UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS MATEMÁTICAS E DA NATUREZA
INSTITUTO DE MATEMÁTICA

Universidade Federal
do Rio de Janeiro

Alguns Aspectos Teóricos dos Métodos de Monte Carlo

Henrique Andrade de Aquino

Rio de Janeiro
2019

Alguns Aspectos Teóricos dos Métodos de Monte Carlo

por

Henrique Andrade de Aquino

Dissertação de Mestrado apresentada ao Programa de Pós-graduação do Instituto de Matemática, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Orientador: Heudson Tosta Mirandola

Aprovada por:

Heudson Tosta Mirandola, IM-UFRJ

Isaia Nisoli, IM-UFRJ

João Antonio Recio da Paixão, IM-UFRJ

Guilherme Ost de Aguiar, IM-UFRJ (suplente)

Agradecimentos

Agradeço primeiro a Deus, criador de todas as coisas, pela vida, por minha família e pelo mundo. Agradeço a Ele também pelo desejo de beleza que me fez trilhar um caminho na matemática.

Agradeço aos meus pais por tudo o que já fizeram em minha vida. Eles me formaram como pessoa, fizeram todo o possível para que eu pudesse chegar até aqui, e principalmente, me criaram em meio ao amor. São também minhas primeiras referências profissionais e humanas, que me inspiraram o gosto pela ciência. Agradeço também aos meus irmãos e a toda a família, por sua importante companhia ao longo do mestrado.

Agradeço ao professor Heudson, por sua orientação. Sempre foi paciente, presente e preocupado em ajudar. Levarei seu exemplo e guardarei seus conselhos, não só na academia, mas como alguém que não desiste diante dos desafios e busca sempre melhorar a si mesmo e aos outros. Agradeço também a toda a comunidade acadêmica da UFRJ e todos os professores que passaram por minha vida.

Agradeço aos meus amigos, que me ajudaram muito a olhar para os estudos e para a vida com seriedade e esperança.

Resumo

Este trabalho apresenta alguns métodos de Monte Carlo de relevância histórica ou prática e estuda as suas propriedades, buscando justificar teoricamente o seu funcionamento. Dentre os métodos, o texto busca tratar a influência da dimensionalidade dos problemas tratados na variância dos estimadores.

Palavras-chave: Monte Carlo; Monte Carlo por Cadeias de Markov; MCMC; Simulação Monte Carlo; Amostrador de Gibbs; Algoritmo Metropolis.

Abstract

This essay presents some Monte Carlo methods, relevant in history or practice, and studies their properties, seeking to justify theoretically their operation. Among these methods, the text deals with the influence of the problems dimensionality to variance in estimation.

Keywords: Monte Carlo; Markov Chain Monte Carlo; MCMC; Monte Carlo Simulation; Gibbs Sampler; Metropolis Algorithm.

Conteúdo

Introdução

0.1 Introdução

0.1.1 Quando os Métodos Numéricos Não Dão Conta

Nessa subseção, abordaremos o uso de métodos de simulação de dados para o cálculo de integrais. A pergunta natural a se fazer aqui é bastante simples:

Por que ir além do uso de métodos numéricos?

Consideramos uma função $f(x)$ de classe C^1 definida sobre o hipercubo fechado d-dimensional, $[0,1]^d = [0,1] \times \ldots \times [0,1]$ de $\mathbb{R}^d$. O objetivo é calcular a integral

$$\mu = \int_{[0,1]^d} f(x)dx, \tag{1}$$

levando-se em consideração a precisão e complexidade do método considerado.

Na abordagem por somas de Riemann, considera-se uma partição $P = \{a_0 = 0 < a_1 < \ldots < a_n = 1\}$ com $a_{i+1} - a_i = \frac{1}{n}$ e toma-se os subcubos $R_I = [a_{i_1}, a_{i_1+1}] \times \ldots \times [a_{i_d}, a_{i_d+1}]$, com multi-índices $I = (i_1, \ldots, i_d)$ satisfazendo $0 \leq i_1, \ldots, i_d < n$. Em cada subcubo R_I, considera-se o volume $\Delta_I = \text{vol}\,(R_I) = \frac{1}{n^d}$ e toma-se um ponto $x_I \in R_I$ qualquer, podendo ser escolhido aleatoriamente. A integral μ é dada pelo limite $\mu = \lim_{n\to\infty} \hat{\mu}(n)$, sendo $\hat{\mu}(n) = \frac{1}{n^d} \sum_I f(x_I)$.

Primeiro, vamos analisar a precisão do estimador $\hat{\mu}(n)$. Do teorema do valor intermediário, segue-se que $\mu_I = \int_{R_I} f(x)dx = f(x_I^*)\Delta_I$, para algum

$x_I^* \in R_I$. E, da desigualdade do valor médio,

$$|\mu_I - f(x_I)\Delta_I| = |f(x_I^*) - f(x_I)|\frac{1}{n^d} \leq M\text{diam}(R_I)\frac{1}{n^d} = M\frac{\sqrt{d}}{n}\frac{1}{n^d},$$

onde $M = \max_{x\in[0,1]^d} \|\nabla f(x)\|$ é a norma máxima do vetor gradiente de $f(x)$ e $\text{diam}(R_I) = \max\{\|x_I - x_I'\| \mid x_I, x_I' \in R_I\} = \frac{\sqrt{d}}{n}$ é o diâmetro de R_I. Segue-se, da desigualdade triangular, que

$$|\mu - \hat{\mu}(n)| \leq \sum_I |\mu_I - f(x_I)\Delta_I| \leq \sum_I \frac{M\sqrt{d}}{n^{d+1}} = \frac{M\sqrt{d}}{n}.$$

Donde, $\hat{\mu}(n) = \mu + O(\frac{1}{n})$, quando $n \to \infty$.

Observe que $\hat{\mu}(n)$ tem complexidade $O(n^d)$, visto que para se realizar a soma $\sum_I f(x_I)$ deve-se fazer pelo menos n^d operações. Para dimensões altas, torna-se completamente inviável calcular o estimador $\hat{\mu}(n)$. Sistemas com alta dimensão nos estados podem ocorrer com bastante frequência. Considere, por exemplo, o volume de tráfego diário de uma certa rodovia, no qual as informações são dadas a cada minuto do dia. Neste caso, as estatísticas serão dadas por funções sobre vetores com 1.440 entradas.

No livro de D. Mackay [12], há uma ilustração bastante interessante do quanto a alta complexidade pode inviabilizar o uso de métodos numéricos. Suponhamos que o espaço de estados possua dimensão $d = 1000$. Considere o reticulado bastante grosseiro dado por $P = \{a_0 = 0 < a_1 = \frac{1}{2} < a_2 = 1\}$. Para calcular $\hat{\mu}(2)$ deve-se evaluar $f(x_I)$ em 2^{1000} estados $x_I \in R_I = [a_{i_1}, a_{i_1+1}] \times [a_{i_d}, a_{i_d+1}]$, com $i_1, \ldots, i_d \in \{0, 1\}$.

> *Suponhamos que cada elétron do universo (que existem num total de 10^{80} ($\approx 2^{266}$), aproximadamente) fosse um supercomputador de 1000 gigahertz, podendo assim evaluar 1 trilhão ($\approx 2^{40}$) de $f(x_{I's})$ a cada segundo. Se todos os 2^{266} computadores rodassem simultanemente por todo o tempo de existência do universo (2^{58} segundos) teríamos no final evaluado apenas 2^{364} estados. Seria necessário esperar por mais $2^{636} \approx 10^{190}$ eras do universo até que se finalize o cálculo de todas as $f(x_{I's})$, para enfim se obter $\hat{\mu}(2)$.*

Há métodos numéricos muito mais precisos para se calcular a integral (1) (Simpson, quadratura, trapézio, Newton-Cotes). No entanto, em todos estes

métodos, a complexidade não diminui. Com isso, torna-se inviável o uso de métodos numéricos para se calcular integrais em dimensões altas.

Uma abordagem alternativa, que independa da dimensão do espaço de estados, será proposta na próxima subseção.

0.1.2 Princípio de Monte Carlo

Segundo [9], "Simulação Monte Carlo é, por essência, a geração de objetos ou processos aleatórios por meio de um computador" (trad. livre). Técnicas de Monte Carlo são todas aquelas que façam uso de amostras de variáveis aleatórias para resolver problemas tanto determinísticos quanto estocásticos, tirando proveito das propriedades probabilísticas das distribuições das quais se amostra. A ideia é a mesma de quando se estima alguma estatística através de amostras reais. No caso da Simulação Monte Carlo, as amostras não vêm de um fenômeno externo ao computador, mas são simuladas através de algum algoritmo.

O Problema das Agulhas de Buffon

Um exemplo bastante ilustrativo e de grande importância histórica é conhecido como o *Problema das Agulhas de Buffon*, proposto pelo naturalista francês George Louis Leclerc, Conde de Buffon, em 1777. É considerado o problema mais antigo conhecido no campo da probabilidade geométrica e trata-se do seguinte problema:

> *Uma agulha de comprimento L cai sobre um papel pautado com linhas retas igualmente espaçadas de distância $D \geq L$. Qual a probabilidade da agulha intersectar uma das linhas do papel?*

É um problema simples e divertido cuja probabilidade p pode ser estimada simplesmente jogando várias agulhas no papel, digamos n agulhas, com n grande, e calculando a proporção $\hat{p}_n$ daquelas que intersectam as linhas (veja Figura 1).

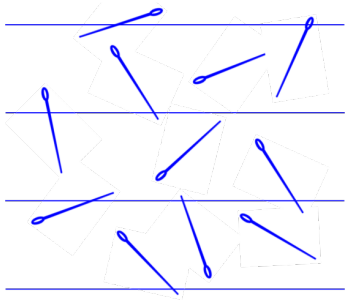

Figura 1: Várias agulhas jogadas sobre um papel pautado.

O interessante é que, com a solução deste problema, pode-se estimar π. De fato, pela conta que faremos a seguir, a probabilidade p da agulha intersectar alguma linha é dada por $p = \frac{2R}{\pi}$, onde $R = \frac{L}{D} \leq 1$. E como $\lim_{n\to\infty} \hat{p}_n = p$, com probabilidade 1, pode-se obter aproximações de π, da seguinte forma

$$\pi = \lim_{n\to\infty} \frac{2R}{\hat{p}_n},$$

com probabilidade 1.

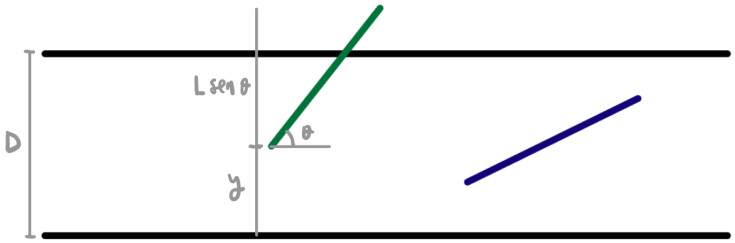

Figura 2: Parâmetros (θ, y) de uma agulha.

Com efeito, considere uma agulha, de comprimento L, caindo sobre num papel pautado com linhas igualmente espaçadas de distância $D \geq L$. Considere $y \in [0, D)$ a distância entre o ponto inferior da agulha e a respectiva linha inferior mais próxima e $\theta \in [0, \pi]$ o ângulo da agulha com a linha horizontal com base nesse ponto (veja a ilustração na Figura 2). Com isso, uma agulha de parâmetros y e θ intersecta a linha se, e somente, $y + L\,\mathrm{sen}(\theta) \geq D$,

ou seja, se, e somente se, (y, θ) pertence à região

$$\mathcal{R} = \{(y, \theta) \mid \theta \in [0, \pi] \ \text{ e } \ D - L\text{sen}(\theta) \leq y \leq D\}$$

(veja a ilustração na Figura 3). E como a agulha cai aleatoriamente sobre

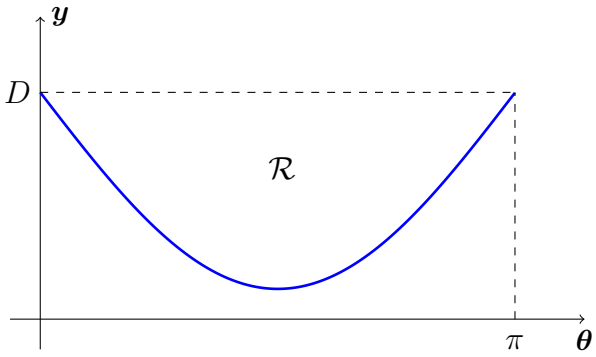

Figura 3: Região $\mathcal{R}$.

o papel, segue-se que a probabilidade de uma agulha (de parâmetros y e θ) intersectar a linha é dada pela razão das áreas

$$\frac{\text{Area}(\mathcal{R})}{\text{Area}([0, \pi] \times [0, D])} = \frac{1}{D\pi} \int_0^{\pi} D - (D - L\text{sen}(\theta))d\theta = \frac{2L}{D\pi}.$$

Agora, vamos estabelecer um algoritmo para simular esse problema no computador. Tome a variável aleatória $A = 1$ se a agulha intersecta a linha e $A = 0$ caso contrário. Tem-se que A é distribuída pela Bernoulli, $A \sim \text{Ber}(p)$, onde $p \in [0, 1]$ denota a probabilidade de se obter $A = 1$. Como as agulhas caem aleatoriamente sobre o papel, segue-se que y e θ têm distribuições uniformes,

$$y \sim \text{Unif}([0, D)) \ \text{ e } \ \theta \sim \text{Unif}([0, \pi)), \tag{2}$$

respectivamente. Assim, escreva $y = Du$ e $\theta = \pi v$, com $u, v \sim \text{Unif}([0, 1])$ independentes. Vimos também que a agulha intersecta a linha se, e somente se, $D - L\text{sen}(\theta) \leq y = Du$, ou seja, $1 - R\text{sen}(\pi v) \leq u$, onde $R = \frac{L}{D}$. Logo, $p = E[A]$, pode ser estimado tomando-se amostras independentes $u_i, v_i \sim \text{Unif}([0, 1])$, com $i = 1, \ldots, n$, e considerando a média amostral

$$\hat{p}_n = \frac{1}{n}(A_1 + \ldots + A_n),$$

onde $A_i = 1$, se $1 - R\text{sen}(\pi v_i) \leq u_i$, e $A_i = 0$, caso contrário. Tomando $\hat{\pi}_n = \frac{2R}{\hat{p}_n}$ estima-se o valor de π, com n grande.

Com este exemplo anterior, podemos entender um pouco melhor de como se faz o uso dos métodos de Monte Carlo:

- Estabelecimento do problema a ser resolvido;
- Identificação das variáveis aleatórias e técnicas de simulação que possam ajudar a resolver o problema;
- Definição de um modelo estatístico para as variáveis aleatórias;
- Simulação de valores envolvidos no problema, considerando a técnica escolhida e o modelo estatístico das variáveis;
- Análise dos resultados, chegando a conclusões para o problema ou decidindo pela repetição dos passos anteriores

Tratemos agora do problema de se calcular a média

$$\mu = E_p[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx, \tag{3}$$

onde $p(x)$ é a distribuição de probabilidade da variável aleatória $X : (\Omega, P) \to \mathcal{X} = X(\Omega) \subset \mathbb{R}^d$. Além disso, se X é uma variável aleatória discreta, a integral (3) é interpretada pela soma $E_p[X] = \sum_x f(x)p(x)$. Observe que (1) é um caso particular, tomando-se $\mathcal{X} = [0,1]^d$ e $p(x) = 1$.

Considere uma amostra iid $X_1, \ldots, X_n \sim p(x)$. Denotemos por $Y_i = f(X_i)$, para todo i, e $\bar{Y}_n = \frac{1}{n}(Y_1 + \ldots + Y_n)$. Pela lei forte dos grandes números,

$$\mathbb{P}(\bar{Y}_n \to \mu) = 1,$$

ou seja, $\bar{Y}_n$ converge para μ quase certamente. Com isso, $\hat{\mu}_n = \bar{Y}_n$ é um estimador de μ, para n grande. Precisamos estimar a precisão desse estimador. Assumindo que a variância $\sigma^2 = \text{Var}[f(X)] = \int (f(x) - \mu)^2 \, p(x) \, dx$ é finita, pelo teorema central do limite,

$$\frac{\sqrt{n}}{\sigma}(\bar{Y}_n - \mu) \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

onde "d" indica convergência em distribuição. Em outras palavras,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}(\bar{Y}_n - \mu) \leq z\right] = \phi(z),$$

onde $\phi(z)$ é a distribuição acumulada da distribuição normal padrão $\mathcal{N}(0,1)$. Com isso, segue-se que $\hat{\mu}_n = \mu + O(\frac{\sigma}{\sqrt{n}})$, quando $n \to \infty$, em distribuição, independentemente da dimensionalidade do espaço de estados.

0.1.3 Por Que Buscar Mais Métodos de Monte Carlo?

Apesar da estimação $\hat{\mu}_n = \mu + O(\frac{\sigma}{\sqrt{n}})$, em distribuição, com n grande, não ser tão boa quanto aquelas obtidas a partir dos métodos numéricos, a não-dependência da dimensionalidade faz deste método uma escolha evidente, principalmente em estados de dimensão alta. No entanto, há outros problemas:

(I) A variância de $f(x)$ pode ser muito grande;

(II) Pode ser computacionalmente caro produzir uma amostra iid grande da distribuição $p(x)$.

Com isso, pode ser difícil tomar uma amostra substancialmente grande de modo a tornar o erro $O(\frac{\sigma}{\sqrt{n}})$ suficientemente pequeno.

Para ilustrar esse problema, vamos calcular o volume da bola unitária em dimensões altas,

$$B = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\},$$

onde $\|x\| = \sqrt{x_1^2 + \ldots + x_n^2}$ denota a norma Euclideana em $\mathbb{R}^d$. O cubo circunscrito é dado por $C = [-1,1]^d$.

Considere uma amostra $x = (x^1, \ldots, x^d)$ tomada aleatoriamente sobre C. Para isso, basta tomar cada coordenada $x^i \sim \text{Unif}([-1,1])$. Segue-se que a probabilidade $p = P(x \in B)$ de x cair em B é dada por

$$p = \frac{\text{vol}\,(B)}{\text{vol}\,(C)} = \int_C \mathbb{1}_B(x)\,\frac{dx}{\text{vol}\,(C)} = E[\mathbb{1}_B(x)],$$

onde $\mathbb{1}_B(x)$ denota a função indicadora de B,

$$\mathbb{1}_B(x) = \begin{cases} 1 & \text{se } x \in B \\ 0 & \text{se } x \notin B \end{cases}$$

e a média $E[\mathbb{1}_B(x)]$ é tomada com respeito à distribuição uniforme $\text{Unif}(C)$. Assim, uma estimativa de p pode ser dada pela média amostral

$$\hat{p}_n = \frac{1}{n}(\mathbb{1}_B(x_1) + \ldots + \mathbb{1}_B(x_n)) = \frac{1}{n}\text{Card}(\{x_i \in B\}),$$

onde $x_1, \ldots, x_n \overset{iid}{\sim} \text{Unif}(C)$. É claro que uma estimativa minimamente aceitável deva ser dada por um número positivo. Logo, é necessário que haja pelo menos uma das amostras caindo sobre B. A pergunta natural então é:

> *Quantas amostras* $\{x_n\} \overset{iid}{\sim} \text{Unif}(C)$ *são esperadas para que exista ao menos uma delas caindo dentro de* B*?*

Como exemplo, considere a dimensão $d = 20$. Para esse cálculo, usamos os fatos de que $\text{vol}\,(C) = 2^d = 2^{20} \approx 1.000.000$ e $\text{vol}\,(B) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \approx 0.025$. Aqui, $\Gamma(\cdot)$ denota a função gamma. Segue-se que

$$P(x \in B) = p = \frac{\text{vol}\,(B)}{\text{vol}\,(C)} \approx 2.5 \times 10^{-9}.$$

Com isso, a cada 1 bilhão de amostras $\{x_n\} \overset{iid}{\sim} \text{Unif}(C)$ espera-se obter, no máximo, 3 delas caindo em B. E quando $d = 50$,

$$P(x \in B) \approx 1.73 \times 10^{-63}.$$

Com isso, a fim de gerar uma amostra grande o suficiente para se obter boas estimativas de p, deve-se buscar outros métodos de Monte Carlo.

Capítulo 1

Métodos Clássicos de Monte Carlo

1.1 Um Pouco de História

Conforme conta o artigo de Eckhardt [3], um dos momentos marcantes da história dos métodos Monte Carlo foi a implementação dessa técnica em computadores (ENIAC - Electronic Numerical Integrator And Computer - veja ilustração na Figura 1.2) em 1947 na pesquisa dos efeitos explosivos da bomba atômica, no Laboratório Nacional de Los Alamos, nos Estados Unidos. O objetivo dessa subseção é contar um pouco do início dessa história.

Stan Ulam repousava em sua casa por motivo de doença. Stan estava jogando uma modalidade do jogo de cartas paciência, denominado *Canfield Solitaire* (veja ilustração do jogo na Figura 1.1), quando se perguntou sobre a probabilidade de se obter um bom jogo de cartas, ou seja, um embaralhamento das cartas cujo jogo pudesse ser finalizado com sucesso.

Depois de tentativas exaustivas de cálculo combinatório para se determinar analiticamente esse valor, ele decidiu estimá-lo empiricamente. Por 100 vezes, embaralhou as 52 cartas e jogou. Depois, simplesmente contou o número de vitórias obtidas. A proporção de jogos vencidos pelo total de jogadas serviu então como uma aproximação da probabilidade de se obter sucesso nesse jogo de cartas. Apenas por curiosidade, numa simulação com 50.000 partidas de Canfield Solitaire, em torno de 71% dos jogos podem ser finalizados. No entanto, dado à complexidade do jogo (as cartas reservas são omitidas e, simultaneamente, três cartas reservas são retiradas) a taxa de

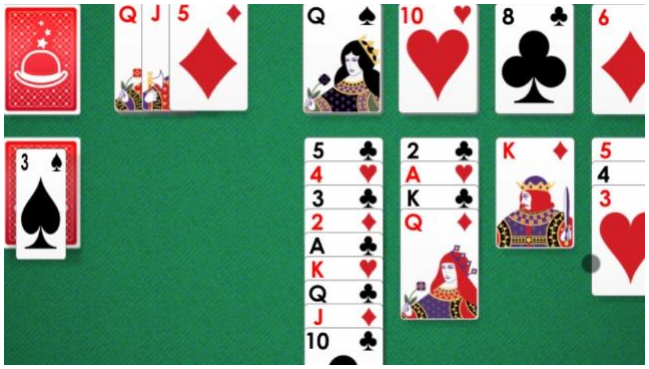

Figura 1.1: Imagem do jogo de Paciência - Canfield Solitaire

partidas vencidas por jogadores de alto nível fica em torno de 35%.

Em seguida, Stam Ulam descreveu suas ideias para John Von Neumann, que logo entendeu o seu grande potencial, aproveitando-se da evolução dos computadores. Depois de pesquisarem sobre o assunto, Von Neumann escreveu para Robert Richtmyer, que na época era o líder da Divisão Teórica do Laboratório Nacional de Los Alamos, onde estavam sendo feitas pesquisas sobre armamento nuclear. Tal método prático logo despertou-lhe interesse que delineou detalhes do uso dessa técnica no estudo da reação em cadeia da difusão e multiplicação de nêutrons na fissão nuclear. O roteiro descrito por Richtmyer foi a primeira implementação do método Monte Carlo num computador (ENIAC). Essa máquina recebeu problemas de matemáticos e físicos, até que em 1949, Metropolis e Ulam publicaram um documento sobre simulação Monte Carlo. Ulam e Von Newman desenvolveram diversos métodos Monte Carlo, tais como, transformação inversa e método da rejeição, a serem vistos ainda neste capítulo.

A ideia de se usar uma quantidade massiva de valores amostrados a partir de uma certa distribuição de probabilidade para se estimar valores determinísticos não era pioneira.

(i) Na Introdução, falamos do problema das Agulhas de Buffon em 1777, onde amostras de uma distribuição Bernoulli são usadas para estimar o valor de π. Inclusive, revisitando esse mesmo problema, Laplace propoe outras formas de se aproximar π a partir de amostras de uma certa distribuição de probabilidade.

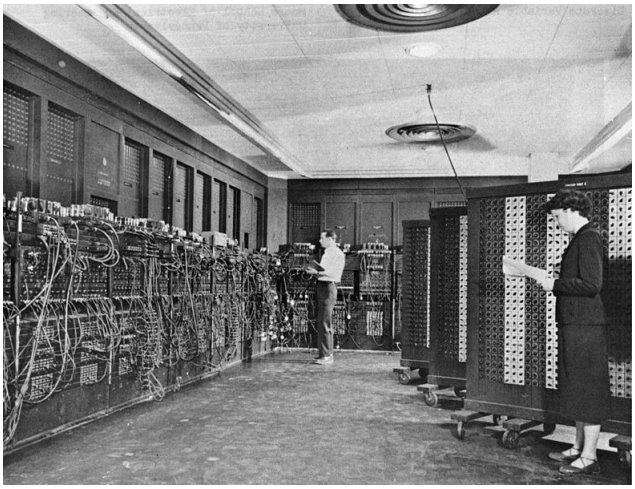

Figura 1.2: ENIAC

(ii) Existe um artigo de Lord Kelvin dezenas de anos antes em que já utilizava técnicas de Monte Carlo numa discussão das equações de Boltzman.

(iii) Segundo [1], Enrico Fermi, nos anos 30, usou métodos Monte Carlo também para cálculos sobre difusão de nêutron, tendo depois criado FERMIAC, uma máquina mecânica para cálculo usando métodos de Monte Carlo (veja Ilustração na Figura 1.3).

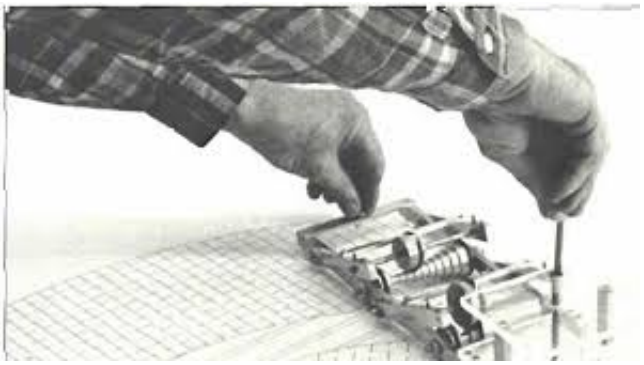

Figura 1.3: FERMIAC

Em 1952, Nicholas Metropolis e Klari Von Neumann trabalharam no aper-

feiçoamento do ENIAC, dando a ela o nome de MANIAC - Mathematical Analyzer, Numerical Integrator, and Computer ou Mathematical Analyzer, Numerator, Integrator, and Computer. Em 1953, Metropolis e os casais Rosenbluth e Teller propuseram o importante algoritmo Metropolis [14].

Indo além da física, os métodos de Monte Carlo se expandiram para aplicações em estatística, ciências biomédicas, redes neurais, visão computacional e inteligência artificial. Em nota anexa de [3], Gary D. Doolenand e John Hendricks contam que, na época da publicação do artigo, a cada segundo aproximadamente 10 bilhões de números aleatórios eram gerados por computadores de todo o mundo. Mais de trinta anos depois, este número certamente cresceu alguns dígitos.

Nas próximas seções desse capítulo, visitaremos alguns métodos historicamente importantes, que ainda apresentam sua utilidade.

1.2 Métodos Clássicos de Simulação

1.2.1 Amostragem Uniforme no Intervalo

Os métodos mais simples de simulação são baseados na manipulação de amostras previamente geradas da distribuição uniforme no intervalo $[0, 1]$. Vale ressaltar aqui que nenhum computador é capaz de gerar números realmente aleatórios, pois sempre processos determinísticos são usados para a geração dos valores. Apesar dessa limitação, os chamados números pseudo-aleatórios podem ser gerados, carregando propriedades que os tornem suficientemente bons para o uso como se fossem aleatórios.

Os algoritmos usados para a geração de tais números baseiam-se em processos caóticos no sentido de que, dado um valor de entrada, o computador retorna um número como amostra, sendo este sensível à chave de entrada - pequenas alterações criam valores possivelmente distintos em larga escala. Uma destas técnicas é aceita se suas amostras passam nos devidos testes de hipóteses relacionados a valores independentes provenientes da distribuição uniforme.

Um exemplo de algoritmo para obter números pseudoaleatórios é o gerador congruencial linear (LCG). Esse método cria uma sequência de valores

usando a operação aritmética

$$X_{n+1} = aX_n + c \mod m,$$

onde a é o multiplicador, c o incremento e m o módulo. A escolha adequada desses valores faz com que o algoritmo percorra os números entre 0 e m de forma aparentemente aleatória. Não queremos nos alongar nesse assunto, pois esses métodos já estão muito bem implementados, usando potências altas de 2 ou números primos grande. Apenas para tomar como exemplo, segue a sequência gerada pelos valores $a = 13$, $c = 1$, $m = 64$ e $X_0 = 1$:

$$(14, 55, 12, 29, 58, 51, 24, 57, 38, 47, 36, 21, 18, 43, 48, 49, 62, 39, 60, 13).$$

A partir de números inteiros pseudoaleatórios, a divisão pelo módulo m faz com que todos os valores se acumulem no intervalo $[0, 1]$, podendo ser usados como vindo da distribuição $\text{Unif}_{[0,1]}$.

Tendo deixado claro que os números gerados nos computadores são sempre determinísticos, a utilidade dos números pseudoaleatórios é inegável, pois se comportam tal como números aleatórios. Então, assim eles serão chamados no resto dessa dissertação, omitindo o prefixo "pseudo".

1.2.2 Amostragem por Transformação Inversa

Uma vez que se entende por satisfatória a geração de amostras de números aleatórios

$$u_1, \ldots, u_N \overset{iid}{\sim} \text{Unif}_{[0,1]}$$

no computador, é natural buscar por métodos que amostrem de outras distribuições de probabilidade. O método da transformação inversa fornece uma maneira simples e direta de se simular a partir de uma distribuição de probabilidade univariada. No entanto, exige-se o conhecimento da inversa F^{-1} da distribuição de probabilidade acumulada $F(x)$, o que pode ser um problema muito caro computacionalmente.

Seja $X : (\Omega, P) \to \mathcal{X} = X(\Omega) \subset \mathbb{R}$ uma variável aleatória unidimensional e $F(x)$ a sua distribuição de probabilidade acumulada (CDF),

$$F(x) = P(X \leq x).$$

Observa-se que:

(a) F é monótona não-decrescente. De fato, se $x \leq y$, então os eventos $[X \leq x] \subseteq [X \leq y]$. Segue-se então $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$. Donde F é monótona não-decrescente.

(b) F é contínua à direita. De fato, considere a sequência $\delta_n > 0$ com $\lim \delta_n = 0$. Defina a cadeia de eventos $U_n = [X \leq x + \delta_n]$ e $U = [X \leq x]$. A função indicadora $f_n(w) = \mathbb{1}_{U_n}(w)$ satisfaz: $\lim f_n(w) = f(w)$, onde $f(w) = \mathbb{1}_U(w)$, para todo w. Além disso, as funções f_n são mensuráveis com $f_n \leq 1$ e a integral $\int f_n(w) dP_w = P(X \leq x + \delta_n) = F(x + \delta_n)$. Como $\int 1\, dP_w = 1$, segue-se, do teorema da convergência dominada, que

$$\lim_{n\to\infty} F(x + \delta_n) = \lim_{n\to\infty} \int f_n(w) dP_w = \int f(w) dP_w = P(X \leq x) = F(x).$$

Segue-se então que F é contínua à direita.

A função $F(x)$ pode ser descontínua. Com efeito, considere X assumindo valores discretos $\mathcal{X} = \{x_0 < x_1 < \ldots\}$. Tem-se que $F(x)$ é uma função escada, assumindo valores constantes no intervalo $[x_k, x_{k+1})$, com

$$\begin{aligned} F(x_{k+1}) - F(x_k) &= P(X \leq x_{k+1}) - P(X \leq x_k) \\ &= P(X \leq x_{k+1}) - P(X < x_{k+1}) \\ &= P(X = x_{k+1}) \end{aligned}$$

(veja ilustração na Figura 1.4).

Defina a inversa generalizada de F, dada por

$$F^-(u) = \inf(\{x \mid F(x) \geq u\}).$$

Teorema 1 (Von Neumann, 1947)**.** *Seja $F(x)$ a distribuição de probabilidade acumulada de uma variável aleatória X. Seja $U \sim \text{Unif}_{[0,1]}$ uma variável aleatória uniformemente distribuída em $[0, 1]$. Então,*

$$X \stackrel{d}{=} F^-(U) \quad e \quad F(X) \stackrel{d}{=} U.$$

Para a segunda igualdade, deve-se assumir que $F(x)$ seja contínua. Aqui, a igualdade "$\stackrel{d}{=}$" indica que ambas variáveis aleatórias possuem a mesma distribuição de probabilidade.

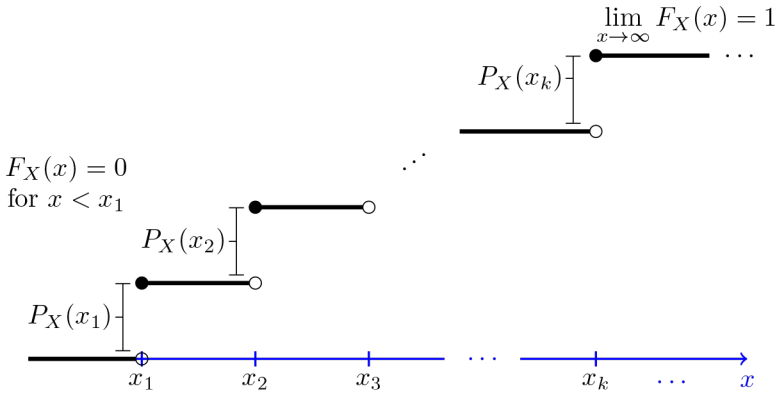

Figura 1.4: CDF $F_X(x)$ associada à variável aleatória discreta X.

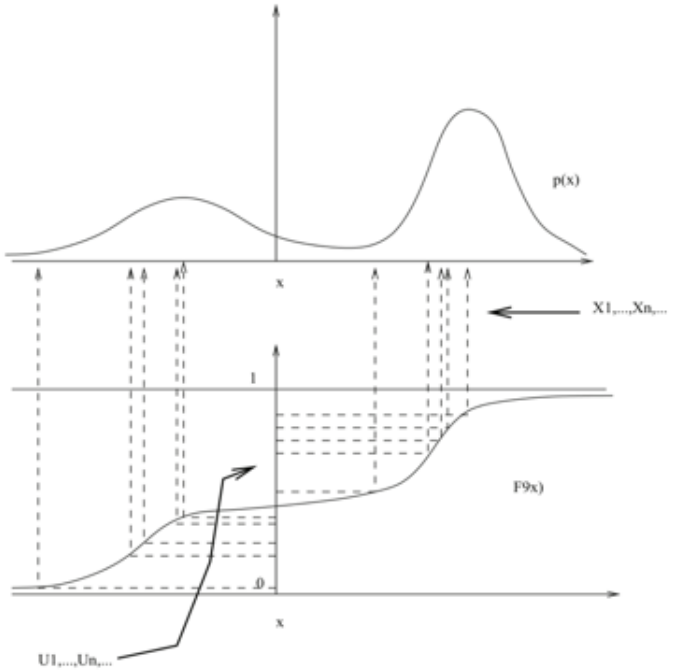

Figura 1.5: Esquema gráfico do método da transformação inversa

Demonstração: Primeiro, afirmamos que $[F^-(u) \leq x] = [u \leq F(x)]$. De fato, como $F(x)$ está definida em toda a reta, da definição de $F^-(u)$, segue-se que $F(x) \geq u$ implica que $F^-(u) \leq x$, ou seja, temos a inclusão $[F(x) \geq u] \subset [x \geq F^-(u)]$. Reciprocamente, tome $F^-(u) \leq x$. Como

F é monótona não-decrescente, $F(F^-(u)) \leq F(x)$. Agora, considere uma sequência $x_k \geq F^-(u)$, com $F(x_k) \geq u$ e $\lim x_k = F^-(u)$. Como F é contínua à direita, temos que $u \leq \lim F(x_k) = F(F^-(u))$. Conclui-se assim que $F(x) \geq F(F^-(u)) \geq u$. Daí, segue-se a igualdade $[F^-(u) \leq x] = [u \leq F(x)]$.

Portanto, se $U \sim \text{Unif}_{[0,1]}$, então $F(x) = P(U \leq F(x)) = P(F^-(U) \leq x)$. Conclui-se que X e $F^-(U)$ possuem as mesmas CDF's, donde $F^-(U) \stackrel{d}{=} X$.

Agora, suponha F contínua. Como $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$, segue-se, do teorema do valor intermediário, que F assume todos os valores em $(0, 1)$. Assim, dado $u \in (0, 1)$, existe x tal que $F(x) = u$. Tem-se então que $F^-(u) = \inf\{x \mid F(x) = u\}$. Logo, existe uma sequência $x_k \geq F^-(u)$ com $F(x_k) = u$ e $\lim x_k = F^-(u)$. Como F é contínua a direita, tem-se que $u = \lim F(x_k) = F(F^-(u))$. E, como $F^-(U) \stackrel{d}{=} X$ e $F(F^-(u)) = u$, para quase todo $u \in [0, 1]$, tem-se que $F(X) \stackrel{d}{=} F(F^-(U)) \stackrel{d}{=} U$. □

1.2.3 O Método de Box-Muller

O método de Box-Muller [2] fornece uma maneira simples de se amostrar da distribuição normal $\mathcal{N}(\mu, \sigma)$. Para simplificar o problema, usando que $X = \mu + \sigma Z$, com $Z \sim \mathcal{N}(0, 1)$, em distribuição, vemos que o problema se reduz a simplesmente amostrar da Gaussiana padrão $Z \sim \mathcal{N}(0, 1)$. O problema é que não se pode aplicar diretamente o método da transformação inversa, visto que a CDF de Z,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt,$$

não admite expressão em termos de funções algébricas. Assim, inverter $\phi(x)$ numericamente pode ser um problema caro computacionalmente.

O método de Box-Muller baseia-se na estratégia clássica de como se calcula a integral $I = \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx$. Com efeito, considere $(x, y) \in \mathbb{R}^2$ escrita em termos de coordenadas polares,

$$x = r \cos\theta \quad \text{e} \quad y = r \sin\theta,$$

com $r \geq 0$ e $\theta \in [0, 2\pi]$. A estratégia é calcular I^2 ao invés de I. Usando que, na mudança de coordenadas polares, $dxdy = rdrd\theta$,

$$I^2 = (\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx)(\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dxdy$$
$$= \int_0^{2\pi}\int_0^{+\infty} e^{-\frac{r^2}{2}} rdrd\theta = 2\pi \int_0^{+\infty} e^{-\frac{r^2}{2}} rdr = 2\pi\big[-e^{-\frac{r^2}{2}}\big]_0^{+\infty} = 2\pi,$$

donde $I = \sqrt{2\pi}$.

Aproveitando-se das ideias dessa conta, escrevamos

$$\pi(x, y) = \mathcal{N}(x \mid 0, 1)\mathcal{N}(y \mid 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

Usando que o determinando Jacobiano $\frac{\partial(x,y)}{\partial(r,\theta)} = r$, a distribuição de (x, y) em termos das coordenadas (r, θ) é dada por

$$g(r, \theta) = \pi(x, y)|\frac{\partial(x, y)}{\partial(r, \theta)}| = \frac{1}{2\pi} e^{-\frac{r^2}{2}} r.$$

E como as distribuições marginais:

(i) $g(r) = \int_0^{2\pi} g(r, \theta)d\theta = e^{-\frac{r^2}{2}} r$,

(ii) $g(\theta) = \int_0^{+\infty} g(r, \theta)dr = \frac{1}{2\pi}$,

obtemos que $g(r, \theta) = g(r)g(\theta)$, donde as variáveis aleatórias r e θ são independentes. Como $g(\theta) = \frac{1}{2\pi}$, tem-se que $\theta \sim \text{Unif}_{[0,2\pi]}$. Para obter amostras de $g(r)$, podemos usar o método da transformação inversa. Para isso, observemos que a CDF

$$G(r) = \int_0^r g(s)\, ds = \int_0^r e^{-\frac{s^2}{2}} s\, ds = 1 - e^{-\frac{r^2}{2}},$$

donde $G^{-1}(u) = \sqrt{-2\ln(1-u)}$. E, como $U \overset{d}{=} 1-U$, para todo $U \sim \text{Unif}_{[0,1]}$, segue-se que $R \overset{d}{=} \sqrt{-2\ln(U)}$, com $U \sim \text{Unif}_{[0,1]}$.

O método de Box-Muller resume-se no seguinte algoritmo, no caso de dimensão $d = 2$

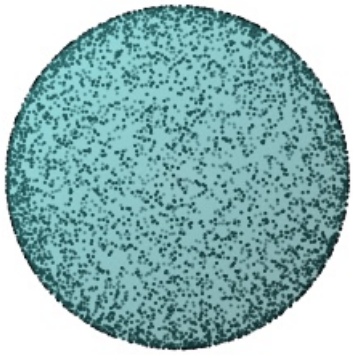

Figura 1.6: Amostragem uniforme sobre a esfera

Algoritmo 1: Método de Box-Muller

Entrada: $u_1, u_2 \overset{iid}{\sim} \text{Unif}_{[0,1]}$
Escreva: $r \leftarrow \sqrt{-2\ln(u_1)}$ e $\theta \leftarrow 2\pi\, u_2$
Faça: $x \leftarrow r\cos\theta$ e $y \leftarrow r\sin\theta$
Saída : $x, y \overset{iid}{\sim} \mathcal{N}(0,1)$

Amostragem Uniforme na Esfera (Muller, 1959)

Em 1959, Muller [15] observou que amostras normais $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ podem ser usadas para se obter amostras uniformes sobre esferas Euclideanas

$$S^{n-1} = \{w \in \mathbb{R}^n \mid \|w\| = 1\}.$$

(veja Figura 1.6). Com efeito, considere n amostras iid's $\{X_k\} \overset{iid}{\sim} \mathcal{N}(0,1)$. Segue-se assim, o vetor aleatório $X = (X_1, \ldots, X_n) \sim N(0, I_n)$, onde I_n denota a matriz identidade $n \times n$.

Escreva X em coordenadas esféricas $X = RW$, onde $R^2 = \|X\|^2 = X_1^2 + \ldots + X_n^2$ e $W = \frac{X}{R}$. Como $\|W\| = \frac{\|X\|}{R} = 1$, temos que W é um vetor aleatório na esfera S^{n-1}.

Afirmação. $W \sim \text{Unif}_{S^{n-1}}$.

Com efeito, a métrica Euclideana de $\mathbb{R}^n$, $g = \langle\,,\rangle = dx_1^2 + \ldots + dx_n^2$ pode ser escrita em termos das coordenadas (r, w), onde $x = rw$, como

$$g = dr^2 + r^2 g_{S^{n-1}},$$

onde $g_{S^{n-1}}$ é a métrica standard de S^{n-1} no ponto w. Assim, a forma elemento de volume de $\mathbb{R}^n$, pode ser escrita, em termos de (r, w), por

$$d\mathrm{vol}_{\mathbb{R}^n} = dx_1 \ldots dx_n = r^{n-1}\, dr\, d\mathrm{vol}_{S^{n-1}},$$

onde $d\mathrm{vol}_{S^{n-1}}$ denota a forma elemento de volume de S^{n-1}. Assim, a distribuição normal multivariada $\mathcal{N}(0, I_n)$, escrita em termos da densidade de medida, se escreve da seguinte forma

$$p(x) d\mathrm{vol}_{\mathbb{R}^n} = (2\pi)^{-n/2} e^{-\frac{\|x\|^2}{2}} d\mathrm{vol}_{\mathbb{R}^n} = (2\pi)^{-n/2} e^{-\frac{r^2}{2}} r^{n-1}\, dr\, d\mathrm{vol}_{S^{n-1}}.$$

Aqui é importante observar que à vezes é conveniente denotar uma distribuição de probabilidade em termos da sua densidade de medida,

$$P(X \in A) = \int_A dP(x).$$

Além disso, vale que $dP(x) = p(x) d\mathrm{vol}_{\mathbb{R}^n}$, sempre que X for um vetor aleatório contínuo sobre um aberto do $\mathbb{R}^n$.

Logo, na mudança de coordenadas, $x = rw$, temos a seguinte distribuição de probabilidade

$$p(r, w) = (2\pi)^{-n/2} e^{-\frac{r^2}{2}} r^{n-1}.$$

Portanto, a distribuição marginal $p(w) = \int_0^\infty p(r, w) dr$ é constante, o que nos dá que $W \sim \mathrm{Unif}_{S^{n-1}}$.

Outros métodos de amostragem uniforme na esfera podem ser vistos em Hicks and Wheeling [6] e em Marsaglia [13].

Amostragem Uniforme na Bola (Muller, 1959)

Novamente, podemos recorrer à geometria para amostrar uniformemente na bola unitária. Novamente, escrevendo $x \in B$ em coordenadas esféricas, $x = rw$, com $r \in [0, 1]$ e $w \in S^{n-1}$, temos a forma elemento de volume,

$$d\mathrm{vol}_B = dx_1 \ldots dx_n = r^{n-1} dr d\mathrm{vol}_{S^{n-1}}.$$

Em particular,

$$\operatorname{vol} B = \int_0^1 \int_{S^{n-1}} r^{n-1} dr d\operatorname{vol}_{S^{n-1}} = \operatorname{vol}(S^{n-1}) \int_0^1 r^{n-1} dr = \operatorname{vol}(S^{n-1}) \frac{1}{n}.$$

Assim, temos as densidades de probabilidade,

$$\pi(x) dP(x) = \frac{d\operatorname{vol}_B}{\operatorname{vol} B} = n r^{n-1} dr \, \frac{d\operatorname{vol}_{S^{n-1}}}{\operatorname{vol}_{S^{n-1}}} = \pi(r) dr \, d \operatorname{Unif}_{S^{n-1}}(w),$$

onde $\pi(r) = n r^{n-1} dr$. Assim, basta amostrar $X = RW$, com $R \sim \pi(r)$ e $W \sim \operatorname{Unif}_{S^{n-1}}$, independentemente. Já vimos que o problema de amostrar uniformemente em S^{n-1} é simples, basta considerar $W = \frac{Z}{\|Z\|}$ com $Z \sim \mathcal{N}(0, I_n)$. Para amostra de $\pi(r)$, basta aplicar o método da transformação inversa, $R \sim U^{\frac{1}{n}}$, com $U \sim \operatorname{Unif}_{[0,1]}$.

1.3 Amostragem por Rejeição

Conforme observamos na Introdução, segundo o artigo de Eckhardt [3], os métodos da transformação inversa e da amostragem por rejeição foram descritos em 1947 nas cartas de Von Neumann para Ulam, motivados principalmente pelo possível uso dessas técnicas em simulação de dados no ENIAC. Em alguns textos, o método da rejeição também é chamado de método da aceitação e rejeição ou, simplesmente, método AR.

Como todo em problema de simulação, pretende-se amostrar de uma distribuição de probabilidade $\pi(x)$, com $x \in \mathcal{X}$. Chamamos $\pi(x)$ de distribuição de probabilidade *objetivo*. Considere $q(x)$, com $x \in \mathcal{X}$, outra distribuição de probabilidade, chamada de distribuição *proposta*, da qual seja possível amostrar e que majore $\pi(x)$, i.e., existe uma constante $M > 0$ tal que $\pi(x) \leq M q(x)$, para todo x. Em geral, sabe-se evaluar apenas versões não-normalizadas de $\pi(x)$ e $q(x)$, ou seja, $\tilde{\pi}(x) \propto \pi(x)$ e $\tilde{q}(x) \propto q(x)$. Logo, considere apenas conhecido $N > 0$ tal que $\tilde{\pi}(x) \leq N \tilde{q}(x)$, para todo $x \in \mathcal{X}$. Assim, defina a função $r(x) = \frac{\tilde{\pi}(x)}{N \tilde{q}(x)} \leq 1$, para todo x.

Neste método, amostras de $q(x)$ são usadas para obter amostras de $\pi(x)$. O procedimento é o seguinte:

- Amostre $x \sim q(x)$ e evalue $r(x) = \frac{\tilde{\pi}(x)}{N \tilde{q}(x)} \leq 1$;

- Amostre $u \sim \text{Unif}_{[0,1]}$;
- Se $u \leq r(x)$, aceite x. Caso contrário, rejeite x e reinicie o processo.

Em outras palavras, considere o seguinte algoritmo:

Algoritmo 2: Amostragem por Rejeição

Entrada: $\tilde{\pi}, \tilde{q}, N$
enquanto *não há amostra aceita* **faça**
  Amostre $x \sim q(x)$ e $u \sim \text{Unif}_{[0,1]}$
  $r(x) \leftarrow \frac{\tilde{\pi}(x)}{N\tilde{q}(x)}$
  se $u < r(x)$ **então**
    aceito a amostra x
Saída : $x \sim \pi(x)$

Teorema 2 (Von Neumann, 1947)**.** *Sejam $x \sim q(x)$ e $u \sim \text{Unif}_{[0,1]}$. Os seguintes itens são válidos:*

(i) Se x foi aceito no método da rejeição então vale também que $x \sim \pi(x)$, i.e, $p(x \mid x \text{ foi aceito }) = \pi(x)$;

(ii) A taxa de aceitação $P(\text{“aceitar” } x)$ é o valor esperado $E_q[r(x)]$. Em particular, escrevendo $r(x) = \frac{\pi(x)}{Mq(x)}$, temos que $P(\text{“aceitar” } x) = \frac{1}{M}$.

Demonstração: Considere as amostras $x \sim q(x)$ e $u \sim \text{Unif}_{[0,1]}$. Defina a variável aleatória $I = I(x, u) \in \{0, 1\}$, dada por

$$\begin{aligned} I &= 1, \quad \text{se } x \text{ foi aceito, i.e., } u \leq r(x); \\ I &= 0, \quad \text{caso contrário.} \end{aligned}$$

Primeiro, vamos calcular a distribuição de probabilidade $p(x \mid I = 1)$. Para isso, considere a função de verossimilhança,

$$P(I = 1 \mid x) = P(U \leq r(x)) = r(x) = \frac{\tilde{\pi}(x)}{N\tilde{q}(x)} \propto \frac{\pi(x)}{q(x)}. \tag{1.1}$$

Assim, pelo Teorema de Bayes,

$$p(x \mid I = 1) \propto P(I = 1 \mid x)q(x) \propto \frac{\pi(x)}{q(x)}q(x) = \pi(x),$$

e, como ambas funções $p(x \mid I = 1)$ e $\pi(x)$ são distribuições de probabilidade, segue-se que $p(x \mid I = 1) = \pi(x)$.

Para provar o item (ii), usando novamente (1.1), temos que a taxa de aceitação $P(\text{“aceitar”}\ \ x) = P(I = 1)$ é dada por

$$P(I = 1) = \int P(I = 1 \mid x)q(x)dx = \int r(x)q(x)dx = E_q[r(x)].$$

E assim o teorema 2 está provado. □

O método da rejeição pode ser usado para amostrar mesmo em dimensões altas, o que não ocorre no método da transformação inversa. E, de fato, ainda é um método bastante usado. As amostras $\{x_k\}$ obtidas pelo método AR são independentes e exatas (no sentido de que $\{x_k\} \sim \pi(x)$). O problema é que o sucesso do método depende fortemente da boa escolha da distribuição proposta $q(x)$, o que pode ser um problema bastante complicado, principalmente em dimensões altas. Para exemplificar esse problema, vamos propor dois exemplos.

Exemplo 1. *Considere as distribuições normais multivariadas $\mathcal{N}(0, I_d)$ e $\mathcal{N}(0, \sigma^2 I_d)$, onde $\sigma = 0,99$ e I_d a matriz identidade $d \times d$.*

Primeiro, observamos que não há dificuldade alguma em se amostrar de ambas distribuições. De fato, escreva $Z = (Z_1, \ldots, Z_d)$ com $Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Segue-se que $Z \sim \mathcal{N}(0, I_d)$ e $X = \sigma Z \sim \mathcal{N}(0, \sigma^2 I_d)$.

Porém, pode-se tentar aplicar o método AR para amostrar da distribuição objetivo $\pi(x) = \mathcal{N}(0, \sigma^2 I_d)$ a partir da distribuição proposta $q(x) = \mathcal{N}(0, I_d)$. Usando que $\mathcal{N}(0, \sigma^2 I_d) = \frac{(2\pi)^{-d/2}}{\sigma^d} \exp(-\frac{\|x\|^2}{2\sigma^2})$, segue-se que a razão $\frac{\pi(x)}{q(x)}$ é dada por

$$\frac{\pi(x)}{q(x)} = \frac{1}{\sigma^d} \exp\big(-\frac{\|x\|^2}{2}[\frac{1}{\sigma^2} - 1]\big) \leq \frac{1}{\sigma^d} =: M,$$

visto que $\frac{1}{\sigma^2} \geq 1$. Assim, a taxa de aceitação $\frac{1}{M} = \sigma^d = 0,99^d$, tende a zero quando $d \to \infty$. Por exemplo, tomando-se $d = 1000$, tem-se $\frac{1}{M} \approx 4,3 \times 10^{-5}$.

Exemplo 2. *Considere o problema de se obter amostras uniformes sobre a bola unitária d-dimensional,* $B = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$.

Já vimos que amostrar $X \sim \text{Unif}_B$ *é um problema simples, basta tomar* $X = RW$, *onde* $R = U^{\frac{1}{d}}$, *com* $U \sim \text{Unif}_{[0,1]}$, *e* $W = \frac{Z}{\|Z\|}$, *com* $Z \in \mathcal{N}(0, I_d)$.

No entanto, vamos tentar abordar esse problema via método da rejeição. Um pouco dessa abordagem já foi feita na Introdução, na tentativa de se estimar o volume de B. *Mantendo as mesmas notações, considere as distribuições* $\pi(x) = \text{Unif}_B = \frac{\mathbb{1}_B(x)}{\text{vol}\, B}$ *e* $q(x) = \text{Unif}_C = \frac{\mathbb{1}_C(x)}{\text{vol}\, C}$, *com* $x \in C = [-1, 1]^d$. *Como* $B \subset C$, *tem-se que* $\mathbb{1}_B \leq \mathbb{1}_C$, *donde* $\frac{\pi(x)}{q(x)} \leq \frac{\text{vol}\, C}{\text{vol}\, B} = \frac{1}{P(x \in B)}$, *para todo* $x \in C$. *Assim,* $\pi(x) \leq Mq(x)$, *para todo* x, *com* $M = \frac{1}{P(x \in B)}$.

É claro que amostrar de $q(x) = \text{Unif}_C(x)$ *é um problema simples, basta tomar amostras de* $X = (X_1, \ldots, X_d)$ *cujas coordenadas* $X_j \overset{iid}{\sim} \text{Unif}_{[-1,1]}$. *Assim, o sucesso do método AR depende da taxa de aceitação que, pelo Teorema 2, é igual a* $\frac{1}{M} = P(x \in B)$. *Na Introdução, vimos que se tomarmos* $d = 20$, *então* $P(x \in B) \approx 2,5 \times 10^{-9}$. *Assim, o método da rejeição não parece ser o mais adequado para simular amostras uniformes sobre* B.

1.4 Integração Monte Carlo

Na Introdução, abordamos o problema de se estimar o valor esperado,

$$\mu = E[f(X)] = \int_{\mathcal{X}} f(x)\pi(x)dx.$$

Vimos que, em dimensões altas (por exemplo, $\mathcal{X} = [0, 1]^d$, com d grande) a abordagem por métodos numéricos tornava-se completamente inviável. A proposta era então considerar amostras aleatórias $\{x_1, \ldots, x_N\} \overset{iid}{\sim} \pi(x)$ e estimar μ pela média amostral,

$$\hat{\mu} = \frac{1}{N} \sum_k f(x_k).$$

E, do teorema central do limite, segue-se $\hat{\mu} = \mu + O(\frac{\sigma}{\sqrt{N}})$, quando $N \to \infty$, onde $\sigma^2 = \text{Var}(f(X))$, independentemente da dimensão de $\mathcal{X}$.

Em integração Monte Carlo busca-se estimar uma integral através de simulação de amostras. As técnicas são pensadas em termos de custo computacional, tamanho efetivo da amostra e redução da variância.

1.4.1 Amostragem por Importância

Em muitos casos, simular uma quantidade significativa de amostras de uma distribuição objetivo $\pi(x)$ pode ser impossível ou muito caro computacionalmente. Com isso, a eficiência de se estimar $\mu = E_\pi[f(X)]$ pela média amostral $\hat{\mu} = \frac{1}{N}\sum_k f(x_k)$ pode ser baixa. Amostragem por importância é uma técnica de integração Monte Carlo onde simula-se μ a partir de amostras de uma distribuição proposta $q(x)$. O primeiro registro dessa técnica apareceu no artigo de Kahn e Harris [7], em 1951. O problema relatado neste artigo era a dificuldade de estimar a probabilidade de uma partícula atravessar um escudo de radiação quando esta probabilidade era da ordem de 10^{-10} a 10^{-6}. E, com a limitação de processamento dos computadores da época, era realmente um problema difícil de se atacar com os métodos da transformação inversa e rejeição.

A técnica de amostragem por importância é de um procedimento bastante simples. Observe que

$$\mu = E_\pi[f(X)] = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx = E_q[f(X)w(X)],$$

onde $w(x) = \frac{\pi(x)}{q(x)}$. Daí, toma-se uma amostra $\{x_k\} \overset{iid}{\sim} q(x)$ e estima-se μ pela seguinte média amostral:

$$\hat{\mu} = \frac{1}{N}\sum_k f(x_k)w(x_k). \tag{1.2}$$

Em resumo, tem-se o seguinte algoritmo:

Algoritmo 3: Amostragem por Importância

Entrada: $f(x)$, $\pi(x)$, $q(x)$, N
Amostre $x_1, \ldots, x_N \overset{iid}{\sim} q(x)$
Faça $w_k \leftarrow \frac{\pi(x_k)}{q(x_k)}$, para todo $k = 1, \ldots, N$
Escreva $\hat{\mu} \leftarrow \frac{1}{N} \sum f(x_i) w_i$
Saída : $\hat{\mu}$

O estimador é não-enviesado, visto que

$$E_q[\hat{\mu}] = \frac{1}{N} \sum_k E_q[f(X)w(X)] = \mu.$$

Já a variância de $\hat{\mu}$ é

$$\mathrm{Var}_q[\hat{\mu}] = \frac{1}{N^2} \sum_k \mathrm{Var}_q[f(X)w(X)] = \frac{1}{N} \mathrm{Var}_q[f(X)w(X)].$$

O sucesso desse método depende da escolha da distribuição proposta $q(x)$ que minimize a variância $\mathrm{Var}_q[f(X)w(X)]$.

Afirmação. *A distribuição proposta $q(x)$ que minimiza $\mathrm{Var}_q[f(X)w(X)]$ é dada por $q(x) \propto |f(x)|\pi(x)$.*

Com efeito, denote por $c = \int |f(x)|\pi(x)dx = E_\pi\big[|f(X)|\big]$ e considere a distribuição $p(x) = \frac{1}{c}|f(x)|\pi(x)$. Assim, $E_q\big[\frac{p(X)}{q(X)}\big] = \frac{1}{c} E_\pi\big[|f(X)|\pi(X)\big] = 1$. Logo,

$$\begin{aligned}
\mathrm{Var}_q\big[f(X)w(X)\big] &= E_q\big[\frac{f(X)^2\pi(X)^2}{q(X)^2}\big] - \mu^2 = c^2 E_q\big[\frac{p(X)^2}{q(X)^2}\big] - \mu^2 \\
&= c^2\big(\mathrm{Var}_q\big[\frac{p(X)}{q(X)}\big] + E_q\big[\frac{p(X)}{q(X)}\big]^2\big) - \mu^2 \\
&= c^2\big(\mathrm{Var}_q\big[\frac{p(X)}{q(X)}\big] + 1\big) - \mu^2 = c^2 \mathrm{Var}_q\big[\frac{p(X)}{q(X)}\big] + c^2 - \mu^2.
\end{aligned}$$

Observe que $c = E_\pi\big[|f(X)|\big] \geq \mu = E_\pi\big[f(X)\big]$ não dependem de $q(x)$. Portanto, $\mathrm{Var}_q\big[f(X)w(X)\big]$ é minimizado tomando $q(x) = \frac{1}{c}|f(x)|\pi(x)$. $\square$

Vejamos como amostragem por importância se aplica na estimativa do volume da bola unitária em dimensão alta. Considerando uma distribuição proposta $q(x)$, com $x \in \mathbb{R}^d$, temos

$$\text{vol}\,(B) = \int \frac{\mathbb{1}_B(x)}{q(x)}\, q(x)\, dx.$$

Para uma boa escolha de distribuição proposta, façamos $q(x)$ próxima, em termos da divergência χ^2, da distribuição $p(x) = \frac{1}{\text{vol}\,(B)}\mathbb{1}_B(x) \approx \delta(x)$, onde $\delta(x)$ é a distribuição delta de Dirac. Assim, tome $q(x) = \mathcal{N}(0, \sigma^2 I_d)$, com $\sigma > 0$ pequeno. Num teste, consideramos as dimensões $d = 50, 60, 70$ e 100. Como

$$P(x \in B) \approx 1.73 \times 10^{-63},$$

o método da rejeição falha completamente. Agora, aplicando Algoritmo 3, com $q(x) = \mathcal{N}(0, \sigma^2)$ e $N = 10^6$ (total de amostras $\{x_k\} \overset{iid}{\sim} q(x)$), temos:

d	σ^2	Volume estimado	Volume verdadeiro	Time
50	0.017	1.7305460128463707e-13	1.7302192458361097e-13	4.04 s
60	0.014	3.0910088860855338e-18	3.096250615296861e-18	4.7 s
70	0.012	2.4304894236882064e-23	2.4322762320344753e-23	5.55 s
100	0.009	2.373938564160651e-40	2.3682021018828293e-40	7.78 s

Observe que a escolha de σ^2 na tabela acima foi feita em termos da dimensão d. Isso se deve ao fato de que a amostragem de $q(x) = \mathcal{N}(0, \sigma^2 I_d)$ pode ficar muito longe da origem, contrariando completamente a nossa intuição da distribuição normal. Com efeito, considere $X \sim q(x) = \mathcal{N}(0, \sigma^2 I_d)$. No livro do Mackay [12], há o seguinte resultado:

Proposição 1. *Considere $X \in \mathcal{N}(0, \sigma^2 I_d)$. Então,*

$$\lim_{d\to\infty} \|X\|^2 \overset{dist}{=} \sigma^2(d + \sqrt{2d}Z),$$

onde $Z \sim N(0, 1)$. Aqui, "dist" denota que o limite é em distribuição.

Assim, para assegurar que haja amostras de $q(x)$ caindo dentro da bola, estamos considerando na tabela acima $\sigma^2 = (d + \sqrt{2d}\,)^{-1}$.

Demonstração: Considere $X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, \sigma^2 I_d)$. Assim, as coordenadas de $X = (X_i)$ satisfazem $\{X_i\} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Logo, a média $E[X_i^2] = \sigma^2$ e a variância $\text{Var}[X_i^2] = E[X_i^4] - E[X_i^2]^2 = 2\sigma^4$. E, pelo teorema central do limite, a norma Euclideana de X satisfaz:

$$\frac{\|X\|^2}{d} = \frac{X_1^2 + \ldots + X_d^2}{d} \approx \sigma^2 + \frac{\sqrt{2}\sigma^2}{\sqrt{d}} Z,$$

em distribuição, onde $Z \sim N(0, 1)$, quando $d \to \infty$. Assim,

$$\|X\|^2 \approx \sigma^2(d + \sqrt{2d}Z), \tag{1.3}$$

em distribuição, quando a dimensão $d \to \infty$. □

Amostragem por Importância Auto-Normalizada

É comum os casos em que, apesar de se poder amostrar da distribuição proposta $q(x)$, podem-se evaluar apenas versões não-normalizadas das distribuições objetivo e proposta, $\tilde{\pi}(x) \propto \pi(x)$ e $\tilde{q}(x) \propto q(x)$, respectivamente. Mesmo assim, pode-se estabelecer uma versão não-normalizada da técnica de amostragem por importância a fim de se estimar a média $\mu = E_\pi[f(X)]$.

Escreva $\tilde{\pi}(x) = c_\pi \pi(x)$, $\tilde{q}(x) = c_q q(x)$ e $\tilde{w}(x) = \frac{\tilde{\pi}(x)}{\tilde{q}(x)} = \frac{c_\pi \pi(x)}{c_q q(x)} = \frac{c_\pi}{c_q} w(x)$. Temos o seguinte:

(i) $E_q[\tilde{w}(X)] = \frac{c_\pi}{c_q} E_q[w(X)] = \frac{c_\pi}{c_q} E_\pi[1] = \frac{c_\pi}{c_q}$;

(ii) $E_q[f(X)\tilde{w}(X)] = \frac{c_\pi}{c_q} E_q[f(X)w(X)] = \frac{c_\pi}{c_q} E_\pi[f(X)] = \frac{c_\pi}{c_q}\mu$.

Assim, segue-se que $\mu = \frac{E_q[f(X)\tilde{w}(X)]}{E_q[\tilde{w}(X)]}$. Isso nos motiva a definição do seguinte estimador:

$$\tilde{\mu} = \frac{\sum_k f(x_k)\tilde{w}(x_k)}{\sum_k \tilde{w}(x_k)},$$

onde $\{x_k\} \overset{iid}{\sim} q(x)$. Em resumo, tem-se o seguinte algoritmo:

Algoritmo 4: Amostragem por Importância Auto-Normalizada

Entrada: $f, \tilde{\pi}, \tilde{q}, N$
Amostre $x_1, \ldots, x_N \sim q(x)$
para $i = 1, \ldots, N$ **faça**
  $\tilde{w}_i \leftarrow \frac{\tilde{\pi}(x_i)}{\tilde{q}(x_i)}$
$\tilde{w} \leftarrow \sum \tilde{w}_i$
$\tilde{z} \leftarrow \sum w_i f(x_i)$
$\tilde{\mu} \leftarrow \tilde{z}/\tilde{w}$
Saída : $\tilde{\mu}$

Observa-se que o estimador é enviesado. Com efeito, considere as variáveis aleatórias $Z = f(X)w(X)$ e $W = w(X)$. Como $\tilde{w}(x) \propto w(x)$, observe que

$$\tilde{\mu} = \frac{\sum_k f(x_k)\tilde{w}(x_k)}{\sum_k \tilde{w}(x_k)} = \frac{\sum_k f(x_k)w(x_k)}{\sum_k w(x_k)} = \frac{\hat{\mu}}{\hat{w}},$$

onde $\hat{\mu} = \frac{1}{N}\sum_k f(x_k)w(x_k)$ e $\hat{w} = \frac{1}{N}\sum_k w(x_k)$ são os estimadores não-enviesados de $E_q[Z] = \mu$ e $E_q[W] = 1$, respectivamente.

Pelo teorema central do limite,

(i) $\hat{\mu} - \mu = O(\frac{\sigma_z}{\sqrt{N}})$ e

(ii) $\hat{w} - 1 = O(\frac{\sigma_w}{\sqrt{N}})$,

onde $\sigma_z = \mathrm{Var}_q[Z]$ e $\sigma_w = \mathrm{Var}_q[W]$, em distribuição, com $N \to \infty$.

Fazendo a expansão de Taylor de $\frac{1}{\hat{w}}$ em torno de $w = 1$, e usando (i) e (ii), tem-se

$$\begin{aligned}
\tilde{\mu} = \frac{\hat{\mu}}{\hat{w}} &= \hat{\mu}\big(1 - (\hat{w} - 1) + (\hat{w} - 1)^2 + O(|\hat{w} - 1|^3))\big) \\
&= \hat{\mu} - \hat{\mu}(\hat{w} - 1) + \hat{\mu}(\hat{w} - 1)^2 + O(\frac{1}{N^{3/2}}) \\
&= \hat{\mu} - (\hat{\mu} - \mu)(\hat{w} - 1) - \mu(\hat{w} - 1) + (\hat{\mu} - \mu)(\hat{w} - 1)^2 + \mu(\hat{w} - 1)^2 \\
&\quad + O(\frac{1}{N^{\frac{3}{2}}}) \\
&= \hat{\mu} - (\hat{\mu} - \mu)(\hat{w} - 1) - \mu(\hat{w} - 1) + \mu(\hat{w} - 1)^2 + O(\frac{1}{N^{\frac{3}{2}}}).
\end{aligned}$$

Assim, usando que $E_q[\hat{\mu}] = \mu$, $E_q[\hat{w}] = 1$, $\mathrm{Cov}_q[\hat{\mu}, \hat{w}] = \frac{1}{N}\mathrm{Cov}_q[Z, W]$ e $\mathrm{Var}_q[\hat{w}] = \frac{1}{N}\mathrm{Var}_q[W]$, segue-se que

$$E_q[\tilde{\mu}] = \mu - \frac{1}{N}\big(\mathrm{Cov}_q[Z, W] - \mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{3/2}}). \tag{1.4}$$

Donde, conclui-se que o estimador $\tilde{\mu}$ é enviesado.

Agora, vamos estimar o erro médio quadrático (MSE) de $\tilde{\mu}$. Para isso, primeiro observe que

$$\begin{aligned} E_q[\tilde{\mu}]^2 &= \big(\mu - \frac{1}{N}\big(\mathrm{Cov}_q[Z, W] - \mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{3/2}})\big)^2 \\ &= \mu^2 - \frac{2\mu}{N}\big(\mathrm{Cov}_q[Z, W] - \mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{\frac{3}{2}}}). \end{aligned} \tag{1.5}$$

Além disso, fazendo novamente a expansão de Taylor de $\frac{1}{\hat{w}}$ em torno de $w = 1$, e usando também que $\hat{\mu}^2 = (\hat{\mu} - \mu + \mu)^2 = \mu^2 + 2\mu(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2$, tem-se

$$\begin{aligned} \tilde{\mu}^2 = \frac{\hat{\mu}^2}{\hat{w}^2} &= \hat{\mu}^2\big(1 - (\hat{w} - 1) + (\hat{w} - 1)^2 + O(\frac{1}{N^{\frac{3}{2}}})\big)^2 \\ &= \hat{\mu}^2\big(1 - 2(\hat{w} - 1) + 3(\hat{w} - 1)^2 + O(\frac{1}{N^{\frac{3}{2}}})\big) \\ &= \big(\mu^2 + 2\mu(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2\big)\big(1 - 2(\hat{w} - 1) + 3(\hat{w} - 1)^2 + O(\frac{1}{N^{\frac{3}{2}}})\big) \\ &= \mu^2(1 - 2(\hat{w} - 1) + 3(\hat{w} - 1)^2) + 2\mu(\hat{\mu} - \mu) - 4\mu(\hat{\mu} - \mu)(\hat{w} - 1) \\ &\quad + (\hat{\mu} - \mu)^2 + O(\frac{1}{N^{\frac{3}{2}}}). \end{aligned}$$

Assim,

$$\begin{aligned} E_q[\tilde{\mu}^2] &= \mu^2 + 3\mu^2\mathrm{Var}[\hat{w}] - 4\mu\mathrm{Cov}_q[\hat{\mu}, \hat{w}] + \mathrm{Var}_q[\hat{\mu}] + O(\frac{1}{N^{\frac{3}{2}}}) \\ &= \mathrm{Var}_q[\hat{\mu}] + \mu^2 + \frac{3\mu^2}{N}\mathrm{Var}_q[W] - \frac{4\mu}{N}\mathrm{Cov}_q[Z, W] + O(\frac{1}{N^{\frac{3}{2}}}) \\ &= \mathrm{Var}_q[\hat{\mu}] + \mu^2 - \frac{\mu}{N}\big(4\mathrm{Cov}_q[Z, W] - 3\mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{\frac{3}{2}}}). \end{aligned} \tag{1.6}$$

Logo, usando (1.4), (1.5) e (1.6), o erro médio quadrado de $\tilde{\mu}$ é dada por

$$\begin{aligned} MSE_q(\tilde{\mu}) = \mathrm{Var}_q[\tilde{\mu}] + (E_q[\tilde{\mu}] - \mu)^2 &= E_q[\tilde{\mu}^2] - E[\tilde{\mu}]^2 + O(\frac{1}{N^2}) \\ &= \mathrm{Var}_q[\hat{\mu}] - \frac{\mu}{N}\big(2\mathrm{Cov}_q[Z, W] - \mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{\frac{3}{2}}}). \end{aligned} \tag{1.7}$$

Com isso, ainda que seja possível calcular o estimador $\hat{\mu}$, este pode ser preterido pelo estimador enviesado $\tilde{\mu}$. Para isso, basta valer a regra de ouro:

$$2\text{Cov}_q[Z, W] - \mu\text{Var}_q[W] > 0.$$

1.4.2 Tamanho Efetivo da Amostra em Amostragem por Importância

Com a má escolha da distribuição proposta $q(x)$, muitos pesos $w_k = \frac{\pi(x_k)}{q(x_k)}$ ficam próximos de zero e então poucos deles realmente são significantes. Com isso, tem-se alta variabilidade nos pesos $w(X)$, comprometendo a qualidade da estimador $\hat{\mu} = \frac{1}{N}\sum_k f(x_k)w(x_k)$. Isso pode ser visto no exemplo anterior (cálculo do volume da bola unitária com amostragem por importância). Com efeito, mantendo as mesmas notações, assumindo d grande, os pesos $w(X) = \frac{\mathbb{1}_B(X)}{q(X)}$, com $X \sim q(X)$, satisfazem

$$w(x) = \frac{\mathbb{1}_B(x)}{q(x)} = \sqrt{2\pi}\sigma^n e^{\frac{1}{2\sigma^2}\|x\|^2}\mathbb{1}_B(x) \approx \sqrt{2\pi}\sigma^n e^{\frac{1}{2}(d+\sqrt{2d}Z)}\mathbb{1}_B(x),$$

em distribuição. Assim, tomando $\sigma^2 = (d + \sqrt{2d})^{-1}$ e $N = 10^6$ (lembre-se que ambos N e σ^2 foram usados na tabela comparativa acima), temos que $w_{\max} = \sqrt{2\pi}\sigma^n e^{\frac{1}{2}(d+\sqrt{2d})}$, visto que existirão amostras com $Z = 1$ e $w_{\text{mediana}} = \sqrt{2\pi}\sigma^n e^{\frac{d}{2}}$, visto que a mediana dos $\|x_{k's}\|^2$ ocorre com $Z = 0$. Portanto, tem-se a razão

$$\frac{w_{\max}}{w_{\text{mediana}}} = e^{\sqrt{2d}}.$$

Essa igualdade aparece no livro do Mackay [12], exatamente na mesma discussão da variabilidade dos pesos $w_{k's}$.

Com isso, se faz necessário estabelecer uma medida de eficiência na amostragem por importância. Para isso, Kong [8] definiu o tamanho efetivo da amostra (ESS) por

$$\text{MSE}_q[\tilde{\mu}] = \frac{1}{\text{ESS}}\text{Var}_\pi[f(X)].$$

Considere o estimador não-enviesado $\bar{\mu} = \frac{1}{M}\sum_{k=1}^{M} f(y_k)$, tomado sobre M amostras $\{y_k\} \overset{iid}{\sim} \pi(x)$. Como

$$\text{Var}_\pi[\bar{\mu}] = \frac{1}{M}\text{Var}_\pi[f(X)] = \frac{\text{ESS}}{M}\text{MSE}_q[\tilde{\mu}],$$

podemos interpretar ESS como o número de amostras iid's de $\pi(x)$ necessário para se produzir o mesmo efeito de $\mathrm{MSE}_q[\tilde{\mu}]$.

Teorema 3. *Vale que*

$$\mathrm{ESS} \approx \frac{N}{1+\mathrm{Var}_q[w(X)]},$$

desde que $\frac{\epsilon}{E_\pi[W]} = E_\pi[(f-\mu)^2(\frac{W}{E_\pi[W]}-1)]$ *seja pequeno.*

Demonstração: Vamos manter as mesmas notações usadas na subseção anterior. Como $\mathrm{Var}_q[\hat{\mu}] = \frac{1}{N}\mathrm{Var}_q[Z]$, usando (1.7), tem-se

$$\begin{aligned}\mathrm{MSE}[\tilde{\mu}] &= \mathrm{Var}_q[\hat{\mu}] - \frac{\mu}{N}\big(2\mathrm{Cov}_q[Z,W] - \mu\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{\frac{3}{2}}}) \\ &= \frac{1}{N}\big(\mathrm{Var}_q[Z] - 2\mu\mathrm{Cov}_q[Z,W] + \mu^2\mathrm{Var}_q[W]\big) + O(\frac{1}{N^{\frac{3}{2}}}). \qquad (1.8)\end{aligned}$$

Considere $Y = f(X)$. Como $E_\pi[Y] = E_q[Z] = \mu$ e $E_q[W] = 1$,

$$\begin{aligned}\mathrm{Cov}_q[Z,W] &= E_q[ZW] - E_q[Z]E_q[W] = E_q[YW^2] - \mu \\ &= E_\pi[YW] - \mu = \mathrm{Cov}_\pi[Y,W] + E_\pi[Y]E_\pi[W] - \mu \\ &= \mathrm{Cov}_\pi[Y,W] + \mu E_\pi[W] - \mu. \qquad (1.9)\end{aligned}$$

Além disso, como $Z = YW$, segue-se

$$\mathrm{Var}_q[Z] = \mathrm{Var}_q[YW] = E_q[Y^2W^2] - \mu^2 = E_\pi[Y^2W] - \mu^2.$$

Agora, observe que

$$\begin{aligned}Y^2W &= \big(Y-\mu+\mu\big)^2\big(W - E_\pi[W] + E_\pi[W]\big) \\ &= \big((Y-\mu)^2 + 2\mu(Y-\mu) + \mu^2\big)\big(W - E_\pi[W] + E_\pi[W]\big) \\ &= (Y-\mu)^2(W - E_\pi[W]) + E_\pi[W](Y-\mu)^2 + 2\mu(Y-\mu)(W - E_\pi[W]) \\ &\quad + 2\mu E_\pi[W](Y-\mu) + \mu^2(W - E_\pi[W]) + \mu^2 E_\pi[W].\end{aligned}$$

E como $\mu = E_\pi[Y]$, aplicando a média em ambos lados,

$$E_\pi[Y^2W] = \epsilon + E_\pi[W]\mathrm{Var}_\pi[Y] + 2\mu\mathrm{Cov}_\pi[Y,W] + \mu^2 E_\pi[W],$$

onde $\epsilon = E_\pi[(Y-\mu)^2(W-E_\pi[W])]$. Donde,

$$\mathrm{Var}_q[Z] = \epsilon + E_\pi[W]\mathrm{Var}_\pi[Y] + 2\mu\mathrm{Cov}_\pi[Y,W] + \mu^2 E_\pi[W] - \mu^2. \qquad (1.10)$$

Aplicando (1.9) e (1.10) em (1.8), tem-se

$$\begin{aligned} N\,\mathrm{MSE}[\tilde{\mu}] &= \epsilon + E_\pi[W]\mathrm{Var}_\pi[Y] + 2\mu\mathrm{Cov}_\pi[Y,W] + \mu^2 E_\pi[W] - \mu^2 \\ &\quad - 2\mu\mathrm{Cov}_\pi[Y,W] - 2\mu^2 E_\pi[W] + 2\mu^2 + \mu^2\mathrm{Var}_q[W] + O(\frac{1}{\sqrt{N}}) \\ &= \epsilon + E_\pi[W]\mathrm{Var}_\pi[Y] - \mu^2 E_\pi[W] + \mu^2 + \mu^2\mathrm{Var}_q[W] + O(\frac{1}{\sqrt{N}}). \end{aligned}$$

E como $E_\pi[W] = E_q[W^2] = \mathrm{Var}_q[W] + 1$, tem-se

$$\mathrm{MSE}[\tilde{\mu}] = \frac{\mathrm{Var}_q[W]+1}{N}\big(\mathrm{Var}_\pi[Y] + \delta\big) + O(\frac{1}{N^{\frac{3}{2}}}),$$

onde o erro $\delta = \frac{\epsilon}{E_\pi[W]} = E_\pi[(f-\mu)^2(\frac{W}{E_\pi[W]} - 1)]$ é assumido pequeno. □

1.4.3 Amostragem por Importância Adaptativa

Lembrando as seções anteriores, a Amostragem por Importância busca resolver o problema de integração

$$\mu = \int f(x)\pi(x)dx = \mathbb{E}_\pi[f(X)]$$

através da aproximação (no caso normalizado) pelo estimador

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)w(x_i),$$

onde $x_i \sim q$ e $w(x) = \frac{\pi(x)}{q(x)}$. É desejável reduzir a variância desse estimador através de uma boa escolha de q, que tal como comentado anteriormente, deve ter um shape próximo ao de $|f(x)|\pi(x)$. Uma opção de abordagem na busca de uma boa distribuição proposta q é definir um modelo $\mathcal{M}$ sobre o qual a distribuição pode variar.

Consideremos $q(x \mid \theta)$ variando em θ sobre um modelo $\mathcal{M}$. Então, fixando n como o tamanho da amostra, podemos escrever o estimador como

$$\hat{\mu}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w(x_i \mid \theta)f(x_i), \text{ com } w(x \mid \theta) = \frac{\pi(x)}{q(x \mid \theta)},$$

onde para cada i, $x_i \sim q(x \mid \theta)$. Seguimos para o cálculo da variância do estimador:

$$\mathbb{V}_{q|\theta}(\hat{\mu}(\theta)) = \mathbb{E}_{q|\theta}\left[\left(\hat{\mu} - \mathbb{E}_{q|\theta}[\hat{\mu}(\theta)]\right)^2\right] = \mathbb{E}_{q|\theta}\left[(\hat{\mu} - \mu)^2\right].$$

Lembrando que $\hat{\mu}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_{q|\theta}\left(f(x_i)w(x_i \mid \theta)\right)$,

$$\mathbb{V}_{q|\theta}\left(\hat{\mu}(\theta)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_{q|\theta}\left(f(x_i)w(x_i \mid \theta)\right).$$

Calculando para apenas um termo,

$$\begin{aligned} \mathbb{V}_{q|\theta}\left(f(X)w(x \mid \theta)\right) &= \mathbb{E}_{q|\theta}\left[(f(X)w(X \mid \theta) - \mu)^2\right] \\ = \mathbb{E}_{q|\theta}\left[f(X)^2 w(X \mid \theta)^2\right] - \mu^2 &= \mathbb{E}_{\pi}\left[f(X)^2 w(x \mid \theta)\right] - \mu^2. \end{aligned}$$

Uma vez que buscamos minimizar a variância a respeito de θ, faz sentido tomar a derivada desta por θ. Então,

$$\begin{aligned} D(\theta) &= N\frac{\partial}{\partial\theta}\mathbb{V}_{q|\theta}(\hat{\mu}(\theta)) = \frac{\partial}{\partial\theta}\mathbb{V}_{q|\theta}(f(X)w(X \mid \theta)) \\ &= \frac{\partial}{\partial\theta}\mathbb{E}_{\pi}\left[f(X)^2 w(X \mid \theta)\right] = \mathbb{E}_{\pi}\left[f(X)^2\frac{\partial}{\partial\theta}w(x \mid \theta)\right] \\ &= \mathbb{E}_{q|\theta}\left[f(X)^2 w(x \mid \theta)\frac{\partial}{\partial\theta}w(x \mid \theta)\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2 w(x_i \mid \theta)\frac{\partial}{\partial\theta}w(x_i \mid \theta). \end{aligned}$$

Agora que sabemos a derivada da variância com relação ao parâmetro do modelo, θ, podemos, a cada iteração, fazer com que θ esteja cada vez mais próximo do ponto de mínimo dessa variância, através de algum algoritmo iterativo de otimização. O exemplo mais simples é usar o método do gradiente descendente:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{n}\sum_{i=1}^{n} f(x_i)^2 w(x_i \mid \theta_t)\frac{\partial w(x_i \mid \theta_t}{\partial\theta_t}.$$

Capítulo 2

MCMC

No início de 2000, a revista Computing in Science and Engineering publicou o artigo "10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century", que, em tradução livre, quer dizer "10 algoritmos com a maior influência no desenvolvimento e na prática de ciência e engenharia no século 20". No topo da lista, figurava o método de simulação Metropolis para Monte Carlo. O artigo ressalta que o método é de grande valor em problemas de dimensão alta, pois com poucas amostras, quando comparado com métodos numéricos, esse algoritmo apresenta resultados próximos dos verdadeiros.

Os métodos de simulação MCMC (Markov Chain Monte Carlo), ou seja, Monte Carlo por cadeias de Markov, consistem em algoritmos de simulação Monte Carlo que usam valores de cadeias de Markov. O uso de um processo estocástico como uma cadeia de Markov pode ter vantagens diante da dimensionalidade do problema e da dificuldade de simular em certas regiões do espaço amostral. Os algoritmos de Metropolis, Metropolis-Hastings e Monte Carlo Hamiltoniano são alguns dos métodos Monte Carlo que pertencem a essa categoria, muito relevante atualmente.

No fim de 2014, dentre os periódicos da Wiley, foi publicado o artigo "Why the Monte Carlo method is so important today" (Kroese et al.), que em tradução livre, seria "Por que o método Monte Carlo é tão importante hoje". O artigo cita engenharia industrial e pesquisa operacional, economia e finanças e estatística computacional como algumas das áreas de aplicação. De fato, Monte Carlo segue uma importante classe de algoritmos ainda nos

dias atuais. Vejamos a teoria por trás de um dos principais representantes dessa classe.

2.1 Cadeias de Markov

2.1.1 Definição

Uma cadeia de Markov é uma sequência de variáveis aleatórias indexadas por um índice temporal que satisfazem uma certa condição de dependência, de forma que, uma vez conhecido o valor de todas as variáveis até certo tempo, o valor da próxima variável depende apenas do valor mais recente. Vale lembrar que esse é um tipo de processo estocástico, uma vez que toda coleção de variáveis aleatórias indexadas pelo tempo é um processo estocástico. Vejamos a definição formal:

Definição 1. *Seja* $(X_0, X_1, \ldots, X_t, ...)$ *uma sequência de variáveis aleatórias definidas em um mesmo espaço amostral* $\mathcal{X}$*. Dizemos que* $(X_t)_{t\in\mathbf{N}}$ *é uma Cadeia de Markov se*

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t),$$

caso $\mathcal{X}$ *seja finito. Caso* $\mathcal{X}$ *seja infinito, a exigência para a definição passa a ser*

$$\mathbb{P}(X_{t+1} \in A \mid X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A \mid X_t = x_t),$$

onde $A \subset \mathcal{X}$*.*

Ou seja, se a probabilidade do próximo estado, condicionada a toda a trajetória até o estado atual depende unicamente do estado atual. Por se tratar da mesma coisa, estaremos sempre omitindo o primeiro termo daqui pra frente, usando apenas $\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t)$ para a chamada probabilidade de transição da cadeia.

Exemplo 3. *Em uma cidade hipotética, a previsão de chuva pode ser reduzida a hipóteses mais simples. Se em determinado dia está chovendo, no dia seguinte há uma probabilidade de* 60% *de continuar chovendo, enquanto há* 40% *de chance do dia seguinte ser ensolarado. Entretanto, caso esteja ensolarado em tal determinado dia, as chances de chuva e sol para o dia seguinte mudam. Uma vez que faz sol, no dia seguinte há* 80% *de chance de fazer sol ainda, sendo* 20% *de chance de chuva.*

Vamos modelar esse problema. Estamos numerando os dias pelo índice d a partir de alguma data arbitrária, e definimos a variável aleatória

$$X_d = \begin{cases} 1, \text{ se faz sol no dia } d. \\ 0, \text{ se chove no dia } d. \end{cases}$$

Se já temos como representar o clima em certo dia d com uma variável X_d, podemos descrever o tempo ao longo de muitos dias por um vetor de tais variáveis. Escrevemos

$$\vec{X_d} = (X_0, X_1, \ldots, X_d).$$

Essas são as variáveis aleatórias. Quando tomam valores, denotamos $\vec{X_d} = \vec{x_d}$, com $\vec{x_d} = (x_0, x_1, \ldots, x_d)$, para representar que $(x_0, X_1, \ldots, X_d)$
$= (x_0, x_1, \ldots, x_d)$.

Conforme a história contada sobre a cidade fictícia, a probabilidade de chuva no próximo dia, dado todo o histórico, depende apenas do clima do dia atual. Podemos escrever

$$\begin{aligned} \mathbb{P}(X_{d+1} = y \mid \vec{X_d} = \vec{x_d}) &= \mathbb{P}(X_{d+1} = y \mid X_0 = x_0, \ldots, X_d = x_d) \\ &= \mathbb{P}(X_{d+1} = y \mid X_d = x_d), \end{aligned}$$

o que bate exatamente com as exigências para que a sequência seja uma cadeia de Markov. Perceba que a probabilidade do clima do próximo dia tampouco depende de quantos dias se passaram desde o início da sequência: Independe se ontem fez chuva ou sol, pois se hoje há chuva, a probabilidade de que amanhã faça sol depende apenas do dia de hoje. Então, podemos representar essa probabilidade apenas em função dos valores do dia atual e do próximo, com

$$P(x, y) := \mathbb{P}(X_{d+1} = y \mid X_d = x).$$

Percebamos que aqui também não importa o valor de d, vale a equação para qualquer que seja o dia, quando tratamos do dia seguinte. Essa característica, de que a probabilidade não depende do índice tem uma definição própria, que em breve trataremos. Agora que a expressão do problema está simplificada, escrevemos os valores

$$P(x, y) = \begin{cases} 0,6, \text{ se } x = 0 \text{ e } y = 1 \\ 0,4, \text{ se } x = 0 \text{ e } y = 0 \\ 0,8, \text{ se } x = 1 \text{ e } y = 1 \\ 0,2, \text{ se } x = 1 \text{ e } y = 0 \end{cases}.$$

Nesse exemplo, o espaço amostral é $\mathcal{X} = \{0, 1\}$, pois são os únicos valores possíveis. Qualquer outro valor teria probabilidade 0.

Uma vez que nosso espaço amostral é finito, podemos representar a cadeia de Markov através de um grafo, conforme na figura 2.1.

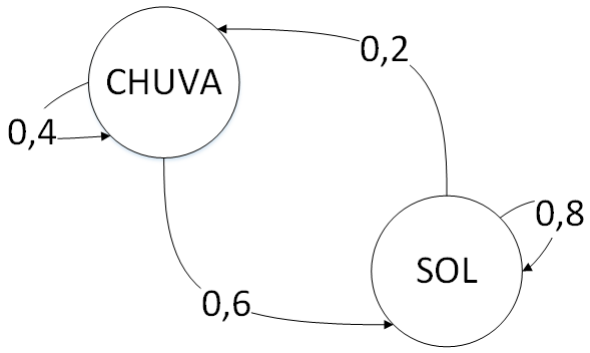

Figura 2.1: Grafo direcionado da cadeia de Markov.

Podemos também tratar a função P como uma matriz, de entradas x e y:

$$P = \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix} = \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix},$$

de tal forma que $P(0, 1)$ retorna o valor da matriz na primeira linha e segunda coluna (e por aí vai).

E se quisermos ir mais além no estudo de probabilidades? Supondo que estejamos interessados em, sabendo que hoje faz sol, calcular a chance de que faça sol depois de amanhã, o problema é, portanto, descrito como

$$P(X_2 = 1|X_0 = 1) = \ ?$$

Como pouco importa a numeração dos dias desde que fixado o espaçamento entre estes, tomo o dia de hoje como 0 e portanto depois de amanhã, daqui a dois dias, será o dia 2. Da Lei da Probabilidade Total, sabemos que

$$\begin{aligned} & P(X_2 = 1|X_0 = 1) \\ = \ & P(X_2 = 1|X_1 = 0)P(X_1 = 0|X_0 = 1) \\ + \ & P(X_2 = 1|X_1 = 1)P(X_1 = 1|X_0 = 1) \\ = \ & 0,6 * 0,2 + 0,8 * 0,8 = 0,12 + 0,64 = 0,76. \end{aligned}$$

Mais do que esse valor numérico, temos algo interessante acontecendo aqui: Para encontrar o valor, multiplicamos a matriz P por ela mesma, e lemos o valor da coluna 1 e linha 1:

$$P(1,1) = (P * P)(1,1) = \left(\begin{pmatrix} 0,4 & 0,6 \\ \mathbf{0,2} & \mathbf{0,8} \end{pmatrix} \begin{pmatrix} 0,4 & \mathbf{0,6} \\ 0,2 & \mathbf{0,8} \end{pmatrix} \right) (1,1) = 0,76.$$

Da mesma forma, podemos concluir que

$$\mathbb{P}(X_2 = y | X_0 = x) = \sum_{z \in \Omega} P(x,z)P(z,y) = (P * P)(x,z),$$

e ainda que

$$\begin{aligned} &\mathbb{P}(X_n = y | X_0 = x) \\ = &\sum_{z_1, z_2, \ldots, z_{n-1} \in \Omega} P(x,z_1)P(z_1,z_2)\ldots P(z_{n-1},y) = (P^n)(x,y). \end{aligned}$$

Assim, faz-se útil a linguagem matricial para a cadeia de Markov. A matriz $P^2 = P * P$ representa portanto as probabilidades de transição de não um, mas dois dias. Dado que hoje temos um certo estado, sol ou chuva, quais são as probabilidades associadas à previsão de tempo para o dia depois de amanhã? A resposta é dada pela matriz

$$P^2 = \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix} \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix} = \begin{pmatrix} 0,28 & 0,72 \\ 0,24 & 0,76 \end{pmatrix}.$$

Fazer esse produto ou também potenciação de matrizes abre um grande leque de possibilidades na área de estudo de cadeias de Markov. O que acontece então se fizermos o produto de um vetor pela matriz? Uma linha de uma matriz de cadeia de Markov representa a distribuição de probabilidade da variável no próximo tempo, condicional a um valor fixo para o tempo presente.

Por exemplo, a primeira linha da matriz deste exemplo contém a função de probabilidade $\mathbb{P}(X_{t+1} = \cdot \mid X_t = 0)$,

$$\mathbb{P}(X_{t+1} = 0 \mid X_t = 0) = 0,4 \text{ e } \mathbb{P}(X_{t+1} = 1 \mid X_t = 0) = 0,6.$$

Por isso, todo vetor linha que compõe a matriz deve somar 1 e ter valores maiores ou iguais a 0. Interpretando isso como uma notação, podemos associar uma probabilidade inicial para a cadeia de Markov. Tudo o que foi dito

até agora sobre cadeias de Markov diz respeito a probabilidades condicionais, mas o que acontece se quisermos obter $\mathbb{P}(X_0 = 1)$ ou $\mathbb{P}(X_3 = 1)$? Voltando ao exemplo, como estamos estudando o clima a partir do dia 0, podemos não saber também se faz sol ou chuva nesse dia, e começar associando uma probabilidade inicial.

Fazemos a hipótese de que

$$\mathbb{P}(X_0 = 0) = 0,3 \text{ e } \mathbb{P}(X_0 = 1) = 0,7.$$

Essa hipótese não interfere em qualquer probabilidade condicional, e portanto não faz com que a sequência de variáveis aleatórias deixe de ser uma cadeia de Markov. Melhor ainda, podemos agora falar, uma vez indicada essa probabilidade inicial, das probabilidades em cada tempo, sem estar se condicionando ao estado em um ou outro determinado tempo.

Vejamos o cálculo então para os tempos 1 e 2 da probabilidade de chuva, a partir dessa probabilidade inicial:

$$\begin{aligned}
\mathbb{P}(X_1 = 0) &= \mathbb{P}(X_1 = 0 \mid X_0 = 0)\mathbb{P}(X_0 = 0) + \mathbb{P}(X_1 = 0 \mid X_0 = 1)\mathbb{P}(X_0 = 1) \\
&= 0,4 * 0,3 + 0,2 * 0,7 = 0,12 + 0,14 = 0,26 \\
\mathbb{P}(X_2 = 0) &= \mathbb{P}(X_2 = 0 \mid X_1 = 0)\mathbb{P}(X_1 = 0) + \mathbb{P}(X_2 = 0 \mid X_1 = 1)\mathbb{P}(X_1 = 1) \\
&= 0,4 * 0,26 + 0,2 * (1 - \mathbb{P}(X_1 = 0)) = 0,4 * 0,26 + 0,2 * 0,74 \\
&= 0,104 + 0,148 = 0,252
\end{aligned}$$

Aqui também a forma matricial nos favorecerá. Denotamos por τ um vetor linha que receberá como entradas os valores da função de probabilidade inicial. Então,

$$\tau = \begin{pmatrix} 0,3 & 0,7 \end{pmatrix}.$$

Com isso, podemos fazer o produto de τ por P para obter o vetor linha com a função de probabilidade $\mathbb{P}(X_1 = \cdot)$ (perceba em destaque o valor já calculado anteriormente, a probabilidade de chuva):

$$\tau P = \begin{pmatrix} \mathbf{0,3} & \mathbf{0,7} \end{pmatrix} \begin{pmatrix} \mathbf{0,4} & 0,6 \\ \mathbf{0,2} & 0,8 \end{pmatrix} = \begin{pmatrix} \mathbf{0,26} & 0,74 \end{pmatrix}.$$

Da mesma forma, podemos calcular o vetor com as probabilidades no tempo 2,

$$\begin{aligned}
\tau P^2 &= \begin{pmatrix} 0,3 & 0,7 \end{pmatrix} \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix} \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0,26} & \mathbf{0,74} \end{pmatrix} \begin{pmatrix} \mathbf{0,4} & 0,6 \\ \mathbf{0,2} & 0,8 \end{pmatrix} = \begin{pmatrix} \mathbf{0,252} & 0,748 \end{pmatrix}.
\end{aligned}$$

O mesmo vale para outros tempos, por conta da mesma operação que se faz através do teorema da probabilidade total para calcular probabilidades condicionais a múltiplos tempos: Também a multiplicação vetorial-matricial aqui nos dá o resultado, de forma a, uma vez suposto que $X_0 \sim \tau$, termos que $X_t \sim \tau P^t$.

2.1.2 Propriedades

Para estudarmos mais a fundo as cadeias de Markov, algumas definições serão úteis:

Definição 2. *Seja (X_t) uma cadeia de Markov tomando valores em $\mathcal{X}$. Dizemos que $\mathcal{X}$ é o espaço de estados da cadeia de Markov (X_t), e chamamos um elemento desse espaço de estados, $x \in \mathcal{X}$, de estado. Caso $\mathcal{X}$ seja finito, dizemos que a cadeia tem espaço de estados finito. Caso tenha infinitos possíveis valores, dizemos que a cadeia tem espaço de estados infinito.*

O estudo de cadeias de Markov varia de acordo com o fato da cadeia ter finitos estados ou não. A princípio, faremos algumas provas para espaço de estados finito.

A próxima definição classifica a dependência das probabilidades da cadeia de transição ao tempo decorrido.

Definição 3. *Seja (X_t) uma cadeia de Markov com espaço de estados $\mathcal{X}$. Dizemos que a cadeia de Markov é homogênea (ou a tempo homogêneo) se*

$$\mathbb{P}(X_{t+1} = y \mid X_t = x) = \mathbb{P}(X_{s+1} = y \mid X_s = x),\ \forall x, y \in \mathcal{X},\ e\ \forall s, t \geq 0.$$

Se a cadeia de Markov não for homogênea, dizemos que é não-homogênea. A equação acima é escrita considerando um espaço de estados finito, mas uma equação análoga serve para uma cadeia em espaço de estados infinito.

Para o estudo de alguns métodos de MCMC, como Metropolis-Hastings, apenas as cadeias homogêneas nos interessam. Aproveitando isso, podemos simplificar a notação com a seguinte definição.

Definição 4. *Seja (X_t) uma cadeia de Markov homogênea com espaço de estados $\mathcal{X}$. Se $\mathcal{X}$ for finito, denotamos*

$$P(x, y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$$

como a matriz de transição da cadeia, usando P ou outra letra maiúscula. Se $\mathcal{X}$ for infinito e $X_{t+1} \mid X_t = x_t$ for uma variável aleatória contínua ou mista, denotamos $K(x, y)$ como o núcleo (ou kernel) de transição, onde

$$\mathbb{P}(X_{t+1} \in A \mid X_t = x) = \int_A K(x, y)dy,$$

podendo-se usar outra letra maiúscula para representar o mesmo.

Daqui em diante, a teoria será desenvolvida considerando espaço de estados finito.

No exemplo visto anteriormente, independente de em que estado estivesse a cadeia, sempre seria possível chegar posteriormente a qualquer outro estado, em algum momento. No contexto, por mais que viesse um dia de chuva, o sol voltaria a aparecer no futuro. Essa característica não é comum a todo tipo de cadeia de Markov. Vejamos o exemplo abaixo:

Exemplo 4 (Ruína do Jogador)**.** *Um exemplo clássico no estudo e ensino das cadeias de Markov é o "Gambler's Ruin", ou "Ruína do Jogador". Trata-se de uma modelagem de um jogador de um jogo de azar onde se tem a seguinte regra:*

Um jogador tem entre 0 *e* n *moedas em cada rodada do jogo. A cada rodada, ele faz uma aposta simples sobre o lançamento de uma moeda justa, ou seja, com probabilidades iguais de vencer e de perder. Caso vença, o jogador ganhará uma moeda a mais, e caso perca, perderá uma de suas moedas. Caso o jogador alcance a quantia de* n *moedas, ele se retira do jogo, assim como faz caso não tenha já nenhuma moeda, indo à falência.*

Então, podemos modelar, a partir da quantidade atual de moedas acumuladas, a probabilidade do jogador estar com certa quantia após uma rodada. Portanto, denotando por X_t a quantidade de moedas que o jogador tem no tempo t, o espaço de estados é $\mathcal{X} = \{0, 1, \ldots, n\}$. Caso k não seja 0 nem n, temos:

$$\mathbb{P}(X_{t+1} = k + 1 \mid X_t = k) = \frac{1}{2} \text{ e } \mathbb{P}(X_{t+1} = k - 1 \mid X_t = k) = \frac{1}{2}.$$

Caso ele tenha nenhuma moeda ou a quantidade máxima de moedas, temos

$$\mathbb{P}(X_{t+1} = n \mid X_t = n) = 1 \text{ e } \mathbb{P}(X_{t+1} = y \mid X_t = n) = 0, \text{ se } y \neq n,$$
$$\mathbb{P}(X_{t+1} = 0 \mid X_t = 0) = 1 \text{ e } \mathbb{P}(X_{t+1} = y \mid X_t = 0) = 0, \text{ se } y \neq 0.$$

Perceba que essa é uma cadeia de Markov porque, de fato, não importa toda a trajetória anterior dele com relação a posse de moedas, apenas o estado atual interferirá na quantidade de moedas que terá na próximas rodada. Assim, podemos representar a cadeia de Markov pela matriz de transição

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & & \dots & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \ddots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix},$$

onde a primeira linha, tal como a primeira coluna, representa o valor 0 de moedas.

Esse exemplo é útil para discutir diferentes comportamentos das cadeias de Markov, mas aqui o destacamos por ser possível chegar a um ponto em que um dos estados fique inalcançável no futuro. Perceba que o único valor não nulo da primeira linha é o da primeira coluna, de probabilidade 1. Então, uma vez que se esteja sem moedas, no próximo estado com probabilidade 1 se permanece sem moedas. Além disso, independente de quanto tempo passe, quantas rodadas aconteçam, o jogador seguirá sem moedas. Vemos uma ilustração do exemplo na figura 2.2.

Então, dado um número natural $t > 0$ qualquer, e $y \in \mathcal{X}$ tal que $y \neq 0$, terei sempre $P^t(0, y) = 0$. É impossível portanto, ir do estado 0 para o estado y, em qualquer tempo, como se não fosse mais possível sair desse estado. O mesmo vale para sair do estado n nesse exemplo. Esse tipo de cadeia tem portanto uma característica diferente da cadeia do Exemplo 3, nos induzindo a outra definição.

Definição 5. *Uma cadeia de Markov (X_t) definida num espaço de estados $\mathcal{X}$, com matriz de transição P, é dita irredutível se*

$$\textit{para todo } x \textit{ e todo } y \in \mathcal{X}, \textit{ existe } t \in \mathbb{N}^* \textit{ tal que } P^t(x, y) > 0.$$

Ou seja, em tempo finito, é possível chegar de um estado a outro qualquer, ao menos com probabilidade positiva. Caso uma cadeia não seja irredutível, ela é dita redutível.

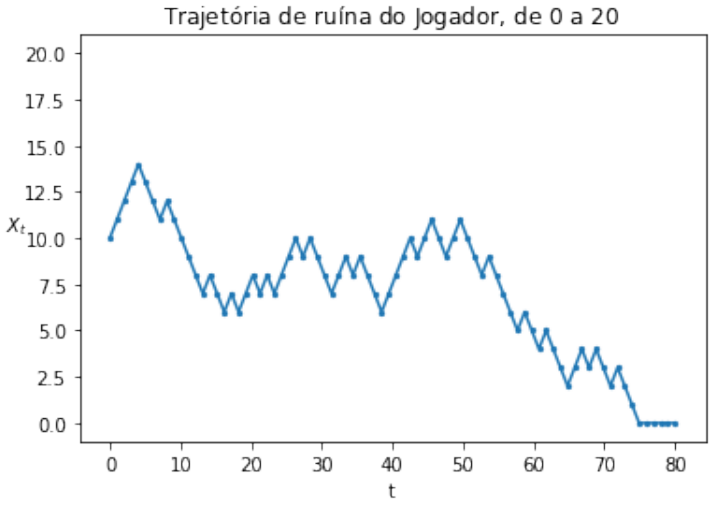

Figura 2.2: Ruína do Jogador.

Esse é o caso do exemplo climático, mas não da Ruína do Jogador. Para a simulação Monte Carlo, será interessante trabalhar com cadeias irredutíveis, uma vez que ficar preso em alguns estados limitaria nossa exploração do espaço de estados.

Outra característica do exemplo de chuva e sol é o fato de que não se fica preso a ciclos. Isso ficará mais claro com o exemplo abaixo, onde isso não funciona tão bem:

Exemplo 5. *Uma cadeia de Markov um tanto simplista é a representada pela matriz*

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

com espaço de estados $\mathcal{X} = \{1, 2\}$.

Por mais que isso possa parecer estranho, essa é uma cadeia de Markov, em que há probabilidades 1 associadas às transições de um estado para o outro. Suponha que saibamos o valor de X_0, e que $X_0 = 1$. Então, a matriz me diz que $\mathbb{P}(1, 2) = 1$. Daí, com probabilidade 1, o próximo estado será 2, e assim por diante. Toda a cadeia, na verdade, já está amarrada, e sabemos que

$$X_0 = 1, X_1 = 2, X_2 = 1, X_3 = 2, \ldots.$$

Percebamos que essa matriz é irredutível, uma vez que sempre é possível estar em qualquer um dois estados em no máximo dois tempos. Mas o exemplo não é muito interessante, e se mostrará não muito útil para métodos MCMC. Mesmo para um tempo t grande, apenas parte do espaço de estados será atingível por vez.

Para evitar que coisas desse tipo ocorram (que em um tempo fixo sempre voltemos a um estado ou um conjunto de estados), uma nova definição surge a seguir:

Definição 6. *Seja (X_t) uma cadeia de Markov irredutível com espaço de estados $\mathcal{X}$. Denotamos*

$$\mathcal{T}(x) = \{t > 0 : P^t(x, x) > 0\},$$

conjunto de tempos de retorno do estado $x \in \mathcal{X}$. Então, dizemos que

$$T(x) = mdc(\mathcal{T}),$$

o máximo divisor comum do conjunto $\mathcal{T}$, é o período do estado x.

A seguinte proposição é um tanto surpreendente:

Proposição 2. *Seja uma cadeia de Markov irredutível num espaço de estados $\mathcal{X}$ finito. Todos os seus estados tem o mesmo período.*

Demonstração: Denotamos P como a matriz de transição da cadeia de Markov. Sejam $x, y \in \mathcal{X}$. Tomamos $T(x)$ como o período de x e $T(y)$ como o período y. Suponhamos, sem perda de generalidade, que $T(x) \leq T(y)$. Como a cadeia é irredutível, existem $r, s > 0$ tais que $P^r(x, y) > 0$ e $P^s(y, x) > 0$. Então, $r + s \in \mathcal{T}(x)$ e $s + r \in \mathcal{T}(y)$. Então, $T(x)$ e $T(y)$ dividem $r + s$.

Tomo $t \in \mathcal{T}(x)$ qualquer. Daí, $P^t(x, x) > 0$. Com isso,

$$\begin{aligned} P^{r+s+t}(y, y) &= \sum_{z \in \mathcal{X}} \sum_{w \in \mathcal{X}} P^s(y, z) P^t(z, w) P^r(w, y) \\ &\geq P^s(y, x) P^t(x, x) P^r(x, y) > 0. \end{aligned}$$

Então, $T(y)$ divide $r + s + t$. Como sabemos já que $T(y)$ divide $r + s$, concluímos que $T(y)$ divide t. Daí, $T(y)$ é um divisor comum de todos os elementos de $\mathcal{T}(x)$. Como $T(y) \geq T(x)$, que por sua vez é máximo divisor

comum de $\mathcal{T}(x)$, temos que $T(y) = T(x)$. Então, de fato x e y tem períodos iguais. □

Uma vez que todos os estados tem o mesmo período, o período passa a ser uma característica da cadeia de Markov como um todo:

Definição 7. *Defino o período de uma cadeia de Markov irredutível como o período de qualquer um de seus estados possíveis. Dizemos que uma cadeia é aperiódica se o seu período for igual a* 1. *Caso contrário, a cadeia é dita periódica.*

No caso do Exemplo 5, a cadeia tinha um período igual a dois. Como pudemos observar, os tempos possíveis de retorno ao estado 1 partindo do próprio eram $2, 4, 6, \ldots$, sendo assim uma cadeia periódica. Já no Exemplo 3, sempre era possível retornar em apenas um passo para o estado atual. Nesse caso, portanto, a cadeia tinha período 1, e pode ser dita aperiódica.

Uma das vantagens já apontadas de se trabalhar com matrizes na representação de cadeias de Markov é a possibilidade de se analisar a evolução da cadeia a longo prazo. Voltando ao Exemplo 3, vejamos como ficam calculadas algumas potências da matriz de transição:

$$P = \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix}, P^2 = \begin{pmatrix} 0,28 & 0,72 \\ 0,24 & 0,26 \end{pmatrix},$$
$$P^3 = \begin{pmatrix} 0,2512 & 0,7488 \\ 0,2496 & 0,7504 \end{pmatrix}, P^5 = \begin{pmatrix} 0,25 & 0,75 \\ 0,25 & 0,75 \end{pmatrix}.$$

A partir da quinta potência, todas as outras são iguais a esta. De fato, uma vez que as colunas de P^5 tem todos os valores iguais e as linhas de P somam 1, o produto não muda a matriz potência.

O que nos revela com essa conta? Que a longo prazo (ou nem tão longo prazo, nesse caso), não importa mais de que estado a cadeia tenha começado, as probabilidades de estar em um estado ou outro são iguais quando decorridas muitas iterações. É como se toda a cadeia tivesse "convergido" para apenas uma distribuição, independente da condição inicial. Isso será melhor definido em breve.

Já que obtemos linhas iguais na matriz da cadeia de Markov, o que acontece se associarmos essa linha como probabilidade inicial da cadeia? Remontando ao exemplo inicial, tomamos

$$\tau = \begin{pmatrix} 0,25 & 0,75 \end{pmatrix}$$

e supomos que $\mathbb{P}(X_0 = \cdot) = \tau(\cdot)$. Então, ao calcular as probabilidades de X_1 segundo essa hipótese, temos

$$\tau P = \begin{pmatrix} 0,25 & 0,75 \end{pmatrix} \begin{pmatrix} 0,4 & 0,6 \\ 0,2 & 0,8 \end{pmatrix} = \begin{pmatrix} 0,25 & 0,75 \end{pmatrix}.$$

Não só estamos multiplicando uma linha que soma 1 pela matriz e encontrando o mesmo vetor linha, como também estamos multiplicando por uma matriz que não tem linhas iguais (valores iguais em cada coluna). Isso revela uma propriedade importante desse vetor de probabilidades iniciais com relação a essa cadeia de Markov. Uma vez que assumimos que as probabilidades iniciais da cadeia são essas, elas se mantém em qualquer tempo t. Isso nos leva a mais uma definição:

Definição 8. *Seja (X_t) uma cadeia de Markov com matriz de transição P. Dizemos que a distribuição (representada em forma vetorial) τ é estacionária para essa cadeia se*

$$\tau P = \tau.$$

Vetores em geral (não necessariamente de probabilidade, somando 1) para os quais valha essa equação são ditos invariantes com relação à matriz P.

Não é por coincidência que o vetor linha que apareceu repetido na matriz quando iterando uma certa quantidade de vezes seja também estacionário para a mesma cadeia. Também não é coincidência que essa matriz tenha de alguma forma "convergido" para a uma matriz que repetisse o mesmo vetor em todas as suas linhas, uma vez que a matriz representasse uma cadeia irredutível e aperiódica.

Proposição 3. *Toda cadeia de Markov irredutível e aperiódica sobre espaço de estados finito tem distribuição estacionária.*

Demonstração: A demonstração que será feita aqui não é tão intuitiva, mas constrói a distribuição estacionária de interesse.

Seja P a matriz de transição de uma cadeia de Markov em espaço finito, e seja μ a distribuição inicial da mesma cadeia. Definimos, para $n > 0$,

$$\nu_n = \frac{1}{n}(\mu + \mu P + \cdots + \mu P^{n-1}).$$

Afirmação. $|\nu_n P(x) - \nu_n(x)| \leq \frac{2}{n}$.

Omitindo alguns passos, temos que

$$\begin{aligned}
& |\nu_n P(x) - \nu_n(x)| \\
= & \left| \sum_{y\in\mathcal{X}} \nu_n(y)P(y,x) - \nu_n(x) \right| \\
= & \left| \sum_{y\in\mathcal{X}} \left[\frac{1}{n} \sum_{i=1}^{n} \mu P^{i-1}(y) \right] - \left[\frac{1}{n} \sum_{i=1}^{n} \mu P^{i-1}(y) \right] \right| \\
= & \frac{1}{n} \left| \sum_{i=1}^{n} \sum_{z\in\mathcal{X}} \left[\left(\sum_{y\in\mathcal{X}} \mu(z)P^{i-1}(z,y)P(y,x) \right) - \mu(z)P^{i-1}(z,x) \right] \right| \\
= & \frac{1}{n} \left| \sum_{i=1}^{n} \sum_{z\in\mathcal{X}} \left[\mu(z)(P^i(z,x) - P^{i-1}(z,x)) \right] \right| \\
\leq & \frac{1}{n} \sum_{z\in\mathcal{X}} \mu(z) | \sum_{i=1}^{n} (P^i(z,x) - P^{i-1}(z,x))| \\
= & \frac{1}{n} \sum_{z\in\mathcal{X}} \mu(z)|P(z,x) - 1| \leq \frac{1}{n} \sum_{z\in\mathcal{X}} 2\mu(z) = \frac{2}{n}.
\end{aligned}$$

Com isso, temos a desigualdade

$$|\nu_n P(x) - \nu_n(x)| \leq \frac{2}{n}. \tag{2.1}$$

Se tivermos uma função limite para a sequência (ν_n) ou para uma de suas subsequências, essa distribuição será estacionária por consequência da desigualdade acima. Então, seguimos a construção sobre (ν_n).

Temos que

$$\begin{aligned}
\nu_{n+1} &= \frac{1}{n+1} \left(\mu + \mu P + \cdots + \mu P^n \right) \\
&= \frac{1}{n+1} \left(\mu + (\mu + \mu P + \cdots + \mu P^{n-1})P \right) = \frac{1}{n+1}(\mu + n\nu_n P).
\end{aligned}$$

Daí,

$$|\nu_{n+1} - \nu_n| = \left| \frac{1}{n+1}(\mu + n\nu_n P) - \nu_n \right| = \left| \frac{\mu}{n+1} + \frac{n}{n+1}\nu_n P - \nu_n \right|.$$

Uma vez que $\mathcal{X}$ é finito, podemos fazer uma construção sobre cada x, até que consideremos todos os estados. Seja $x \in \mathcal{X}$. Então, $\nu_n(x) \in [0,1]$. Por ser uma sequência num compacto, $(\nu_n(x))_{n\in\mathbb{N}}$ tem alguma subsequência $(\nu_{n_k}(x))_{k\in\mathbb{N}}$ convergente. Por sua vez, para um outro $y \in \mathcal{X}$, a sequência $(\nu_{n_k}(y))_{k\in\mathbb{N}}$ também está no compacto $[0,1]$ e portanto tem subsequência convergente. Fazendo isso progressivamente para todos os estados de $\mathcal{X}$, temos finalmente uma sequência $(m_k)_{k\in\mathbb{N}}$ tal que $(\nu_{m_k}(x))_{k\in\mathbb{N}}$ converge para todo x de $\mathcal{X}$.

Finalmente, definimos

$$\nu(x) := \lim_{k\to\infty} \nu_{m_k}(x).$$

Então, por (2.1), temos que

$$|\nu_{m_k} P(x) - \nu_{m_k}(x)| \leq \frac{2}{m}.$$

Quando m tende a infinito, temos

$$|\nu P(x) - \nu(x)| = 0.$$

Ou seja, ν é distribuição estacionária para P.

□

Proposição 4. *A distribuição estacionária de uma cadeia de Markov irredutível e aperiódica sobre espaço de estados finito é única.*

Demonstração: Seja P a matriz de transição de uma cadeia de Markov irredutível e aperiódica e definida sobre espaço finito. Suponhamos que π_1 e π_2 são distribuições estacionárias para a cadeia. Lembrando que a cadeia está definida sobre um espaço de estados finito, consideremos x tal que

$$\frac{\pi_1(x)}{\pi_2(x)}$$

seja minimal. Se existe y tal que

$$\frac{\pi_1(y)}{\pi_2(y)} > \frac{\pi_1(x)}{\pi_2(x)}$$

e $P^k(y,x) > 0$, para algum $k \in \mathbb{N}$, então desenvolvemos com

$$\begin{aligned}
\frac{\pi_1(x)}{\pi_2(x)} &= \sum_{y\in\mathcal{X}} \frac{\pi_1(y)}{\pi_2(x)} P^k(y,x) = \sum_{y\in\mathcal{X}} \frac{\pi_1(y)}{\pi_2(y)} \frac{\pi_2(y)}{\pi_2(x)} P^k(y,x) \\
&> \sum_{y\in\mathcal{X}} \frac{\pi_1(x)}{\pi_2(x)} \frac{\pi_2(y)}{\pi_2(x)} P^k(y,x) = \frac{\pi_1(x)}{\pi_2(x)} \frac{1}{\pi_2(x)} \sum_{y\in\mathcal{X}} \pi_2(y) P^k(y,x) \\
&= \frac{\pi_1(x)}{\pi_2(x)} \frac{\pi_2(x)}{\pi_2(x)} = \frac{\pi_1(x)}{\pi_2(x)}.
\end{aligned}$$

Então, uma vez que P é irredutível, existe $k \in \mathbb{N}$ tal que $P^k(y,x) > 0$. Daí, para todo $y \in \mathcal{X}$, $\frac{\pi_1(y)}{\pi_2(y)} = \frac{\pi_1(x)}{\pi_2(x)}$. Tendo isso em mente, concluímos que

$$1 = \sum_{y\in\mathcal{X}} \pi_1(y) = \sum_{y\in\mathcal{X}} \frac{\pi_1(x)\pi_2(y)}{\pi_2(x)} = \frac{\pi_1(x)}{\pi_2(x)}.$$

Ou seja, a distribuição estacionária é única. □

O resultado que afirma que há de fato uma convergência da cadeia de Markov à sua distribuição estacionária é citado na subseção 2.2.4. Entretanto, podemos ver uma intuição já aqui.

Seja P a matriz de transição de uma cadeia de Markov em espaço de estados finito. Suponhamos que, para todo $x \in \mathcal{X}$, valha que

$$\lim_{t\to\infty} P^t(x,y) = \pi(y)$$

para todo $y \in \mathcal{X}$.

Então, sendo essa, para cada y, uma convergência de uma sequência de números reais, podemos usar que toda subsequência converge igualmente. Daí,

$$\begin{aligned}
\pi(y) &= \lim_{t\to\infty} P^{t+1}(x,y) = \lim_{t\to\infty} \sum_{z\in\mathcal{X}} P^t(x,z) P(z,y) \\
&= \sum_{z\in\mathcal{X}} P(z,y) \lim_{t\to\infty} P^t(x,z) = \sum_{z\in\mathcal{X}} \pi(z) P(z,y) = \pi P(y).
\end{aligned}$$

Ou seja, π é estacionária.

Com isso, concluímos que se a cadeia convergir para uma distribuição, essa distribuição será a distribuição estacionária.

Outro conceito que será útil mais à frente é o conceito de equilíbrio detalhado.

Definição 9. *Dizemos que uma distribuição de probabilidade μ e uma cadeia de Markov com matriz de transição P satisfazem as equações de equilíbrio detalhado se*

$$\mu(x)P(x,y) = \mu(y)P(y,x), \text{ para todo } x, y \in \mathcal{X}.$$

Quando calculamos um dos termos dessa equação, há uma certa intuição envolvida. Associando a μ a probabilidade de se estar num elemento do espaço de estados inicialmente, e tomando por P a matriz de transição da cadeia de Markov, estamos dizendo que $\mathbb{P}(X_0 = x) = \mu(x)$ e que $\mathbb{P}(X_1 = y \mid X_0 = x) = P(x,y)$. Portanto,

$$\mu(x)P(x,y) = \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \mid X_0 = x) = \mathbb{P}(X_0 = x, X_1 = y).$$

Então, o que está sendo estudado aqui é a probabilidade de se estar num estado x e em seguida estar num estado y. Então, interpreto as equações de equilíbrio detalhado como garantir que "estar em um estado e depois no outro é a mesma coisa que estar no outro e ir para o um".

O seguinte resultado torna essas equações uma ferramenta muito útil.

Proposição 5. *Seja (X_t) uma cadeia de Markov com espaço de estados $\mathcal{X}$ e matriz de transição P e seja π uma função de probabilidade de distribuição definida em $\mathcal{X}$. Se valem as equações de equilíbrio detalhado para esse par, ou seja,*

$$\pi(x)P(x,y) = \pi(y)P(y,x), \forall x, y \in \mathcal{X},$$

então π é a distribuição estacionária da cadeia (X_t).

Demonstração: Primeiramente, uma vez que temos a proposição anterior e tenhamos provado que π é uma distribuição estacionária da cadeia, ela será "a" distribuição estacionária, por essa ser única.

Assumo então, por hipótese, as equações de equilíbrio detalhado para π e P. Para determinar se π é distribuição estacionária para a cadeia, precisamos

verificar o valor de πP, o produto entre o vetor linha da distribuição e a matriz de transição, nesta ordem. Tomamos $y \in \mathcal{X}$. Temos

$$\pi P(y) = \sum_{x \in \mathcal{X}} \pi(x) P(x, y) = \sum_{x \in \mathcal{X}} \pi(y) P(y, x) = \pi(y) \sum_{x \in \mathcal{X}} P(y, x) = \pi(y).$$

A segunda igualdade da linha acima vem das equações que assumimos. Já a última igualdade vem do fato de, fixado y, $P(y, x)$ ser uma distribuição de probabildade e portanto somar 1.

Logo, π é distribuição estacionária para P. □

Definição 10. *Uma cadeia de Markov com matriz de transição P é simétrica se*

$$P(x, y) = P(y, x)$$

para todo x e todo y no espaço de estados.

É bem direto perceber que toda cadeia simétrica em espaço finito tem como distribuição estacionária a distribuição uniforme. Tomando P como matriz de transição de uma cadeia definida sobre um espaço finito $\mathcal{X}$, temos que a distribuição uniforme sobre esse é dada por $\pi(x) = \frac{1}{|\mathcal{X}|}$. Então,

$$\pi(x)P(x, y) = \frac{1}{|\mathcal{X}|} P(x, y) = \frac{1}{|\mathcal{X}|} P(y, x) = \pi(y)P(y, x),$$

quaisquer que sejam x e y em $\mathcal{X}$.

Com isso, visitamos todos os conceitos necessários para entender os algoritmos de Metropolis e Metropolis-Hastings.

2.2 O Algoritmo Metropolis-Hastings

2.2.1 Problema e Proposta

Aqui, todo o estudo sobre cadeias de Markov tem por objetivo nos ambientar a esse objeto que servirá como ferramenta a ser usada em Monte Carlo. Usando dados simulados a partir de cadeias de Markov, muitos problemas são resolvidos pela abordagem de Monte Carlo, por isso o nome, MCMC. Mas onde uma sequência de variáveis pode nos ajudar?

Como foi visto anteriormente, à medida em que a cadeia de Markov evolui com o tempo, em que se acessa X_t para t grande, esta se aproxima da distribuição estacionária. Então, se criarmos uma cadeia de Markov que tenha

como distribuição estacionária nossa distribuição de interesse, ao simular essa cadeia, depois de passada uma certa quantidade de iterações teremos praticamente amostras da distribuição de interesse.

Anteriormente, no estudo das cadeias de Markov, começamos com uma cadeia e então obtemos a sua distribuição estacionária. No uso em MCMC, o caminho é inverso, partimos de uma distribuição estacionária. Criamos essa cadeia de Markov para a distribuição através das equações de equilíbrio detalhado.

Seja π uma função de probabilidade, representando também a distribuição que descreve. Criar diretamente uma cadeia de Markov pode ser complicado, até porque não há intuição de como montar uma matriz de transição que satisfaça as equações. Partimos então de uma cadeia de Markov da qual já saibamos amostrar.

No primeiro caso, partiremos de uma cadeia simétrica. Relembrando, uma cadeia de Markov com matriz de transição P é dita simétrica se

$$P(x, y) = P(y, x), \forall x, y.$$

Tomando portanto uma matriz transição de Markov Q simétrica, desejamos criar uma outra cadeia de Markov nos inspirando no algoritmo da amostragem por rejeição. Definimos a cadeia de Markov através de sua matriz de transição P com

$$P(x, y) = Q(x, y)a(x, y), \text{ se } x \neq y,$$

onde a função a é uma função de aceitação ou não da amostra de Q para a nossa cadeia. Isso quer dizer que amostraremos da cadeia Q, mas aceitaremos ou não esta amostra de acordo com a função a. Caso não haja aceitação, preservaremos o estado atual.

Queremos que as equações de equilíbrio detalhado sejam válidas para a nossa distribuição π de interesse e a cadeia Q, portanto,

$$\pi(x)a(x, y)Q(x, y) = \pi(x)P(x, y) = \pi(y)P(y, x) = \pi(y)a(y, x)Q(y, x).$$

Então, a condição primordial para que a satisfaça a equação acima é

$$\frac{a(x, y)}{a(y, x)} = \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}, \forall x, y.$$

Da mesma forma, tenho

$$\frac{a(y,x)}{a(x,y)} = \frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}, \forall x, y.$$

Como em muitos métodos, é conveniente que a aceitação seja a maior possível. Por outro lado, sendo uma função de aceitação, a taxa de aceitação é uma probabilidade, e deve estar entre 0 e 1. Sabendo que o valor da função de aceitação não pode ser maior que 1, tomamos

$$a(x,y) = \min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}.$$

Então, basta agora verificar as condições de equilíbrio detalhado:

$$\pi(x)P(x,y) = \pi(x)a(x,y)Q(x,y) = \pi(x)\min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}Q(x,y).$$

Supondo, sem perda de generalidade, que $pi(y)Q(y,x) \geq \pi(x)Q(x,y)$, temos que

$$\pi(x)P(x,y) = \pi(x)\min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}Q(x,y) = \pi(x)Q(x,y).$$

Por outro lado, com essa nossa suposição, temos que $a(y,x) = \frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}$. Daí,

$$\pi(y)P(y,x) = \pi(y)\frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}Q(y,x) = \pi(x)Q(x,y).$$

Então, de fato, com a cadeia construída dessa forma, temos as equações de equilíbrio detalhado.

O que foi visto até agora é apenas um vislumbre do método, que veremos mais claramente a seguir. Vale destacar que ainda não foi feita qualquer conta sobre o caso em que uma amostra é rejeitada. Quando supomos que Q é simétrica, a função de aceitação se torna mais simples:

$$a(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

A utilização tanto de uma quanto da outra cadeia para geração de valores aleatórios é explicitada na seção na próxima seção.

2.2.2 Descrição

O método de Metropolis para obter uma amostra de uma distribuição π é feito da seguinte forma:

Primeiro, escolhemos uma cadeia de Markov com transição Q simétrica da qual saibamos amostrar no mesmo espaço de π. Tomamos x_0 nesse mesmo espaço como chute inicial. Definimos $a(x, y) = \frac{\pi(y)}{\pi(x)}$. Depois, repetimos os passos a seguir por uma quantidade n de vezes:

Algoritmo 5: Metropolis

Entrada: π, Q, N, x_0
Definimos $a(x, y) \leftarrow \pi(y)/\pi(x)$
para $k = 1, \ldots, N$ **faça**
  $\tilde{x} \sim Q(x_{k-1}, \cdot)$
  $u \sim \text{Unif}_{[0,1]}$
  se $u \leq a(x_{k-1}, \tilde{x})$ **então**
    $x_k \leftarrow \tilde{x}$
  senão
    $x_k \leftarrow x_{k-1}$
Saída : $x^{(1)}, \ldots, x^{(t)}$

Ao final, obtemos x_n como uma amostra aproximada de π, se n for suficientemente grande.

Vale citar um outro algoritmo, que não exige que Q seja simétrico, o algoritmo de Metropolis-Hastings. Trata-se então de uma versão mais geral do algoritmo de Metropolis. A grande diferença está na definição da função de aceitação

$$a(x, y) = \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)},$$

que contorna a assimetria da cadeia proposta.

Então, os passos do algoritmo são seguidos da mesma forma.

Algoritmo 6: Metropolis-Hastings

Entrada: π, Q, N, x_0
Definimos $a(x,y) \leftarrow (\pi(y)Q(y,x))/(\pi(x)Q(x,y))$
para $k = 1, \ldots, N$ **faça**
  $\tilde{x} \sim Q(x_{k-1}, \cdot)$
  $u \sim \text{Unif}_{[0,1]}$
  se $u \leq a(x_{k-1}, \tilde{x})$ **então**
    $x_k \leftarrow \tilde{x}$
  senão
    $x_k \leftarrow x_{k-1}$
Saída : $x^{(1)}, \ldots, x^{(t)}$

Como comentamos acima, ao seguir um método de Monte Carlo por Cadeia de Markov, como é o caso do Metropolis-Hastings, estamos criando uma cadeia de Markov, com sua transição de Markov correspondente, para simulá-la até que se chegue a amostras da distribuição objetivo. Qual é então o núcleo de transição dessa cadeia de Markov gerada pelo algoritmo? Devemos considerar que é possível permanecer no estado atual numa próxima iteração não só sorteando esse estado pela transição proposta, mas também rejeitando o valor sorteado. Logo, a transição é

$$P(x,y) = Q(x,y)a(x,y) + \sum_{z \in \mathcal{X}} Q(x,z)(1 - a(x,z)),$$

que também pode ser escrita como

$$P(x,y) = Q(x,y)a(x,y) + \mathbb{E}_{Q(x,\cdot)}[1 - a(x,Z)].$$

Tendo uma taxa de aceitação a adequada, a cadeia de Markov acima pode ser usada na geração de valores da distribuição de interesse. Já vimos duas taxas de aceitação possíveis, usadas no método de Metropolis e Metropolis-Hastings. Outras versões, porém, também permitem o funcionamento do algoritmo, fazendo valer as condições de equilíbrio detalhado.

Suponhamos que uma transição P seja tal que $P(x,y) = \pi(y)S(x,y)$, onde S é uma função simétrica. Então,

$$\pi(x)P(x,y) = \pi(x)\pi(y)S(x,y) = \pi(y)\pi(x)S(y,x) = \pi(y)P(y,x),$$

ou seja, as equações de equilíbrio detalhado são preservadas.

Daí, temos um novo critério na construção da cadeia. Queremos $P(x,y) = \pi(y)S(x,y)$, mas construímos a cadeia de transição P através de uma cadeia proposta Q com a fórmula

$$P(x,y) = Q(x,y)a(x,y),$$

onde $a(x,y)$ é a taxa de aceitação, sendo $a(x,y) \leq 1$, desconsiderando o caso em que $y = x$, por não interferir na simetria. Então, temos que

$$a(x,y) = \frac{P(x,y)}{Q(x,y)} = \frac{\pi(y)S(x,y)}{Q(x,y)}.$$

Tomando então, uma taxa de aceitação como acima com S simétrica, vale, para $x \neq y$,

$$P(x,y) = Q(x,y)a(x,y) = Q(x,y)\frac{\pi(y)S(x,y)}{Q(x,y)} = \pi(y)S(x,y).$$

Enquanto isso, para $x = y$,

$$P(x,y) = \pi(y)\frac{P(x,y)}{\pi(y)},$$

onde $\frac{P(x,y)}{\pi(y)}$ é simétrico por ser apenas evaluado em $x = y$. Essa é então, uma condição suficiente para que uma função possa ser usada como taxa de aceitação. Outro exemplo de taxa de aceitação, a partir disso, é portanto a seguinte proposta por Barker em 1965:

$$a(x,y) = \frac{1}{1 + \frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}}.$$

De fato, essa obedece à condição

$$\begin{aligned} \frac{1}{1 + \frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}} &= \frac{1}{\frac{\pi(y)Q(y,x) + \pi(x)Q(x,y)}{\pi(y)Q(y,x)}} \\ &= \frac{\pi(y)}{Q(x,y)} \frac{Q(x,y)Q(y,x)}{\pi(y)Q(y,x) + \pi(x)Q(x,y)}, \end{aligned}$$

onde a segunda fração é função simétrica de x e y.

2.2.3 Tamanho Efetivo da Amostra

Assim como o método da rejeição, o Metropolis-Hastings trabalha com a rejeição de amostras de uma distribuição proposta. Por outro lado, para que as ferramentas de cadeias de Markov sejam aproveitadas, quando há rejeição de valores, o método não só deixa de tomar o novo valor para parte do algoritmo, como também permanece onde está: O valor atual é repetido.

Essa repetição de valores pode levar à alta correlação. As amostras de um método MCMC muitas vezes são geradas eficientemente, porém a custo da independência: Não se pode dizer que as amostras são independentes entre si. Essa falta da independência influencia nos erros de estimação.

Supondo que P seja a transição de Markov de uma cadeia cuja distribuição estacionária seja π, distribuição objetivo em questão no método MCMC, vale que

$$P^t(x, y) \xrightarrow{t\to\infty} \pi(y),$$

em um sentido a ser explicitado mais à frente.

Então, usar esse tipo de método, podemos começar iterando o algoritmo por uma certa quantidade de iterações, a fim de obter a convergência. Essas primeiras amostras são desconsideradas, como que queimadas, através do que se chama de "burn-in". Assim, depois do "burn-in", temos amostras da distribuição π. O estudo da convergência desses métodos está voltado a buscar o valor ideal do "burn-in".

Assumo $x^{(1)}, \ldots, x^{(n)}$ gerados de P depois do "burn-in", tendo aproximadamente $X^{(i)} \sim \pi, \forall i \geq 1$. Daí,

$$\begin{aligned}\mathbb{P}(X_2 = x^{(2)}) &= \sum_{z\in\mathcal{X}} \mathbb{P}(X_2 = x^{(2)} \mid X_1 = z)\mathbb{P}(X_1 = z) \\ &= \sum_{z\in\mathcal{X}} \pi(z)P(z, x^{(2)}) = \pi(x^{(2)}),\end{aligned}$$

por π ser estacionária para P. E assim, vale para todo $i \geq 1$.

Essas amostras são tomadas para resolver um problema, por exemplo, de inferência,

$$\mu = \mathbb{E}_\pi[f(X)].$$

Nesse caso, de fato, em cada iteração i após o "burn-in":

$$\mathbb{E}[f(X^{(i)})] = \int_{\mathcal{X}} f(x^{(i)})\pi(x^{(i)})dx^{(i)} = \mathbb{E}_{\pi}[f(X^{(i)}] = \mu.$$

Da mesma forma, a variância sobre cada amostra é a mesma do problema,

$$\sigma^2 = \mathbb{V}_{\pi}(f(X^{(i)})).$$

Obtida a amostra $x = (x^{(1)}, \ldots, x^{(n)})$, a partir de um método MCMC, um estimador para μ é:

$$\bar{f}(x) = \frac{1}{n}(f(x^{(1)}) + \cdots + f(x^{(n)})).$$

Esse estimador é não-enviesado, com $\mathbb{E}[\bar{f}(X)] = \mu$. Entretanto, uma vez que as amostras são correlacionadas, o cálculo da variância exige mais atenção, não sendo apenas a soma das variâncias de cada amostra. De fato, a escolha pelos métodos de MCMC ganha na geração de grandes amostras da distribuição objetivo através de um método que não tenha interferência da dimensão e que combine formas de reduzir a variância, mas perde com relação ao fato das amostras não serem mais IID (independentes e identicamente distribuídas), pela perda da independência.

Na amostragem por importância, as amostras, não sendo da distribuição objetivo, mas tendo pesos em vista para corrigir a estimação, faziam com que o estimador tivesse uma variância maior do que aquela de uma média feita sobre a mesma quantidade de amostras da própria distribuição objetivo. Aqui em MCMC, o uso de amostras correlacionadas faz com que o estimador tenha variância maior do que um que use amostras IID. Tanto num caso quanto no outro, podemos avaliar como seria a variância se alternativamente o estimador usasse amostras IID da distribuição objetivo. Essa comparação dá origem ao conceito de tamanho efetivo da amostra.

Primeiro, se os valores amostrados $x = x^{(1)}, \ldots, x^{(n)}$ fossem independentes, além de identicamente distribuídos (o que já é característica do MCMC para um burn-in grande), a variância seria

$$\mathbb{V}(\bar{f}(x)) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} f(X^{(i)})) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(f(X^{(i)})) = \frac{1}{n}Var_{\pi}(X^{(1)}) = \frac{\sigma^2}{n}.$$

Essa conta é feita assumindo que há independência. No caso de MCMC, há uma certa correlação entre amostras da mesma cadeia de Markov.

Agora, voltamos à hipótese de que $X^{(1)}, \ldots, X^{(n)}$ vem dos sorteios do MCMC, que tem sim uma certa correlação entre si. Então, a variância do estimador é

$$\begin{aligned}
& \mathbb{V}(\bar{f}(x)) \\
= \; & \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} f(X^{(i)})) = \mathbb{E}[\{\frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) - \mathbb{E}[\bar{f}(x)]\}^2] \\
= \; & \mathbb{E}[\{\frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) - \mu\}^2] = \mathbb{E}[\{\frac{1}{n}\sum_{i=1}^{n} [f(X^{(i)}) - \mu]\}^2] \\
= \; & \mathbb{E}[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} [f(X^{(i)}) - \mu][f(X^{(j)}) - \mu]] \\
= \; & \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}[[f(X^{(i)}) - \mu]^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[[f(X^{(i)}) - \mu][f(X^{(j)}) - \mu]] \\
= \; & \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(f(X^{(i)})) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[(f(X^{(i)}) - \mu)(f(X^{(j)}) - \mu)].
\end{aligned}$$

Uma vez que temos $X^{((i))} \sim \pi$, segue que:

$$\frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(f(X^{(i)}) = \frac{1}{n^2}\mathbb{V}_\pi(f(X)) = \frac{\sigma^2}{n}.$$

Enquanto isso, usando o fato da cadeia do Metropolis-Hastings ser homogênea no tempo, temos, tomando $j = k + i$,

$$\mathbb{P}(X^{(j)} \mid X^{(i)}) = \mathbb{P}(X^{(k+i)} \mid X^{(i)}) = \mathbb{P}(X^{(k+1)} \mid X^{(1)}).$$

Daí,

$$\begin{aligned}
& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}[(f(X^{(i)}) - \mu)(f(X^{(j)}) - \mu)] \\
= \; & \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{E}[(f(X^{(i)}) - \mu)(f(X^{(j)}) - \mu)] \\
= \; & \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{k=n-i}^{n} \mathbb{E}[(f(X^{(i)}) - \mu)(f(X^{(i+k)}) - \mu)].
\end{aligned}$$

Para cada i e cada k,

$$\begin{aligned}
& \mathbb{E}\left[(f(X^{(i)})-\mu)(f(X^{(i+k)})-\mu)\right] \\
= \; & \mathbb{E}\left[\mathbb{E}\left[(f(X^{(i)})-\mu)(f(X^{(i+k)})-\mu) \mid X^{(i)}\right]\right] \\
= \; & \mathbb{E}\left[\mathbb{E}\left[(f(X^{(1)})-\mu)(f(X^{(k+1)})-\mu) \mid X^{(1)}\right]\right] \\
= \; & \mathbb{E}\left[(f(X^{(1)})-\mu)(f(X^{(k+1)})-\mu)\right] \\
= \; & \mathrm{Cov}(X^{(1)}, X^{(k+1)}) = \sigma^2\rho(X^{(1)}, X^{(k+1)}) = \sigma^2\rho_k.
\end{aligned}$$

Assim, concluímos que

$$\begin{aligned}
\mathbb{V}(\bar{f}(x)) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(f(X^{(i)})) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}[(f(X^{(i)})-\mu)(f(X^{(j)})-\mu)] \\
&= \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{k=n-i}^{n}\mathbb{E}[(f(X^{(i)})-\mu)(f(X^{(i+k)})-\mu)] \\
&= \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{k=n-1}^{n}\sigma^2\rho_k = \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{k=1}^{n-1}\sum_{i=1}^{n-k}\sigma^2\rho_k \\
&= \frac{\sigma^2}{n}\left[1 + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\rho_k\right] = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho_k\right].
\end{aligned}$$

Ou seja, ainda que as amostras sejam da mesma distribuição objetivo π, que não deixa de ter a mesma variância, a variância do estimador sobre essas amostras, que são dependentes entre si, é diferente.

Em vez de ser apenas a variância da distribuição dividida pelo tamanho da amostra, é dividida por uma termo que podemos interpretar como o tamanho da amostra, a menos da correlação. Vejamos:

$$\mathbb{V}(\bar{f}(x)) = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho_k\right] = \frac{\sigma^2}{ESS(f)},$$

onde

$$ESS(f) = \frac{n}{\left[1 + 2\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho_k\right]}.$$

Esse é o Tamanho Efetivo da Amostra.

Se por acaso, $\rho_k \approx \rho$ constante para todo k, então

$$\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho_k \approx \rho\sum_{k=1}^{n-1}(1-\frac{k}{n}) = \rho\left((n-1) - \frac{n(n-1)}{2n} = \frac{(n-1)\rho}{2}\right),$$

e com isso,

$$ESS(f) \approx \frac{n}{1+(n-1)\rho},$$

que para n grande, é aproximadamente $\frac{1}{\rho}$.

2.2.4 Estudo da Convergência

Sabemos que uma cadeia de Markov ergódica converge para a sua distribuição estacionária. Mas em que sentido converge, a que velocidade? Para isso, vale a pena definir um conceito de distância entre distribuições. Mais geralmente o conceito está definido para medidas.

Definição 11. *Sejam μe ν medidas definidas sobre o espaço $\mathcal{X}$. A distância de variação total entre μ e ν é definida como*

$$\|\mu-\nu\|_{TV} = \frac{1}{2}\sum_{x\in E} \mid \mu(x) - \nu(x) \mid = \sum_{x\in E}(\mu(x)-\nu(x))^+,$$

onde $^+$ representa a função

$$x^+ = \begin{cases} 0, & se\ x \leq 0 \\ x, & se\ x > 0 \end{cases}.$$

Vamos interpretar essa definição. Se duas distribuições de probabilidade com funções de probabilidade μ e ν são iguais, então $\|\mu-\nu\|_{TV} = 0$. Por outro lado, para que a distância seja máxima, as massas de probabilidades devem estar localizadas em conjuntos disjuntos.

Sejam μ e ν medidas de probabilidade. Suponha que $\mathcal{X} = A \cup B$, onde $A = \{x \in \mathcal{X} \mid \mu(\{x\}) > 0\}$ e $B = \{x \in \mathcal{X} \mid \nu(\{x\}) > 0\}$, com A e B disjuntos. Então, $\mu(A) = 1$ e $\mu(B) = 1$. Daí, $\|\mu-\nu\|_{TV} = 1$.

Em qualquer ponto do espaço, as funções de probabilidade não podem ser muito diferentes se a distância de variação total for baixa.

Em [10], a partir do conceito de distância de variação total, a convergência de uma cadeia de Markov pode ser controlada.

Teorema 4. *Seja P a transição de uma cadeia de Markov irredutível e aperiódica, cuja distribuição estacionária é dada por π. Então, existem α entre 0 e 1 e $c > 0$ tais que*

$$\max_{x \in \mathcal{X}} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t.$$

Portanto, a convergência da cadeia para sua distribuição estacionária é exponencial. O cálculo do tempo mínimo para que a distância de variação total esteja abaixo de um valor controlado foge ao conteúdo dessa obra.

De qualquer forma, basta que a distribuição proposta explore bem o espaço de estados para que os resultados sejam satisfatórios. Por outro lado, a exploração de espaços de alta dimensionalidade e de estrutura complicada é um desafio relevante. Frente a isso, outros métodos são propostos.

2.3 Amostrador de Gibbs

Ao longo desse texto, discutimos diferentes métodos para lidar com dimensões cada vez mais altas. Ainda que outros métodos de Monte Carlo via cadeias de Markov apresentem evolução quando comparados aos métodos clássicos, explorar o domínio da distribuição objetivo ao realizar amostragem pode ser ainda um desafio. Quando essa exploração acontece em alta dimensão, pode ser que em algumas coordenadas (direções) o domínio seja espesso demais ou de difícil mobilidade. Nesse caso, evoluir a amostragem de diferentes formas para diferentes direções é uma alternativa.

O amostrador de Gibbs faz a atualização da cadeia coordenada a coordenada, permitindo que se regule quais direções variar mais, ou menos. Esse método foi descrito pela primeira vez pelos irmãos Donald e Stuart Geman em [5], como aplicação em recuperação de imagens. A relação da mecânica estatística com o estudo fez com que o método levasse o nome do físico e matemático Josiah W. Gibbs.

2.3.1 Descrição

Queremos amostrar de uma distribuição de probabilidade com densidade π sobre um espaço $\mathcal{X}$. Supomos que um ponto x do espaço $\mathcal{X}$ possa ser escrito como

$$x = (x_1, \ldots, x_d),$$

onde x_i pode ser um escalar ou um vetor. A hipótese essencial para o uso desse algoritmo é de que seja possível amostrar das condicionais completas, ou seja, denotando $x_{[-i]} = (x_1, \ldots, x_{i-1}, x_{x+1}, \ldots, x_d)$, possa se amostrar de $\pi(X_i \mid x_{[-i]})$, para todo $i \in \{1, \ldots, d\}$. A partir daí, estando em um valor de amostra $x^{(t)} = (x_1^{(t)}, \ldots, x_d^{(t)})$ o algoritmo decorre como

- escolho $i \in \{1, \ldots, d\}$ e
- amostro $x_i' \sim \pi(X_i \mid x_{[-i]}^{(t)})$.

A descrição acima não é suficiente. pois não está claro como i é escolhido e nem o que fazer o valor x_i'. Assim foi escrito porque o Amostrador de Gibbs tem dois tipos, pois o i pode ser escolhido aleatoriamente a cada passo, ou sistematicamente, seguindo uma rotina determinística. E dependendo dessa decisão, o valor de x_i' é incorporado à amostra de uma maneira diferente. Essa escolha difere os dois tipos de amostrador de Gibbs, descritos abaixo.

Na primeira abordagem, em cada iteração o algoritmo sorteia qual será a coordenada alterada naquele passo. Partimos de $x^{(t-1)}$, sorteamos i de uma distribuição $\rho(i)$ pré-definida, e então amostramos a i-ésima coordenada $x_i^{(t)}$ da próxima amostra $x^{(t)}$ a partir da distribuição condicional total $\pi(\cdot \mid x_{[-i]}^{(t-1)})$. De uma iteração para a outra é feita apenas uma alteração, em uma coordenada sorteada, por vez, até que se completem as N iterações pré-determinadas.

Algoritmo 7: Amostrador de Gibbs com Scan Aleatório

Entrada: $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$, $\pi(\cdot \mid x_{[-i]})$, para todo i de 1 a d, N, $\rho(\cdot)$.
para $t = 1, \ldots, N$ **faça**
  $i \sim \rho(\cdot)$
  $x_i^{(t)} \sim \pi(\cdot \mid x_{[-i]}^{(t-1)})$
  $x^{(t)} \leftarrow (x_1^{(t-1)}, \ldots, x_{i-1}^{(t-1)}, x_i^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)})$
Saída : $x^{(1)}, \ldots, x^{(N)}$

Assim, a cadeia anda nas diferentes direções aleatoriamente. Note que, se cada direção tiver uma probabilidade positiva associada, com t tendendo

a infinito, com probabilidade 1 cada direção i terá sido escolhida.

O outro tipo de Amostrador de Gibbs, com Scan Sistemático, tem definida previamente a sequência na qual as coordenadas serão atualizadas, em vez de sortear a coordenada que será a próxima a ser atualizada. Para simplificar, podemos reordenar as coordenadas de forma que a primeira será atualizada primeiro, depois a segunda e assim por diante. Para isso, supomos que uma coordenada não será atualizada duas vezes antes que outra volte a ser atualizada.

Então, a partir de $x^{(t-1)}$, primeiro sorteamos $x_1^{(t)}$ da distribuição condicional total com relação ao tempo anterior $\pi(\cdot \mid x^{(t)_{[-1]}})$. Depois, o passo seguinte já é sorteado a partir da condicional a um estado atualizado $\pi(\cdot \mid (x_1^{(t)}, x_2^{(t-1)}, \ldots, x_d^{(t-1)}))$. Uma coordenada é atualizada por vez, até que se complete todas e recomece o ciclo.

Algoritmo 8: Amostrador de Gibbs com Scan Sistemático

Entrada: $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$, $\pi(\cdot \mid x_{[-i]})$, para todo i de 1 a d, N.
para $t = 1, \ldots, N$ **faça**
  $x_1^{(t)} \sim \pi(\cdot \mid x_{[-1]}^{(t-1)})$
  para $i = 2, \ldots, d-1$ **faça**
    $x_i^{(t)} \sim \pi(\cdot \mid (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)}))$
  $x_d^{(t)} \sim \pi(\cdot \mid x_{[-d]}^{(t)})$
  $x^{(t)} \leftarrow (x_1^{(t)}, \ldots, x_d^{(t)})$
Saída : $x^{(1)}, \ldots, x^{(N)}$

Note que, no SGS, a casa iteração t, d alterações são feitas, enquanto no RGS, a cada iteração, apenas uma mudança ocorre no estado, em uma de suas coordenadas. Então, devemos tomar cuidado com a escolha de N, uma vez que não podemos comparar a velocidade através de N pela forma como os dois algoritmos são diferentes entre si.

Além da exploração de espaços geometricamente complicados, uma importância do amostrador de Gibbs é que ele facilita a simulação, quando é menos complicado amostrar de distribuições "condicionais cheias" (cada

coordenada condicional a todas as outras) do que amostrar diretamente da distribuição objetivo.

2.3.2 Justificativa do Método

O Amostrador de Gibbs com Scan Aleatório (Random-Scan Gibbs Sampler, ou RGS) pode ser interpretado como um tipo particular de algoritmo de Metropolis-Hastings, com matriz de transição e taxa de aceitação específicas.

Proposição 6. *Seguindo o algoritmo do Amostrador de Gibbs com Scan Aleatório para uma distribuição objetivo π, associando a cada índice de coordenada i uma probabilidade α_i de escolha ($\rho(i) = \alpha_i$), tem-se a seguinte transição de Markov:*

$$P(x, y) = \sum_i \pi(y_i \mid x_{[-i]})\delta(y_{[-i]} - x_{[-i]})\alpha_i,$$

para a qual valem as equações de equilíbrio detalhado.

Demonstração: Note que $\pi(y_i \mid x_{[-i]})$ é uma distribuição conhecida por hipótese do algoritmo. De fato, pelo Teorema da Probabilidade Total, temos, para a probabilidade de transição em uma iteração do algoritmo,

$$\begin{aligned}\mathbb{P}(X_t = x^{(t)} \mid X_{t-1} = x^{(t-1)}) &= \sum_i \mathbb{P}(X_t = x^{(t)} \mid i, X_{t-1} = x^{(t-1)})\mathbb{P}(i) \\ &= \sum_i \mathbb{P}(X_t = x^{(t)} \mid i, X_{t-1} = x^{(t-1)})\alpha_i.\end{aligned}$$

Uma vez escolhido o i, o algoritmo aponta que deve-se sortear $x_i^{(t)} \sim \pi(\cdot \mid x_{[-i]}^{(t-1)})$, e substituir o valor na coordenada i. Então, apenas se todas as coordenadas menos uma, no caso a própria i, forem iguais, haverá probabilidade positiva dessa transição.

Ou seja, podemos escrever

$$\mathbb{P}(X_t = x^{(t)} \mid i, X_{t-1} = x^{(t-1)}) = \pi(x_i^{(t)} \mid x_{[-i]}^{(t-1)})\delta(x_{[-i]}^{(t)} - x_{[-i]}^{(t-1)}).$$

E com isso, concluímos a equação do enunciado, sobre a transição.

É possível ir além,

$$\begin{aligned}
&\mathbb{P}(X_t = x^{(t)} \mid i, X_{t-1} = x^{(t-1)}) = \pi(x_i^{(t)} \mid x_{[-i]}^{(t-1)})\delta(x_{[-i]}^{(t)} - x_{[-i]}^{(t-1)}) \\
= \;& \frac{\pi(x_i^{(t)}, x_{[-i]}^{(t-1)})}{\pi(x_{[-i]}^{(t-1)})}\delta(x_{[-i]}^{(t)} - x_{[-i]}^{(t-1)}) \\
= \;& \frac{\pi(x^{(t)})}{\pi(x_{[-i]}^{(t)})}\delta(x_{[-i]}^{(t)} - x_{[-i]}^{(t-1)}),
\end{aligned}$$

onde, no caso em que $x_{[-i]}^{(t)} = x_{[-i]}^{(t-1)}$, temos $x^{(t)}$ definido como $(x_1^{(t-1)}, \ldots, x_{i-1}^{(t-1)}, x_i^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)})$.

Para que a notação seja mais explícita, já sabendo que o algoritmo tem a transição de Markov descrita como acima, denotamos $P(x, y) = \mathbb{P}(X_t = y \mid i, X_{t-1} = x)$. Definimos $S_i(x, y) = \frac{\delta(y_{[-i]} - x_{[-i]})}{\pi(x_{[-i]})}$. Então, caso $y_{[-i]} \neq x_{[-i]}$,

$$S_i(x, y) = \frac{\delta(y_{[-i]} - x_{[-i]})}{\pi(x_{[-i]})} = 0.$$

Se $y_{[-i]} = x_{[-i]}$,

$$S_i(x, y) = \frac{\delta(y_{[-i]} - x_{[-i]})}{\pi(x_{[-i]})} = \frac{\delta(x_{[-i]} - y_{[-i]})}{\pi(y_{[-i]})} = S_i(y, x).$$

Portanto, S é uma função simétrica, e pode-se escrever P como

$$P(x, y) = \sum_i \mathbb{P}(y \mid i, x)\mathbb{P}(i) = \sum_i \pi(y)S_i(x, y)\alpha_i = \pi(y)\sum_i S_i(x, y)\alpha_i.$$

E, sendo cada S_i simétrica, a função $S = \sum_i S_i(x, y)$ é simétrica. Logo, a transição de Markov $P(x, y)$ é reescrita como o produto de $\pi(y)$ por uma função simétrica. Segundo o raciocínio desenvolvido anteriormente a respeito das taxas de variação alternativas na seção 2.2.2, por ser dessa forma, para P valem as equações de equilíbrio detalhado. De fato,

$$\pi(x)P(x, y) = \pi(x)\pi(y)S(x, y) = \pi(y)\pi(x)S(y, x) = \pi(y)P(y, x), \forall x, y \in \mathcal{X}.$$

□

Já criando uma cadeia de Markov cuja distribuição estacionária é π, por conta do equilíbrio detalhado, o algoritmo não exige algum tipo de taxa de

aceitação, visto que a própria cadeia já serve para o algoritmo. Vale perceber que, se fosse calculada uma taxa como no Metropolis-Hastings comum, teria

$$a(x,y) = \min\left\{\frac{\pi(y)P(y,x)}{\pi(x)P(x,y)}, 1\right\},$$
$$\frac{\pi(y)P(y,x)}{\pi(x)P(x,y)} = \frac{\pi(y)\pi(x)S(y,x)}{\pi(x)\pi(y)S(x,y)} = 1, \text{ e}$$
$$a(x,y) = 1, \forall x, y \in \mathcal{X}.$$

Então, este é um caso de Metropolis-Hastings onde sempre se aceita o valor sorteado.

O resultado acima tem o seu análogo para o SGS:

Proposição 7. *Seguindo o algoritmo do Amostrador de Gibbs com Scan Sistemático, com distribuição objetivo π, tem-se uma transição de Markov com distribuição estacionária π.*

Demonstração: De uma iteração para a outra no amostrador de Gibbs sistemático, são sorteados valores novos para cada coordenada dentre as d dimensões, um depois do outro. Da descrição do método, tenho que a matriz de transição é

$$P(x,y) = \pi(y_1 \mid x_2, \ldots, x_d)\pi(y_2 \mid y_1, x_3, \ldots, x_d)\ldots\pi(y_d \mid y_1, \ldots, y_{d-1}),$$

onde $x = (x_1, \ldots, x_d)$ e $y = (y_1, \ldots, y_d)$.

Nesse caso, o amostrador de Gibbs não gera uma cadeia reversível, mas sim uma cadeia que tem π como sua distribuição estacionária. A conta segue:

$$\begin{aligned}
(\pi P)(y) &= \sum_x \pi(x)P(x,y) \\
&= \sum_{x_1}\cdots\sum_{x_d} \pi(x_1, \ldots, x_d)\pi(y_1 \mid x_2, \ldots, x_d)\ldots\pi(y_d \mid y_1, \ldots, y_{d-1}) \\
&= \sum_{x_1}\cdots\sum_{x_d} \pi(x_1, \ldots, x_d, y_1)\pi(y_2 \mid y_2, x_3, \ldots, x_d)\ldots\pi(y_d \mid y_1, \ldots, y_{d-1})
\end{aligned}$$

e assim até que se obtenha

$$(\pi P)(y) = \sum_{x_1}\cdots\sum_{x_d} \pi(x_1, \ldots, x_d, y_1, \ldots, y_d).$$

E daí,

$$\begin{aligned}(\pi P)(y) &= \sum_{x_1} \cdots \sum_{x_d} \pi(x_1, \ldots, x_d, y_1, \ldots, y_d) \\ &= \pi(y_1, \ldots, y_d) = \pi(y),\end{aligned}$$

obtendo que π é distribuição estacionária para d.

A mesma prova pode ser feita para o caso contínuo, com o núcleo de transição dado por

$$K(x, y) = \pi(y_1 \mid x_2, \ldots, x_d) \ldots \pi(y_d \mid y_1, \ldots, y_{d-1}),$$

e fazendo a conta para

$$(\pi K)(y) = \int \cdots \int \pi(x) K(x, y) dx_1 \ldots dx_d.$$

A convergência da cadeia, vem, portanto, das mesmas propriedades de cadeias de Markov que garantem a convergência do algoritmo Metropolis. □

2.4 Monte Carlo Hamiltoniano

2.4.1 O Problema

Os métodos de Monte Carlo por cadeia de Markov tem como base a criação de uma cadeia de Markov que satisfaça as condições necessárias para gerar, após uma certa quantidade de iterações, amostras de uma distribuição desejada. Para criar essa cadeia adequada, o método de Metropolis-Hastings propõe a utilização de outra cadeia, já conhecida e da qual se saiba amostrar a princípio, e um passo de aceitação ou rejeição para que a cadeia se associe à distribuição objetivo. Entretanto, como visto anteriormente, o sucesso desse método depende de uma boa escolha da distribuição proposta.

Se a transição provocar baixas probabilidades de aceitação, o método levará tempo para gerar novas amostras e o tamanho efetivo será baixo. Por outro lado, se a distribuição de transição tiver sua massa de probabilidade muito concentrada em torno do estado atual, também teremos problema com a dependência entre as amostras, e a cadeia levará tempo demais para explorar o espaço.

Uma forma de aproveitar informações da distribuição proposta para MCMC e explorar bem o seu espaço de estados é apresentada pelo algoritmo Monte Carlo Hamiltoniano. Além das propostas serem distantes, elas tem alta probabilidade de aceite.

O problema segue o mesmo: queremos gerar amostras da distribuição π, e gostaríamos que a correlação não fosse tão alta. A seguir, o método é descrito.

2.4.2 Descrição

O método de Monte Carlo usa uma transição proposta baseada na simulação de um sistema Hamiltoniano, da física estatística. A fundamentação teórica completa vai além do escopo desse trabalho, pois exige o estudo de geometria simplética.

Seja π_x a distribuição de interesse. Queremos amostrar valores da π_x sobre o seu espaço amostral, $\mathcal{X}$. Como em muitos casos, sabemos apenas valores do núcleo da densidade $\tilde{\pi}_x$, tendo portanto $\pi_x(x) = \frac{\tilde{\pi}_x(x)}{Z}$. Definimos a função

$$U(x) = -\log(\tilde{\pi}_x(x)),$$

que interpretaremos como "energia potencial" do sistema. Por outro lado,

$$\pi_x(x) \propto e^{-U(x)}.$$

A intenção aqui é tratar um ponto no espaço amostral como uma partícula física. Queremos que a cadeia de Markov se mova pelo espaço amostral gerando valores de π, portanto, fazendo seus movimentos segundo uma lei para a qual valham as equações do equilíbrio detalhado. Cada posição dessa partícula será um valor de $\mathcal{X}$. Para descrever esse movimento, associaremos um momento p, uma segunda variável com a mesma dimensão d de x, a auxiliar na simulação da física. A esse momento, ou quantidade de movimento, estará associada uma energia cinética,

$$K(p) = \frac{1}{2} p^T G^{-1} p,$$

sendo G uma matriz $d \times d$ de covariância, ou seja, simétrica positiva simétrica. Essa matriz induz uma normal multivariada centrada na origem, com expressão dada por

$$\pi_p(p) = ((2\pi)^d |G|)^{-\frac{1}{2}} e^{-K(p)} \propto e^{-K(p)}.$$

Diremos que $H(x,p) = U(x)+K(p)$ é o Hamiltoniano do sistema, sistema esse que é caracterizado por x e p. Dessa forma, a distribuição conjunta de x e de p seria

$$\pi(x,p) = \pi_x(x)\pi_p(p) = e^{-U(x)}((2\pi)^d|G|)^{-\frac{1}{2}}e^{-K(p)} \propto e^{-H(x,p)}.$$

A ideia agora é percorrer sobre o espaço dos estados $z = (x,p)$ de forma a manter constante os valores de $H(x,p)$. Para isso, escrevendo $(x,p) = (x^1,\ldots,x^d,p_1,\ldots,p_d)$, seguiremos as equações de Hamilton,

$$\frac{dx^i}{dt} = \frac{\partial H}{\partial p_i},$$
$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x^i}.$$

A solução desse sistema de equações não é encontrada analiticamente, mas um método numérico permite que se ande sobre as curvas de nível desse sistema. O método usado no HMC é conhecido como "passo de leapfrog" ou também integrador de Störmer–Verlet. A partir de um estado $z^k = (x^k, p_k)$, o próximo é dado pelas equações

$$p_{k+\frac{1}{2}} = p_k - \frac{\tau}{2}\frac{\partial H}{\partial x}(x^k),$$
$$x^{k+1} = x^k + \tau\frac{\partial H}{\partial p}(p_{k+\frac{1}{2}}),$$
$$p_{k+1} = p^{k+\frac{1}{2}} - \frac{\tau}{2}\frac{\partial H}{\partial x}(x^k).$$

Esses passos são boas aproximação (ordem de τ^2 para x e ordem de τ^3 para p) por motivos relacionados à geometria do método. Visto que o erro de aproximação para as curvas dadas pelas equações depende de τ, não escolhemos τ muito grande como passo de "integração numérica" usando o leapfrog. Por outro lado, dar passos muito pequenos entre uma amostra e outra levaria a amostras muito correlacionadas, caindo em problemas comuns ao Metropolis-Hastings e tendo a cadeia de Markov semelhante a um passeio aleatório (o caso particular do algoritmo Metropolis que tem como distribuição proposta um passeio aleatório é o chamado Random Walk Metropolis). A solução, portanto, é fazer vários passos de leapfrog antes de considerar uma nova amostra.

Depois de percorrido o movimento que simula a mecânica Hamiltoniana de H, fazendo o que seria a nossa amostragem proposta, resta um passo de aceitação ou rejeição, para garantir as equações de equilíbrio detalhado. Tais equações vem do interesse na reversibilidade da cadeia: Que se possa chegar ao ponto de origem a partir do ponto de chegada do algoritmo a menos de uma compensação da taxa de aceitação. Para isso, antes de aplicar o passo de rejeição, multiplicamos o momento p obtido pelo leapfrog por -1.

O uso do Leapfrog para a aproximação do sistema Hamiltoniano é essencial para as equações de equilíbrio detalhado, pois é uma aproximação que mantém o volume.

Entre pontos z e w de um espaço de estados geral, a função de aceitação deve ser da forma $\alpha(z,w) = \pi(w)\frac{S(z,w)}{Q(z,w)}$, onde S é uma função simétrica, π é a distribuição objetivo, nesse caso $\pi(z) = \pi(x,p)$, e $Q(z,w)$ é a transição proposta, que descreve a probabilidade de ir para w uma vez que se esteja em z. No caso da cadeia auxiliar do HMC, isso envolve o sorteio de um momento em uma distribuição normal centrada na origem (que é simétrica, uma vez que p e $-p$ tem a mesma probabilidade de escolha), e a evolução determinística (numérica) segundo os passos de leapfrog. Então, a transição proposta no HMC pode ser dita simétrica, e nesse caso,

$$\alpha(z,w) = \pi(w)S(z,w) = \pi(w)Q(z,w)\frac{S(z,w)}{Q(z,w)} = \pi(w)\frac{\tilde{S}(z,w)}{Q(z,w)},$$

onde $\tilde{S}(z,w) = S(z,w)Q(z,w)$. Sabendo disso, a probabilidade de aceitação no Monte Carlo Hamiltoniano é

$$\alpha((x^k,p^{k'}),(x^\star,p^\star)) = \min\{1, \exp\left[-H(x^\star,p^\star) + H(x^k,p^{k'})\right]\}.$$

Tendo tudo isso bem claro, podemos descrever o método explicitamente.

Para amostrar valores $\{x^{(1)},\ldots,x^{(t)}\}$ de π_x, vamos amostrar valores de $\pi(x,p)$, incluindo p, variável de momento. Definimos $H(x,p) = K(p)+U(x)$, onde $K(p) = \frac{1}{2}p^TG^{-1}p$ e $U(x) = -\log\tilde{\pi}_x(x)$, com $\tilde{\pi}_x \propto \pi_x$. Calculamos analiticamente as derivadas parciais $\frac{\partial H}{\partial p} = \frac{\partial K}{\partial p}$ e $\frac{\partial H}{\partial x} = \frac{\partial U}{\partial x}$. O algoritmo se segue:

Algoritmo 9: Monte Carlo Hamiltoniano

Entrada: τ, l, t, G, e derivadas
Inicio com $(x^{(0)}, p^{(0)})$
para $i = 1, \ldots, t$ **faça**
  $(x_0^{(i)}, p_0^{(i)}) \leftarrow (x^{(i-1)}, p^{(i-1)})$
  $p_0^{(i)} \sim \mathcal{N}(0, G)$
  para $k = 1, \ldots, l$ **faça**
    $p_{k'}^{(i)} \leftarrow p_{k-1}^{(i)} - \frac{\tau}{2}\frac{\partial H}{\partial x}(x_{k-1}^{(i)})$
    $x_k^{(i)} \leftarrow x_{k-1}^{(i)} + \tau\frac{\partial H}{\partial p}(p_{k-1}^{(i)})$
    $p_k^{(i)} \leftarrow p_{k'}^{(i)} - \frac{\tau}{2}\frac{\partial H}{\partial x}(x_k^{(i)})$
  Sorteio $u \sim \text{Unif}[0, 1]$
  se $u \leq \alpha((x^{(i-1)}, p_0^{(i)}), (x_l^{(i)}, p_l^{(i)}))$ **então**
    $x^{(i)} \leftarrow x_l^{(i)}$
    $p^{(i)} \leftarrow -p_l^{(i)}$
  senão
    $x^{(i)} \leftarrow x^{(i-1)}$
    $p^{(i)} \leftarrow -p_0^{(i)}$
Saída : $x^{(1)}, \ldots, x^{(t)}$

Esse algoritmo usa informações que temos sobre a distribuição objetivo para moldar a distribuição proposta.

Capítulo 3

Conclusão

Nesse trabalho, vimos algumas técnicas de Monte Carlo e percebemos como diferentes métodos são capazes de superar obstáculos provenientes de algoritmos mais simples. Em particular, as diferenças foram notáveis em problemas de alta dimensionalidade, tanto pela dificuldade em se explorar o espaço amostral da distribuição de interesse, como no controle da variância, que se torna cada vez mais difícil em altas dimensões. O desafio está sempre ligado à variância, e aumentar a quantidade das amostras não costuma ser computacionalmente viável.

Os métodos estatísticos clássicos podem ser satisfatórios em dimensões baixas, onde métodos numéricos também são suficientes e os espaços amostrais são mais simples. O algoritmo de aceitação e rejeição apresentou a dificuldade na geração de uma amostra grande, caso a taxa de aceitação seja baixa. Já o método da amostragem por importância tinha o desafio de uma possível amostra grande com poucos resultados relevantes, levantando o problema do tamanho efetivo da amostra (ESS).

Métodos de Monte Carlo via Cadeias de Markov evitam algumas das dificuldades com dimensão alta, mas também tem seu limite, e cada vez mais os algoritmos tem de ser aperfeiçoados. Vimos que um ponto baixo dos métodos de MCMC é a dependência dos valores gerados, uma vez que pode não haver amostra independente, e um método baseado em Metropolis terá sua correlação atrelada à correlação da cadeia de Markov usada como proposta. Isso influencia diretamente o tamanho efetivo da amostra.

Conseguimos provar a convergência de alguns métodos, mostrando que estes tem embasamento teórico que dá sustentação ao seu uso. Alguns, entretanto, ao longo da pesquisa, mostraram requerer um estudo mais profundo

de geometria e estatística. Com isso, revelam-se possibilidades de trabalhos futuros.

Alguns dos principais textos usados como referência na formação dessa dissertação foram os livros [4], [11], [16], além do artigo [1]. Alguns textos de geometria também foram desbravados, mas não tiveram seu conhecimento explorado no conteúdo final da dissertação.

3.1 Trabalhos Futuros

Afim de compreender mais profundamente o funcionamento dos métodos ligados à mecânica estatística, podemos estudar geometria simplética e preparar um material voltar à sua aplicação na estatística computacional. A justificativa do método de Monte Carlo Hamiltoniano exige esse passo.

Também ligado à geometria Riemanniana estão os métodos de Langevin adaptado por Metropolis e Monte Carlo Hamiltoniano por variedades Riemannianas. Esses métodos permitem a exploração de espaços amostrais de distribuições levando mais ainda em conta suas características geométricas, e podem ser um bom objeto de estudo futuramente.

Tendo compreendido melhor os métodos já inventados, podemos propor melhorias e alternativas. Esse é um possível objetivo futuro, para mais longo prazo. Também estudar a aplicação da estatística computacional no aprendizado de máquina e na própria estatística pode dar melhor intuição sobre os caminhos a seguir e as questões às quais nos engajar.

Bibliografia

[1] Nando e Doucet Arnaud e Jordan Michael I Andrieu, Christophe e De Freitas. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[2] M. E. Box, G. E. P. e Muller. A note on the generation of random normal deviates. *The annals of Mathematical Statistics*, 29:610 – 611, 1958.

[3] Roger Eckhardt. Stan ulam, john von neumann, and the monte carlo method. *Los Alamos Science*, 15(131-136):30, 1987.

[4] Hedibert F Gamerman, Dani e Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 2006.

[5] Donald Geman, Stuart e Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.

[6] R. F. Hicks, J. S.; Wheeling. An efficient method for generating uniformly distributed points on the surface of an n-dimensional sphere. *Comm. Assoc. Comput. Mach.*, 43:13 – 15, 1959.

[7] T. E. Kahn, H. e Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27 – 30, 1951.

[8] A. Kong. A note on importance sampling using standardized weights. Technical report, Department of Statistics, University of Chicago, 1992.

[9] Tim Kroese, Dirk P e Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

[10] Yuval Levin, David A e Peres. *Markov Chains and Mixing Times*, volume 107. American Mathematical Soc., 2017.

[11] Jun S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Science & Business Media, 2008.

[12] David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.

[13] G. Marsaglia. Choosing a point from the surface of a sphere. *Ann. Math. Stat.*, 43:645 – 646, 1972.

[14] A.W.; Rosenbluth M.N.; Teller A.H.; Teller E. Metropolis, N.; Rosenbluth. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 6:1087–1902, 1953.

[15] M. E. Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Comm. Assoc. Comput. Mach.*, 2:19 – 20, 1959.

[16] George Robert, Christian e Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.