ &= 1\end{aligned}$$

logo

$$\Pr[X \geq (1+1)\mu] \geq \left(\frac{e^1}{(1+1)^{1+1}}\right)^{20} = \frac{e^{20}}{2^{40}} = 0.00044$$

Definição C.2 (**Evento com Alta Probabilidade**). Seja n um parâmetro de um modelo probabilístico (por ex. n° de jogadas de um dado) e $A(n)$ um evento no respectivo espaço amostral, definimos que $A(n)$ ocorre *alta probabilidade* quando

$$\Pr[A(n)] \geq 1 - \frac{1}{n^\alpha}, \text{ para algum } \alpha > 1$$

note que

$$\lim_{n\to\infty} \Pr[A(n)] = 1 \text{ e } \lim_{n\to\infty} \Pr\left[\overline{A}(n)\right] = 0$$

Exemplo C.6. Seja Y_i uma v.a indicadora se uma jogada de uma moeda honestas e cara, e seja a sequência i.i.d

$$X = Y_1 + \cdots Y_n$$

ou seja qual o valor de λ tal que

$$\Pr[X \geq \mu + \lambda] \leq \frac{1}{n^\alpha}$$

.

Podemos usar uma variação da desigualdade de Chernoff:

$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\delta^2 \frac{\mu}{3}}$$

então

$$\begin{aligned}(1+\delta)\mu &= \mu + \lambda\\ 1+\delta &= 1 + \frac{\lambda}{\mu}\\ \delta &= \frac{\lambda}{\mu}\end{aligned}$$

portanto

$$\Pr[X \geq \mu + \lambda] \leq e^{-\left(\frac{\lambda}{\mu}\right)^2 \frac{\mu}{3}}$$

e como

$$e^{-\left(\frac{\lambda}{\mu}\right)^2 \frac{\mu}{3}} = e^{-\frac{2\lambda^2}{3n}}$$

temos que

$$\lambda = \sqrt{\frac{3}{2} n \ln n}$$

Então se $\lambda = \sqrt{\frac{3}{2} n \ln n}$ então $\Pr[X \geq \mu + \lambda] \leq \frac{1}{n}$; e se $n = 1000$ temos que $\mu 500$ e

$$\lambda = \sqrt{\frac{3}{2} n \ln n} = \sqrt{1500 \ln 1000} = 101.8$$

logo

$$\Pr[X \geq 500 + 102] = \Pr[X \geq 602] \leq 0.001$$

ou seja observar 602 caras ou mais é bastante raro ao jogar uma moeda honesta 1000 vezes.

C.2 Limitante da União

O *Limitante da União* é um limitante útil quando se tem muitos eventos que não necessariamente são mutuamente exclusivoS.

Sejam A e B dois eventos em um espaço amostral S; temos que

$$\begin{aligned} \Pr[A \vee B] &= \Pr[A] + \Pr[B] - \Pr[A \wedge B] \\ &\leq \Pr[A] + \Pr[B] \end{aligned}$$

Seja A_i uma sequência de eventos num espaço amostral para $i \in [1..n]$, temos que

$$\begin{aligned} P\left[\bigcup_i A_i\right] &= \sum_{i=1}^{n} \Pr[A_i] - P\left[\bigcap_i A_i\right] \\ &\leq \sum_{i=1}^{n} \Pr[A_i] \end{aligned}$$

Se A_i forem identicamente distribuídos (possuem a mesma probabilidade) então

$$\sum_{i=1}^{n} \Pr[A_i] = n \Pr[A_1]$$

caso contrário

$$\sum_{i=1}^{n} \Pr[A_i] \leq n \max \{i \in [1..n] \mid \Pr[A_i]\}$$

Exemplo C.7. Seja $S = (X_1, X_2, X_3)$ uma sequência de v.a.s *i.i.d* cujo $X_i \in [1..6]$ (jogar um dado honesto 3 vezes).

Qual a probabilidade de algum $X_i = 6$?

$$\begin{aligned} P\left[\bigvee_{i=1}^{3} X_i = 6\right] &\leq \sum_{i=1}^{3} \Pr[X_i = 6] \\ &= \frac{3}{6} \\ &= \frac{1}{2} \end{aligned}$$

Qual é a probabilidade exata?

$$\begin{aligned} P\left[\bigvee_{i=1}^{3} X_i = 6\right] &= 1 - P\left[\bigwedge_{i=1}^{3} X_i \neq 6\right] \\ &= 1 - \prod_{i=1}^{3} \Pr[X_i \neq 6] \\ &= 1 - \left(\frac{5}{6}\right)^3 \\ &\approxeq 1 - 0.58 \\ &= 0.42 \end{aligned}$$

ou seja o limitante forneceu uma boa aproximação.

Exemplo C.8. Seja $S = (X_1, \ldots, X_{10})$ uma sequência de v.a.s *i.i.d* cujo $X_i \in [1..6]$ (jogar um dado honesto 10 vezes).

Qual a probabilidade de algum $X_i = 6$?

$$P\left[\bigvee_{i=1}^{10} X_i = 6\right] \leq 10\frac{1}{6} = \frac{5}{3}$$

o que não é nada útil.

Ou seja, o *Limitante de União* é útil quando as parcelas possuem probabilidades pequenas e/ou quando número de parcelas é pequeno.

D Lei dos Grandes Números, Erro e Confiança

No campo da teoria das probabilidades e estatísticas, diversos conceitos fundamentais moldam nossa compreensão sobre a aleatoriedade e a incerteza inerentes aos fenômenos naturais e processos estocásticos. Nesta seção, exploraremos quatro tópicos cruciais que desempenham um papel central na inferência estatística e na interpretação de resultados probabilísticos.

D.1 Fração Relativa dos Resultados

Para compreendermos a Lei dos Grandes Números, é fundamental explorar o conceito de fração relativa dos resultados. Em uma sequência de n lançamentos de um dado, consideramos as seguintes variáveis:

$$Y_i := \text{variável aleatória indicadora de que o resultado é o número 1}$$
$$N_1(n) := \text{número de vezes que o resultado é o número 1, ou seja, } N_1(n) = \sum_{i=1}^{n} Y_i$$
$$F_1(n) := \text{fração de vezes que o resultado é 1, ou seja, } F_1(n) = \frac{N_1(n)}{n}$$

A questão que surge é: qual é o valor de $F(10), F(100)$ e $F(1000)$? Vale ressaltar que $F(n)$ é uma variável aleatória.

D.2 Lei dos Grandes Números

A **Lei dos Grandes Números** afirma que $F_1(n)$ converge para sua respectiva probabilidade à medida que n se aproxima do infinito. Em outras palavras, é a conexão direta entre a teoria e a prática.

Intuitivamente, a Lei dos Grandes Números nos assegura que, ao lidarmos com uma quantidade "grande" de resultados aleatórios, a fração de ocorrências de um evento específico converge para a probabilidade desse evento. Em resumo, a probabilidade se revela como uma ferramenta prática e confiável na análise de eventos aleatórios em larga escala.

Vamos apresentar algumas defninições e teoremas necessários para a compreensão da Lei.

Definição D.1 (Média Amostral). Dado X_i para $i \in [1..n]$, uma sequência de variáveis aleatórias i.i.d (*independentes e identicamente distribuídas*) temos que a **média amostral** é

$$\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} X_i := \text{ média aritmética dos } n \text{ valores observados}$$

Teorema D.1. A esperança da média amostral $\bar{M}_n$ de uma sequência i.i.d de n variáveis aleatórias X_i é μ, onde $\mu = \mathbb{E}[X_i]$.

Demonstração.

$$\begin{aligned}
\mathbb{E}\big[\bar{M}_n\big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\
&= \frac{1}{n} n\mu \\
&= \mu.
\end{aligned}$$

□

Teorema D.2. A variância da média amostral $\bar{M}_n$ de uma sequência i.i.d de n variáveis aleatórias X_i é $\frac{\sigma^2}{n}$, onde $\sigma^2 = \mathrm{Var}[X_i]$.

Demonstração.

$$\begin{aligned}
\mathrm{Var}\big[\bar{M}_n\big] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=}^{n} X_i\right] \\
&= \left(\frac{1}{n}\sum_{i=}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right)^2 \\
&= \left(\frac{1}{n}\sum_{i=}^{n} X_i - \frac{1}{n}\mathbb{E}[sum_{i=1}^{n} X_i]\right)^2 \\
&= \left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)^2 \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

□

Teorema D.3. Seja $\bar{M}_n$ a média amostral de uma sequência de n variáveis aleatórias X_i e $\epsilon > 0$. Se $\mu = \mathbb{E}[X_i]$ e $\sigma^2 = \mathrm{Var}[X_i]$ são finitos então

$$\Pr\big[|\bar{M}_n - \mu| \geq \epsilon\big] \leq \frac{\sigma^2}{\epsilon^2 n}.$$

Demonstração. Basta usar a desigualdade de *Chebyshev*, isto é

$$\Pr\big[|\bar{M}_n - \mu| \geq k\sigma_{\bar{M}_n}\big] \leq \frac{1}{k^2}$$

podemos fazer $k\sigma_{\bar{M}_n} = \epsilon$ logo

$$\begin{aligned} k &= \frac{\epsilon}{\sqrt{\mathrm{Var}\big[\bar{M}_n\big]}} \\ &= \frac{\epsilon}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{\epsilon\sqrt{n}}{\sigma} \end{aligned}$$

e portanto

$$\Pr\big[|\bar{M}_n - \mu| \geq \epsilon\big] \leq \frac{1}{k^2} = \frac{\sigma^2}{\epsilon^2 n}.$$

□

Teorema D.4 (Lei Fraca dos Grandes Números). Dado a média amostral $\bar{M}_n$ de uma sequência de n variáveis aleatórias X_i, se $\mu = \mathbb{E}[X_i]$ é finito, para qualquer $\epsilon > 0$ temos que

$$\lim_{n\to\infty} \Pr\big[|\bar{M}_n - \mu| < \epsilon\big] = 1$$

Observação. Este teorema é frequentemente referido como "convergência em probabilidade" e reflete que a probabilidade de M_n estar a uma distância ϵ da média tende a 1, para qualquer ϵ positivo. Isso significa que, à medida que o número de observações n aumenta, a média amostral M_n se aproxima cada vez mais a probabilidade do evento, com alta probabilidade, mesmo para valores pequenos de ϵ, como $\epsilon = 10^{-10}$.

Demonstração. Basta usar utilizar o resultado do Teorema D.3, isto é

$$\Pr\big[|\bar{M}_n - \mu| < \epsilon\big] = 1 - \Pr\big[\bar{M}_n - \mu| \geq \epsilon\big] \geq 1 - \frac{\sigma^2}{\epsilon^2 n}$$

e é trivial que

$$\lim_{n\to\infty} \left(1 - \frac{\sigma^2}{\epsilon^2 n}\right) = 1.$$

□

Teorema D.5 (Lei Forte dos Grandes Números). Dado a média amostral $\bar{M}_n$ de uma sequência de n variáveis aleatórias X_i, se $\mu = \mathbb{E}[X_i]$ é finito temos que

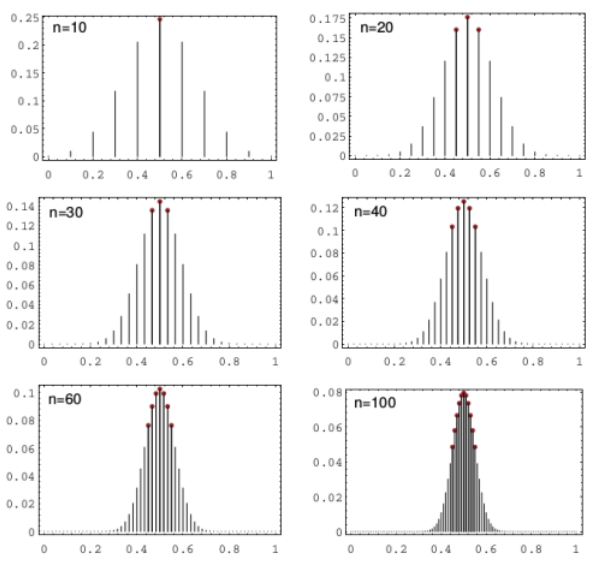
$$\lim_{n\to\infty} \Pr\big[\bar{M}_n = \mu\big] = 1.$$

Observação. Este teorema é comumente conhecido como "convergência quase certa", representando um resultado significativamente mais robusto do que a Lei Fraca dos Grandes Números. Neste contexto, $\bar{M}_n$ realmente converge para a média esperada.

Exemplo D.1. Dado M_n a média amostral de n variáveis aleatórias indicadores X_i cujo tais variáveis indicam se o resultado de uma jogada i de uma moeda honesta resultou em cara. Ou seja, temos

$$\mathbb{E}\big[\bar{M}_n\big] = \frac{1}{2} \quad \text{e} \quad \mathrm{Var}\big[\bar{M}_n\big] = \frac{\sigma^2}{n} = \frac{1}{4n}$$

observe a curva normal com base em $\bar{M}_n$ quando o n cresce



note que $\bar{M}_n$ fica mais centrada (converge) quando n cresce.

D.3 Erro e Confiança na Lei dos Grandes Números

Seja M_n uma variável aleatória representando a média amostral de uma distribuição. Podemos utilizar a desigualdade de Chebyshev para calcular a probabilidade de M_n estar dentro de um intervalo específico. Para uma precisão ϵ e confiança β, a relação é expressa por:

$$\Pr\big[\bar{M}_n \in [\mu - \epsilon, \mu + \epsilon]\big] \geq \beta$$

Essa expressão indica que, com uma probabilidade de pelo menos β, a média amostral $\bar{M}_n$ estará contida no intervalo definido pela o valor esperado do evento μ com uma precisão de ϵ.

Dada a precisão ϵ, a confiança β, bem como os parâmetros μ e σ^2 da distribuição, podemos determinar o tamanho necessário da amostra, n, para garantir essa relação. A expressão fundamental é:

$$n = \frac{\sigma^2}{\epsilon^2(1-\beta)}$$

A relação entre confiança e precisão é notável nesse contexto. A confiança tem uma influência linear no tamanho de n, enquanto a precisão exerce uma influência quadrática. Aumentar a precisão demanda um aumento mais significativo no tamanho da amostra em comparação ao aumento da confiança. Essa consideração é vital ao planejar experimentos e estudos estatísticos.

Exemplo D.2. Suponha que tenhamos uma moeda enviesada com uma probabilidade de cara sendo de 45%. Para verificar se a moeda é realmente enviesada, quantas vezes precisamos lançá-la?

Assumindo uma precisão desejada de $\epsilon = 0.01$ e uma confiança de $\beta = 0.95$, definimos $\mu = 0.45$ e $\sigma^2 = 0.45 \cdot 0.55$. Portanto, queremos garantir que:

$$\Pr[M_n \in [0.44, 0.46]] > 1 - \frac{0.45 \cdot 0.55}{(0.01)^2 n} = 0.95$$

Isso implica em:

$$n = \frac{0.2475}{0.0001 \cdot 0.05} = 49500$$

Concluímos que, para determinar com uma precisão de 1% e uma confiança de 95% se a moeda é enviesada, precisamos lançá-la 49500 vezes e verificar se a média amostral está dentro da faixa especificada.

E Geração de Amostras Aleatórias

Nesta seção, exploraremos a questão crucial da geração de amostras aleatórias, com destaque para um algoritmo renomado nesse domínio: o Método Alias, particularmente eficaz na geração de amostras não uniformes. Luc Devroye aborda de maneira abrangente a geração de variáveis aleatórias não uniformes em seu livro seminal [16].

A necessidade de gerar amostras aleatórias é abrangente em diversos contextos, como a implementação de métodos de Monte Carlo, simulação de sistemas aleatórios, desenvolvimento de jogos, criação de algoritmos e outras aplicações. A compreensão de métodos eficientes para essa geração é essencial em uma ampla gama de campos, tornando o Método Alias uma ferramenta significativa nesse contexto.

E.1 Geração de um dado aleatório

Considere o problema de gerar uma amostra de um lançamento de um dado honesto[10] de k faces, seja D a variável aleatória que denota o valor da face.

A abordagem para a geração de uma amostra é a seguinte: ao considerarmos um intervalo contínuo de 0 a 1, dividimos esse intervalo em k faixas. Em seguida, geramos uma variável uniforme $u \sim U(0, 1)$ (consideremos que a geração é de tempo constante) e determinamos o intervalo que contém o valor gerado por u, retornando a correspondente categoria; veja a ilustração da abordagem em Algoritmo 1.

```
1 gera_amostra_dado(int k) {
2     u = unif(0, 1);
3     i = to_int(k * u) + 1;  // to_int(r): retorna parte inteira de r
4     return i;
5 }
```

Algoritmo 1: Gerando amostra de um dado com k faces

se considerarmos que `to_int` é de tempo constante então conseguimos **gerar uma amostra de um dado aleatório de k faces em tempo constante!**

E.2 Método Alias

Considerando agora o problema de gerar amostras de um dado enviesado de k faces. Seja D a variável aleatória, onde a probabilidade é definida como

$$\Pr[D = i] = \frac{w_i}{W} \quad \text{e} \quad W = \sum_{i=1}^{k} w_i$$

Vamos agora ilustrar o Método Alias. A ideia por trás do método é utilizar a *memória* como artifício para ganharmos em eficiência, os passos são:

- **Pré-processamento**:

[10] Um dado honesto é o dado que suas faces possuem a mesma probabilidade.

1. Alocamos um vetor de tamanho W
2. Preenchemos w_i posições com o valor i

- **Geração da Amostra**:

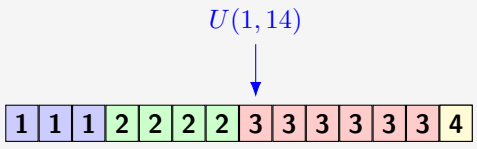
1. Escolhemos um inteiro uniforme em $[1..W]$
2. Acessamos o vetor para retornar a face

Portanto o pré-processamento tem complexidade $O(W)$ e a geração da amostra tem complexidade $O(1)$.

Exemplo E.1. Considerando um dado enviesado com 4 faces onde

$$w = \begin{pmatrix} 3 & 4 & 6 & 1 \end{pmatrix},$$
$$W = 3 + 4 + 6 + 1 = 14$$

veja a ilustração do vetor abaixo:



no exemplo acima a uniforme gerou a amostra 3.