

# Teoria: Sistemas de Processamento de Informação

Nestor Caticha

24 de maio de 2013

# Capítulo 1

## Monte Carlo

### 1.1 Integração Numérica em espaços de alta dimensão

Considere o método de integração numérica mais simples, chamado método do trapézio (ver de Vries). Aproximamos a integral

$$I = \int_a^b f(x) dx$$

por

$$I_T = \frac{1}{N} \left( \frac{1}{2} f(x_1) + \sum_{i=2}^{N-1} f(x_i) + \frac{1}{2} f(x_N) \right), \quad (1.1)$$

podemos mostrar que o erro cometido é proporcional a  $h^2$ , onde  $h = (b - a)/N$ , escrevemos então que

$$I = I_T + \vartheta(h^2).$$

Esta estimativa do erro também vale para integrais multidimensionais. Métodos mais sofisticados, baseados neste (e.g. estilo Romberg-Richardson), levam a melhorias no expoente de  $h$ , mas como veremos a seguir, não suficientes.

O custo computacional no cálculo de uma integral é proporcional ao número de vezes que a rotina que calcula o integrando é chamada dentro do programa. Na fórmula do trapézio acima este número de chamadas é  $N$ . Suponhamos um problema típico de Mecânica Estatística, por exemplo um gás dentro de uma caixa. Temos da ordem de  $k = 10^{23}$  moléculas mas digamos que para poder lidar com o problema temos somente  $k = 20$ . Uma aproximação drástica, mas veremos não suficiente. Neste caso é necessário lidar com integrais do tipo

$$Z = \int g(\{r_{ix}, r_{iy}, r_{iz}\}) dr_1^3 dr_2^3 \dots dr_k^3$$

uma integral em  $d = 3k = 60$  dimensões. Suponhamos que o volume da caixa seja  $V = L^3$ , e dividimos cada uma dos  $d$  eixos em intervalos de tamanho  $h$ .

Isto significa uma grade com

$$N = \left(\frac{L}{h}\right)^d$$

pontos. Suponhamos que escolhemos um  $h$  extremamente grande, tal que  $L/h = 10$ , ou seja cada eixo será dividido em somente 10 intervalos. Assim temos

$$N = 10^{60}$$

pontos na grade e esperamos ter um erro talvez da ordem de  $10^{-2}$ . O quê significa um número tão grande como  $10^{60}$ ? Suponhamos que a máquina que dispomos é muito veloz, ou que a função que queremos integrar é muito simples, tal que cada chamada à subrotina demore somente  $10^{-10}$  segundos. O tempo que demorará para calcular  $I_T$  é  $10^{50}$  s. Para ver que isso é muito basta lembrar que a idade do universo é da ordem de  $4 \cdot 10^{17}$  s, portanto nosso algoritmo levará da ordem de  $10^{31}$  idades do universo. Não precisamos muito mais para que nos convençamos a procurar outro método de integração. Variantes do método de trapézio não ajudam muito. Infelizmente o que temos disponível, o Monte Carlo não é muito preciso, mas é muito melhor que isso.

## 1.2 Monte Carlo

### 1.2.1 Teorema Central do Limite: revisitado

Considere uma variável  $X$  com valores  $x$  em um intervalo dado e distribuição  $P(x)$ . Assumimos que os valores médios  $\bar{x}$  e  $\overline{x^2}$  existem e são finitos.<sup>1</sup> A variância  $\sigma_x$  é definida por

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2$$

que também é finita.

Considere ainda uma sequência de  $N$  amostragens independentes de  $X$ :  $\{x_i\}_{i=1, \dots, N}$ , e outra variável  $Y$  com valores  $y$  dados por

$$y = \frac{1}{N} \sum_{i=1}^N x_i$$

Assintoticamente, isto é para  $N$  grande, a distribuição de  $y$  se aproxima de uma distribuição gaussiana, podemos escrever que aproximadamente

$$P(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\bar{y})^2}{2\sigma_y^2}}$$

A aproximação é boa na região central da gaussiana e melhora quando  $N$  cresce. Mais detalhes no futuro ( ou em aulas anteriores). O valor médio de  $y$  e sua variancia são

$$\bar{y} = \bar{x} \text{ e } \sigma_y = \frac{\sigma_x}{\sqrt{N}}$$

<sup>1</sup>Definimos os momentos  $\overline{x^n} = \int x^n P(x) dx$

Notem que se o objetivo for encontrar o valor esperado de  $x$ , que é  $\bar{x}$ , e não for possível realizar a integral, podemos estimar  $\bar{x}$  a partir de  $y$  (isso pode ser generalizado para o cálculo de  $\bar{f} = \int fP(x)dx$ .) Qual é vantagem sobre simplesmente fazer uma medida (amostragem) de  $x$ ? É que neste último caso o erro seria da ordem de  $\sigma_x$ , enquanto que a estimativa baseada em  $y$  terá erro estimado em  $\sigma_y = \sigma_x/\sqrt{N}$ , portanto **o erro da estimativa é independente da dimensão de  $x$** . Para grandes dimensões isso é uma grande vantagem. O problema é que para reduzir o erro por um fator 2 é necessário trabalhar 4 vezes mais duro. E isso para o caso em que as variáveis são independentes e condicional que sabemos gerar as amostras.... . O erro pode ser diminuído não só aumentando  $N$  mas também se mudarmos  $\sigma_x$ . Esse é o objetivo da técnica de amostragem por importância.

**Exercício :** Considere uma variável aleatória  $X$  que toma valores  $-\infty < x < \infty$ , com probabilidade  $P(x)$ . é dado que  $\sigma_x^2 = \overline{x^2} - \bar{x}^2$  é finito. Dado  $y = \frac{1}{N} \sum_{i=1}^N x_i$  mostre, a partir de

$$\begin{aligned} P(y) &= \int P(y, x_1, x_2 \dots x_N) \prod_{i=1}^N dx_i \\ &= \int P(y|x_1, x_2 \dots x_N) P(x_1, x_2 \dots x_N) \prod_{i=1 \dots N} dx_i \\ &= \int P(y|x_1, x_2 \dots x_N) \prod_{i=1}^N P(x_i) dx_i \\ P(y) &= \int \dots \int dx_1 \dots dx_N \delta \left( y - \frac{1}{N} \sum_{i=1}^N x_i \right) \prod_{i=1}^N P(x_i) \end{aligned}$$

que  $P(y)$  é aproximada por uma gaussiana para  $N$  grande. Determine a variância de  $y$ .

**Exercício: Distribuição de Cauchy** Considere o problema acima, exceto que  $\sigma_x^2 = \overline{x^2} - \bar{x}^2$  é infinito pois  $P(x) = \frac{b}{\pi(b^2+x^2)}$ . Encontre a distribuição  $P(y)$  de  $y$ , Note que não é gaussiana para nenhum valor de  $N$ . As integrais necessárias são relativamente fáceis de calcular pelo método dos resíduos.

### 1.2.2 Monte Carlo

A idéia básica é aproximar uma integral  $I$  por  $I_{MC}$

$$I = \int_a^b f(x) dx \simeq I_{MC} = \frac{1}{N} \sum_{i=1}^N f(x_i) \quad (1.2)$$

onde os  $\{x_i\}$  são escolhidos aleatoriamente de forma independente da distribuição uniforme em  $[a, b]$ . Se a integral de  $f^2$  existir e for finita, e se as amostras  $f(x_i)$  forem estatisticamente independentes - e isto é um grande *se* - então o

erro da estimativa MC acima será dado por

$$\sigma_{I_{MC}} = \frac{\sigma_f}{\sqrt{N}}$$

e podemos estimar  $\sigma_f$  a partir dos dados da amostragem

$$\sigma_f^2 \approx \frac{1}{N} \sum f^2(x_i) - \left[ \frac{1}{N} \sum f(x_i) \right]^2.$$

Embora eq. (1.2) possa ser usada para o cálculo da integral, em geral é necessário reduzir a variancia da função  $f$ . Isso é possível através de uma mudança de variáveis, que nem sempre pode ser implementada analiticamente e será descrita a seguir<sup>2</sup>.

O método que iremos descrever não é útil, em geral, para realizar estimativas de Monte Carlo, mas servirá para motivar e sugerir novos caminhos. Imagine uma integral da forma

$$I = \int f(x)w(x)dx,$$

em geral essa separação do integrando em duas funções é muito natural. Tipicamente  $x$  é um vetor em um espaço de muitas dimensões mas  $f(x)$  só depende de algumas poucas componentes de  $x$ , enquanto que  $w(x)$  depende de todas. Suponha que  $w(x)$  esteja normalizado. i.e:

$$\int w(x)dx = 1$$

Ilustraremos a separação em uma dimensão, tomemos o intervalo de integração  $(0, 1)$  e façamos a seguinte mudança de variáveis

$$y(x) = \int_0^x w(z)dz \tag{1.3}$$

$$y(0) = 0, \quad y(1) = 1$$

então  $dy = w(x)dx$  e a integral toma a forma

$$I = \int f(x(y))dy$$

e a aproximação Monte Carlo é

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N} \sum_{i=1}^N f(x(y_i)) \tag{1.4}$$

---

<sup>2</sup>Uma forma trivial de conseguir a redução de  $\sigma_f$  é considerar variações da identidade  $\int_0^1 f(x)dx = \int_0^1 g(x)dx$ , onde  $g(x) = \frac{1}{2}(f(x) + f(1-x))$ . Note que o cálculo de  $g$  é duas vezes mais caro que o de  $f$ , portanto devemos ter  $\frac{\sigma_f^2}{2\sigma_g^2} > 1$  para ter ganho efetivo

onde os valores de  $y_i$  serão amostrados de uma distribuição uniforme no intervalo  $(0, 1)$ . Depois basta calcular a função que relaciona  $y$  e  $x$  (eq. [1.3]). A função inversa permite calcular o valor de  $x$  onde deverá ser calculada a função  $f(x)$ . Este método assume que saibamos fazer a integral da equação 1.3, mas não é em geral possível fazê-lo de forma analítica.

### 1.2.3 Exemplos analíticos.

Ao realizar um cálculo MC teremos, tipicamente, acesso a um gerador de números aleatórios distribuídos uniformemente em  $(0, 1)$ . O objetivo é, aqui de forma analítica e posteriormente, de forma numérica, mostrar como gerar números aleatórios distribuídos de acordo com uma distribuição dada a partir da distribuição disponível. Apresentaremos dois casos muito úteis que podem ser feitos de forma analítica.

Se duas variáveis (em e.g.  $R^N$ ) tem uma relação funcional  $y = \sigma(x)$ , então suas densidades de probabilidade estão relacionadas assim

$$P_Y(y)dy = P_X(x)dx$$

$$P_Y(y)dy = P_X(x) \left| \frac{\partial x}{\partial y} \right| dy \quad (1.5)$$

onde  $\left| \frac{\partial x}{\partial y} \right|$  é o jacobiano da transformação e  $dy = \prod_i dy_i$ . No caso de interesse numérico temos aproximadamente

$$P_Y(y)dy = dy, \quad 0 \leq y_i < 1, i = 1 \dots N$$

e zero fora.

### Distribuição Exponencial

Suponha que queremos gerar amostras de uma distribuição exponencial. i.e  $P_X(x) = \exp(-x)$ . Integrando a eq. (1.5) obtemos

$$y(x) = \int_0^{y(x)} P_X(x) \frac{dx}{dy} dy$$

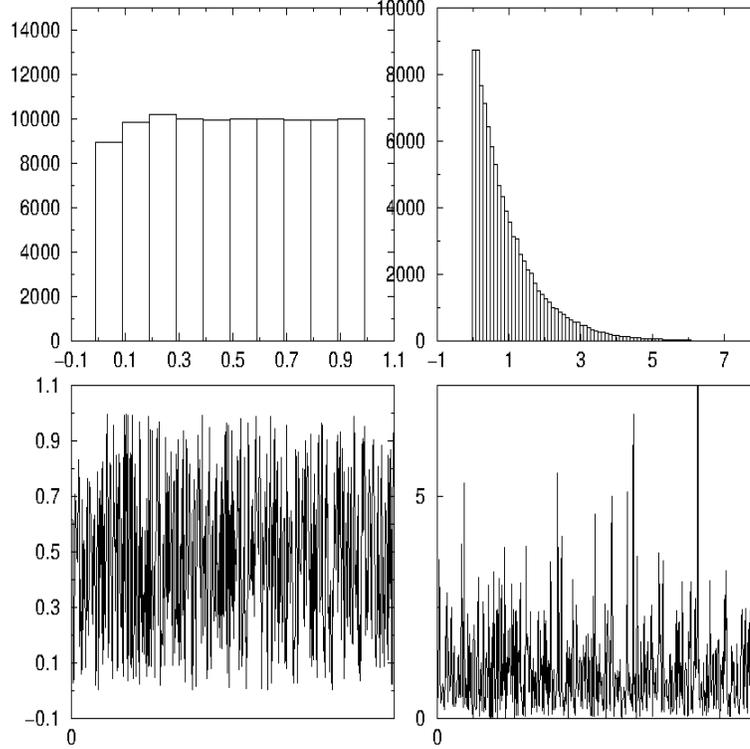
$$y(x) = \int_0^x P_X(x) dx = \int_0^x e^{-z} dz$$

$$y(x) = 1 - \exp(-x)$$

ou  $x = -\ln(y)$  terá a distribuição exponencial desejada, pois se  $y$  é uniforme em  $(0, 1)$  então  $1 - y$  também o é. Portanto é suficiente para gerar números distribuídos exponencialmente usar uma função que gera números aleatórios de distribuição uniforme `RAND(SEED)` e somente uma linha de (pseudo-) código

```
x=-log( RAND(SEED))
```

Compare na figura a distribuição uniforme (esquerda) e a a exponencial (direita) (abaixo : série temporal, acima : histogramas)



### Distribuição Normal

Para gerar números distribuídos de acordo com a distribuição normal é tentador gerar um número grande de amostras de  $P_Y(y)$  e definir  $x = \frac{1}{\sqrt{N}} \sum y_i - \frac{\sqrt{N}}{2}$ , que terá distribuição gaussiana (aproximadamente). O problema é o custo computacional, pois requer  $N$  chamadas da função RAN para gerar uma só amostra de  $x$ . Portanto nunca gere números aleatórios gaussianos dessa maneira. Mais fácil, do ponto de vista computacional é partir da equação (1.5). O método de Box-Muller, mostrado a seguir é muito mais eficiente, pois gera dois números gaussianos para duas chamadas da função geradora de uniformes. Dados  $y_1$  e  $y_2$  obtemos  $x_1$  e  $x_2$  a partir da transformação:

$$\begin{aligned} x_1 &= \sqrt{-2 \ln y_1} \cos 2\pi y_2 \\ x_2 &= \sqrt{-2 \ln y_2} \sin 2\pi y_2 \end{aligned}$$

mostraremos que a sua distribuição conjunta será  $P_X(x_1, x_2) = \frac{1}{2\pi} \exp(-(x_1^2 + x_2^2)/2)$ . Integrando a eq.(1.5) temos:

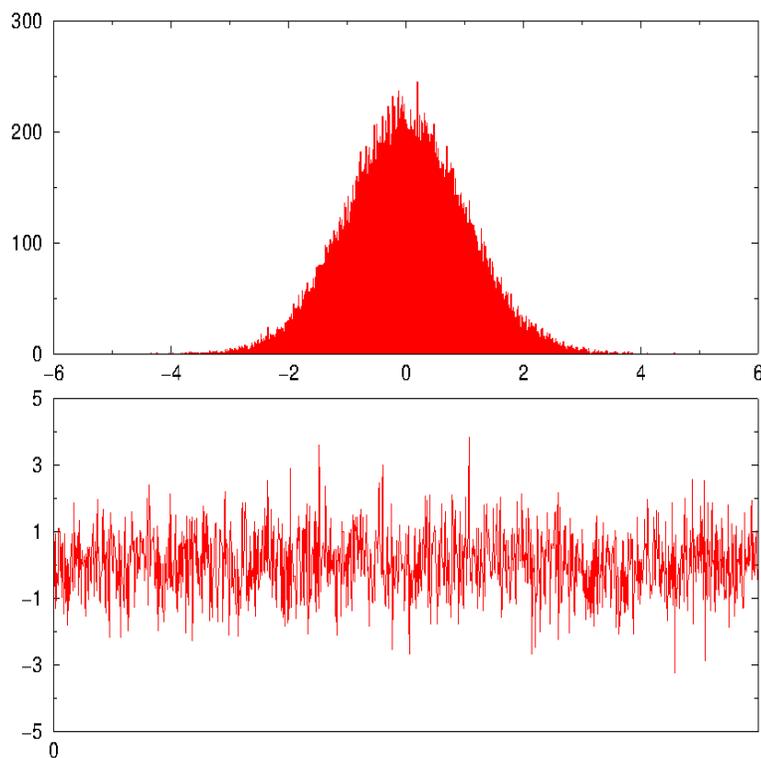
$$\int \int P_Y(y(x_1, x_2)) \left| \frac{\partial y}{\partial x} \right| dy_1 dy_2 = \int \int P_X(x) dx_1 dx_2$$

segue o resultado pois o jacobiano é:

$$J = \left| \frac{\partial y}{\partial x} \right| = \frac{y_1}{2\pi} = \frac{1}{2\pi} e^{-\frac{x_1^2 + x_2^2}{2}}$$

Usando este método obtemos a figura que segue, abaixo temos a série temporal e acima o histograma dos desvios normais:

Estes resultados de muita utilidade na simulação de distribuições gaussianas multivariadas, a ser discutidas posteriormente.



### 1.2.4 Métodos Estáticos: rejeição

Raramente é possível realizar as integrais que permitem descobrir a transformação exata de variáveis e devemos então encontrar uma forma gerar diretamente os  $x$  com a distribuição  $w(x)$ . Os métodos que apresentaremos podem ser divididos em duas classes, estáticos e dinâmicos. Na primeira os números são gerados independentemente um dos outros<sup>3</sup>, enquanto que na segunda classe,

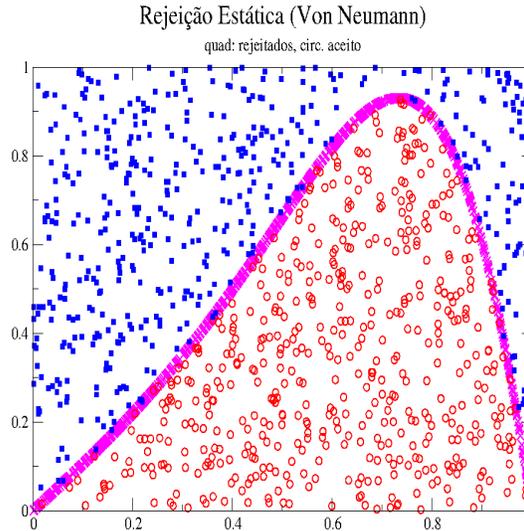
<sup>3</sup>Tão independentemente quanto o gerador de números pseudo-aleatórios o permitir.

construiremos um processo dinâmico que usara informação anterior para gerar o próximo número.

Suponhamos que a região onde  $w(x) \neq 0$  está contida em  $(a, b)$  e que ela é limitada, tal que  $w(x) < c$ . No método de rejeição estático geramos dois NAU  $\xi$  e  $\eta$  e definimos

$$\rho = a + (b - a)\xi, \quad \varphi = c\eta$$

o valor de  $\rho$  será aceito como o novo valor de  $x$  se  $\varphi \leq w(\rho)$  e rejeitado se não.



A seqüência de números aceitos  $x$  são as abscissas dos círculos na figura acima.

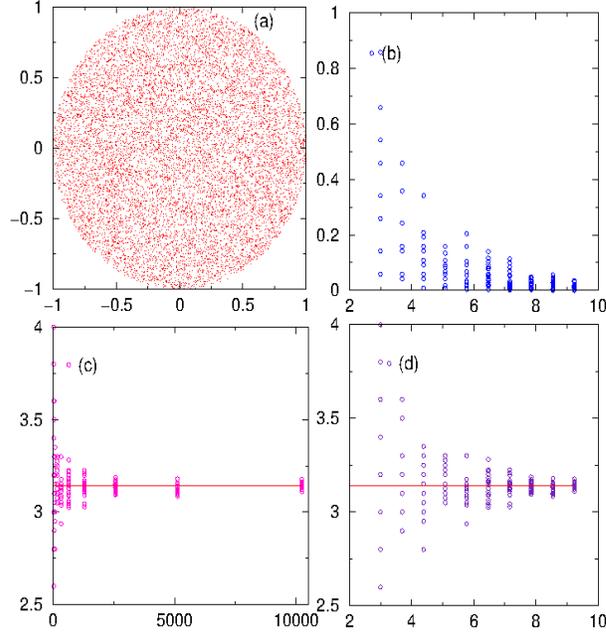
### 1.2.5 Círculo

Exemplo: calcule  $\pi$

A figura mostra os resultados de algumas simulações para estimar  $\pi$ . Foram gerados  $N_{MC}$  pares de números aleatórios  $(x, y)$ . Se  $z = x^2 + y^2 \leq 1$  então o ponto é aceito, de outra forma é rejeitado. Os resultados foram obtidos para  $N_{MC} = 10 * 2^m$  passos de Monte Carlo, com  $m = 2, 4, \dots, 20$ . O resultado (figura (a) abaixo esq. acima) mostra os pares aceitos. Continuando no sentido horário, temos os resultados respectivamente :

- (b) do erro absoluto contra  $\log(N_{MC})$
- (d) resultado de  $\pi_{MC} = (\text{numero aceito} / \text{numero total})$  contra  $\log(N_{MC})$
- (c) resultado de  $\pi_{MC} = (\text{numero aceito} / \text{numero total})$  contra  $N_{MC}$ , os gráficos mostram os resultados de 20 corridas independentes. A dispersão dos pontos nos dá uma idéia dos erros estatísticos. As barras horizontais mostram o valor 3.14159

: (a) pontos aceitos, (b) erro abs, (c) pi vs N, (d) pi vs logN



### 1.2.6 Rejeição estática em espaços de alta dimensionalidade não funciona

Suponha uma melancia hipercúbica na caixa  $\{0, L\}^N$ . A espessura da casca é  $\varepsilon/2$ . Qual é a probabilidade de ao escolher um ponto cujas coordenadas são independentes e uniformemente distribuídas em  $[0 - L]$ , cair na casca? A probabilidade de ao escolher uma das coordenadas do ponto não cair na casca é  $1 - \varepsilon$ . Ao escolher as  $N$  coordenadas, a probabilidade de não cair na casca é  $(1 - \varepsilon)^N = \exp(-cN)$ , onde  $c = -\log(1 - \varepsilon) > 0$ . Portanto a casca domina e a probabilidade de escolher um ponto ao acaso que seja casca vai para 1 quando  $N \rightarrow \infty$ <sup>4</sup>.

Isto mostra que se escolhermos amostras de forma independente nunca sairemos da casca, mas a massa da distribuição pode estar em outras regiões e a estimativa da integral e seu erro serão da mesma magnitude. O problema do método está na independência entre as diferentes amostras. Para corrigir isto precisamos dos métodos dinâmicos.

<sup>4</sup>Alessandro Moura me disse para nunca comprar uma melancia em  $N$  dimensões. Ele atribuiu a história a outra pessoa mas a referência se perdeu no tempo

### 1.3 Métodos Dinâmicos

A idéia por trás dos processos de Monte Carlo dinâmicos é a de um processo estocástico em tempo discreto. Desta forma a amostra em um dado instante depende do passado do processo. Discutiremos nestas aulas somente processos Markovianos de ordem 1, tal que a nova configuração a ser considerada só depende da configuração atual mas não das anteriores. A ferramenta matemática necessária para este método é o teorema de Perron-Frobenius.

#### 1.3.1 O Teorema de Perron Frobenius para matrizes de Markov

Considere um processo estocástico representado por um conjunto de variáveis aleatórias  $\{X_i\}$  onde o índice  $i$  pode ser considerado como um tempo discreto. O valor da variável  $X_i$  é  $x_i$  que toma valores num conjunto  $L = \{\alpha_1, \alpha_2, \dots, \alpha_K\}$ . A probabilidade do evento  $\hat{X}_{0,n} = \{X(t_n) = x(t_n), X(t_{n-1}) = x(t_{n-1}), \dots, X(t_0) = x(t_0)\}$  é denotada por  $\mathbb{P}(\hat{X}_{0,n})$ . A regra do produto para essa sequência leva a

$$\mathbb{P}(\hat{X}_{0,n}) = \mathbb{P}(X_n = x_n | \hat{X}_{0,n-1}) \mathbb{P}(\hat{X}_{0,n-1})$$

Para o caso Markoviano em que a única informação relevante é o último valor de  $x$

$$\mathbb{P}(\hat{X}_{0,n}) = \mathbb{P}(x_n | x_{n-1}) \mathbb{P}(\hat{X}_{0,n-1})$$

que pode ser estendido a

$$\mathbb{P}(\hat{X}_{0,n}) = \prod_{i=1,n} \mathbb{P}(x_i | x_{i-1}) \mathbb{P}(x_0)$$

Seja  $\Gamma$  a matriz de transição de Markov. É uma matriz quadrada  $K \times K$  com elementos não negativos:

$$\Gamma_{ij} = \mathbb{P}(x_n = \alpha_j | x_{n-1} = \alpha_i),$$

é a probabilidade de transição (1-passo) do estado  $i$  para o  $j$ . Consideramos o caso em que estes elementos de matriz não dependem do tempo. Note que  $\sum_j \Gamma_{ij} = 1$ , mas  $\sum_i \Gamma_{ij}$  não é obrigatoriamente 1.

A probabilidade  $P(x_n)$  é obtida marginalizando sobre todas as variáveis  $X_i$ ,  $i = 0, \dots, n-1$ , em notação de matriz

$$P_n = \Pi_0 \Gamma^n \tag{1.6}$$

onde  $P_n$  e  $\Pi_0$ , vetores (linha) de dimensão  $K$ , são respectivamente as probabilidades no instante  $n$  e 0.

Mostraremos a seguir vários resultados que coletados são um caso particular do teorema de Perron-Frobenius:

1.  $\Gamma$  tem um autovetor à direita  $v_1$  que é o vetor coluna com todas as entradas iguais a 1.

2. O autovalor associado a esse autovetor é  $\lambda^{PF} = 1$  de multiplicidade algébrica e geométrica 1
3. Todos os outros autovalores  $\lambda_i$  de  $\Gamma$  satisfazem  $|\lambda_i| < 1$
4. O autovetor à esquerda  $u^{PF}$  associado a  $\lambda^{PF}$  tem todas as componentes não negativas. No caso de matrizes  $\Gamma$  irredutíveis ( $(\Gamma^k)_{i,j} > 0$  para algum  $k > 0$ ) todas são positivas.  $u^{PF}$  pode ser normalizado de modo que a soma das componentes seja 1.
5. Para cada um dos outros autovetores de  $\Gamma$  à esquerda (com  $|\lambda| < 1$ ) a soma das componentes é zero.
6. O vetor  $P_n$  tende exponencialmente rápido com  $n$  para  $u^{PF}$

**Prova de 1 e 2:** Como os elementos de  $\Gamma$  são probabilidades (sobre o segundo índice) temos  $\sum_j \Gamma_{ij} = 1$ . que  $v^1 = (1, \dots, 1)^T$  pode ser escrito

$$\sum_j \Gamma_{ij}(v^1)_j = (v^1)_i$$

ou  $\Gamma v^1 = v^1$ , portanto  $v^1$  é autovetor e seu autovalor associado é 1.

**Prova de 3.** Seja  $v = (v_1, v_2, \dots, v_K)^T$  um autovetor à direita com autovalor associado  $\lambda$

$$\lambda v_i = \sum_j \Gamma_{ij} v_j.$$

Como nem toda componente de  $v$  pode ser nula, existe uma componente que satisfaz  $|v_k| \geq |v_i|$  para todo  $i \neq k$ . Tomando o módulo da equação de autovalores, temos uma primeira desigualdade

$$|\lambda v_i| = \left| \sum_j \Gamma_{ij} v_j \right| \leq \sum_j \Gamma_{ij} |v_j| \quad (1.7)$$

e uma segunda desigualdade é obtida majorando  $|v_j|$  por  $|v_k|$

$$|\lambda v_i| = \left| \sum_j \Gamma_{ij} v_j \right| \leq \sum_j \Gamma_{ij} |v_j| \leq \sum_j \Gamma_{ij} |v_k| = |v_k| \quad (1.8)$$

pois a somatória é 1 por normalização das probabilidades. Assim temos que, para todo  $i$

$$|\lambda v_i| \leq |v_k| \quad (1.9)$$

e em particular para  $i = k$

$$|\lambda| |v_k| \leq |v_k| \quad (1.10)$$

Dois casos são possíveis. Primeiro, se  $|\lambda| = 1$ , então as duas desigualdades acima são igualdades. A desigualdade 1.7 mostra que todos os termos  $v_i$  tem a mesma fase. A segunda desigualdade em 1.8 mostra que os  $v_i$  também tem o mesmo módulo. A conclusão é que se  $|\lambda| = 1$  então o vetor  $v$  so pode ser um múltiplo

de  $(1, \dots, 1)^T$  e que  $\lambda = 1$ , é portanto o único o autovalor 1, i.e. simples. Note que vale o inverso: se  $v$  for um múltiplo de  $v^1$  implica  $|\lambda| = 1$ .

O segundo caso ocorre se  $v$  não for um múltiplo de  $(1, \dots, 1)^T$ , então não pode valer a igualdade:  $|\lambda| < 1$ .

**Prova de 4:** que o autovetor à esquerda  $u^{PF}$  de autovalor 1 tem as componentes não negativas. Para qualquer autovetor à esquerda temos

$$\lambda u_j = \sum_i u_i \Gamma_{ij} \quad (1.11)$$

tomando valor absoluto dos dois lados

$$|\lambda| |u_j| = \left| \sum_i u_i \Gamma_{ij} \right| \quad (1.12)$$

a desigualdade triangular

$$|\lambda| |u_j| = \left| \sum_i u_i \Gamma_{ij} \right| \leq \sum_i |u_i| \Gamma_{ij} \quad (1.13)$$

somando sobre todo os  $j$

$$|\lambda| \sum_j |u_j| \leq \sum_i |u_i| \sum_j \Gamma_{ij} = \sum_i |u_i| \quad (1.14)$$

Se  $|\lambda| < 1$  não diz nada, mas se  $\lambda = 1$  significa que a desigualdade acima é uma igualdade. portanto a expressão 1.13 é uma igualdade e todas as componentes tem a mesma fase, que podemos tomar igual a zero. Segue que todas as componentes são não negativas. Ainda mais, se  $\Gamma$  for irredutível todas as componentes de  $u^{PF}$  serão positivas. Neste caso qualquer estado  $j$  pode ser atingido a partir de qualquer estado  $i$  em  $k$  passos.  $k$  pode depender do par  $i, j$ .

**Prova de 5.** Consideremos outro autovetor à esquerda  $u$  com autovalor  $|\lambda| < 1$ .

$$\lambda u_j = \sum_i u_i \Gamma_{ij} \quad (1.15)$$

$$\lambda \sum_j u_j = \sum_{ij} u_i \Gamma_{ij} = \sum_i u_i \sum_j \Gamma_{ij} \quad (1.16)$$

$$\lambda \sum_j u_j = \sum_i u_i \quad (1.17)$$

e portanto para qualquer autovetor à esquerda com autovalor menor que 1, a soma das componentes deve ser zero.

**Prova de 6:** Convergência para o equilíbrio. Denote, para  $a = 2, \dots, K$ , os autovalores menores que 1 por  $\{\lambda_a\}$  e por  $\{u^a\}$  os autovetores associados à esquerda. Note que acima provamos

$$\sum_i u_i^{PF} = 1, \quad \sum_i u_i^a = 0; \quad (1.18)$$

A condição inicial  $\Pi_0$  pode ser escrita na base dos autovetores à esquerda

$$\Pi_0 = c_0 u^{PF} + \sum_{a=2}^K c_a u^a \quad (1.19)$$

Somando as componentes dos vetores acima, por normalização temos que  $1 = c_0 \times 1 + 0$  ou seja

$$\Pi_0 = u^{PF} + \sum_{a=2}^K c_a u^a \quad (1.20)$$

e usando a equação (1.6)

$$P_n = \Pi_0 \Gamma^n = u^{PF} + \sum_{a=2}^K c_a \lambda_a^n u^a \quad (1.21)$$

e podemos mostrar a convergência em qualquer norma apropriada

$$\|P_n - u^{PF}\| = \left\| \sum_{a=2}^K c_a \lambda_a^n u^a \right\| \quad (1.22)$$

ordene os autovetores de forma que  $|\lambda_2| \geq |\lambda_a|$  para todos os  $a > 2$ :

$$\|P_n - u^{PF}\| = |\lambda_2|^n \left\| \sum_{a=2}^K c_a \left(\frac{\lambda_a}{\lambda_2}\right)^n u^a \right\| < C e^{-n/\tau} \rightarrow 0 \quad (1.23)$$

onde  $\tau = 1/\ln |\lambda_2|^{-1}$  é o tempo característico de termalização.

### 1.3.2 Construção do processo estocástico

Dada uma matriz é típico o problema de obter seus autovetores. Aqui o objetivo é construir um processo estocástico com distribuição de equilíbrio associada igual ao  $w(x)$  dado. Ou seja, dado um vetor de Perron-Frobenius  $w(x)$  queremos encontrar a matriz de transição. A distribuição de equilíbrio ou invariante ou estacionária deve satisfazer a condição de estacionaridade

$$w(x) = \int w(z) \Gamma(x|z) dz, \quad (1.24)$$

mas se não for estacionária teremos a relação entre a probabilidade no instante  $t$  e no seguinte  $t + 1$  dada por

$$P_{n+1}(x) = \int P_n(z) \Gamma(x|z) dz,$$

Dado que as probabilidades de transição são normalizadas  $1 = \int \Gamma(z|x) dz$  segue que

$$\Delta P_n(x) = P_{n+1}(x) - P_n(x) = \int P_n(z) \Gamma(x|z) dz - P_n(x) \int \Gamma(z|x) dz,$$

$$\Delta P_n(x) = P_{n+1}(x) - P_n(x) = \int [P_n(z)\Gamma(x|z) - P_n(x)\Gamma(z|x)] dz \quad (1.25)$$

A interpretação é imediata, a variação da probabilidade, de um instante para o outro, tem duas contribuições, de entrada e saída. O primeiro termo  $[P_n(z)\Gamma(x|z)] dz$  representa a probabilidade de uma caminhada aleatória em um volume  $dz$  em torno de  $z$  no instante  $n$ , que fizeram a sua transição para  $x$  no instante  $n + 1$ . O segundo termo representa a saída, isto é a caminhada em  $x$  que escapa para  $z$ . A integral leva em conta todas as contribuições do espaço. É óbvio a partir das eqs. [1.24, 1.25]

$$\Delta w(x) = \int [w(z)\Gamma(x|z) - w(x)\Gamma(z|x)] dz = 0,$$

o que sugere uma condição

$$w(z)\Gamma(x|z) = w(x)\Gamma(z|x) \quad (1.26)$$

que se a matriz de probabilidade de transições satisfizer então  $w(x)$  será estacionária. Esta condição, chamada de **balanceamento detalhado**, não é necessária, mas só suficiente. Além de haver motivações físicas para impô-la como condição deve ser ressaltado que é talvez a forma mais fácil de realizar o objetivo para construir a matriz de transição. Com qualquer escolha que satisfaça a condição eq. [1.26]  $w(x)$  é um ponto fixo da dinâmica. Mas a pergunta que resta é sobre a estabilidade. É razoável esperar a estabilidade dado que se em  $n$ ,  $P_n(x) > w(x)$ , o número de caminhantes que sairão da região de  $x$  para  $z$  será maior que o que sairiam se a probabilidade fosse  $w(x)$ . Analogamente, se em  $n$ ,  $P_n(x) < w(x)$  então o número será menor.

Há várias maneiras de satisfazer a equação [1.26]. Embora todas levem a algoritmos corretos, no sentido que

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N} \sum_{n=1}^N f(x_n) \quad (1.27)$$

é uma aproximação que melhora para maiores valores de  $N$ , algumas serão eficientes enquanto outras não. Diferentes escolhas levam a diferentes sequências, e a pergunta relevante é: quanta informação nova é trazida por uma nova amostragem? A função de autocorrelação normalizada, que é fundamental para poder julgar a eficiência do MC, é definida por

$$C(k) \equiv \frac{\langle f_n f_{n+k} \rangle - \langle f \rangle^2}{\langle f_n f_n \rangle - \langle f \rangle^2}$$

onde

$$\langle f \rangle = \int f(x)w(x)dx$$

$$\langle f_n f_{n+k} \rangle = \int \int f(x_n) f(x_{n+k}) w(x_n) \Gamma^k(x_{n+k}|x_n) dx_{n+k} dx_n$$

e

$$\Gamma^k(x_{n+k}|x_n) = \int \dots \int \Gamma(x_{n+k}|x_{n+k-1})\Gamma(x_{n+k-1}|x_{n+k-2})\dots\Gamma(x_{n+1}|x_n)dx_{n+k-1}dx_{n+k-2}\dots dx_{n-1}$$

é a probabilidade de transição em  $k$  passos. É óbvio que não é, em geral, possível calcular a autocorrelação, mas podemos estimá-la a partir das amostras colhidas:

$$C_{MC}(k) \equiv \frac{\langle f_n f_{n+k} \rangle_{MC} - \langle f \rangle_{MC}^2}{\langle f_n f_n \rangle_{MC} - \langle f \rangle_{MC}^2}$$

onde definimos a média (empírica) sobre a amostra de dados

$$\langle f_n f_{n+k} \rangle_{MC} = \frac{1}{N-k} \sum_{i=1}^{N-k} f(x_i) f(x_{i+k})$$

Tipicamente -mas não sempre -  $C(k)$  tem um decaimento exponencial:

$$C(k) = e^{-k/\tau}$$

$\tau$  é tempo de correlação exponencial e mede a eficiência do processo em gerar configurações aleatórias independentes distribuídas de acordo com  $w(x)$ . Agora podemos escrever

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N} \sum_{n=1}^N f(x_n) \pm \sigma_f \sqrt{\frac{2\tau}{N}}$$

onde assumimos que depois de um tempo (em unidades de 1 passo MC) aproximadamente  $2\tau$  as novas amostras serão estatisticamente independentes e o número efetivo de amostras será reduzido por esse fator.

Outro tempo importante e diferente é  $\tau_R = |\log \lambda_2|$ , o tempo de relaxação para o equilíbrio associado ao segundo autovalor no problema de Perron-Frobenius. Este mede quanto tempo demora para que o processo estocástico perca memória das condições iniciais e os  $x$  sejam efetivamente representativos de  $w(x)$ . Do ponto de vista de eficiência é razoável não considerar e.g. os primeiros  $K\tau_R$  passos gerados pelo processo. Qual é o valor de  $K$ ? A resposta geral vai estar errada em alguns casos interessantes, é preciso realizar testes numéricos sobre a estabilidade das respostas. Mas  $K$  deve ser pelo menos maior que 5. Se  $C(k)$  efetivamente decair exponencialmente esses dois tempos são relacionados. Perto de transições de fase críticas ou em fases de vidro de spin devemos ter cuidado dobrado pois estes tempos divergem com o tamanho do sistema.

### 1.3.3 Algoritmo de Metropolis

O processo de geração dos números  $x_n$  será separado em duas partes. Em primeiro lugar definimos a probabilidade de *tentativa de mudança*  $T(x_T|x_n)$ , que determina a probabilidade de estando no tempo  $n$  em  $x_n$ , seja escolhido

o ponto  $x_T$  como candidato ao próximo passo da sequência. Uma vez gerado  $x_T$  passamos à segunda parte, que é onde se decide se é feita a transição  $x_n \rightarrow x_{n+1} = x_T$ , ou seja  $x_T$  é aceito ou se não. Neste caso de rejeição fazemos a transição trivial  $x_n \rightarrow x_{n+1} = x_n$ , de forma que  $x_n$  é incluído novamente na sequência, Isto é feito introduzindo a *matriz de aceitação*  $A(x_{n+1}|x_T)$ . Ou seja

$$\Gamma(x|z) = A(x|z)T(x|z)$$

e a condição de balanceamento detalhado, para todo par de pontos  $x \neq z$  toma a forma

$$A(x|z)T(x|z)w(z) = A(z|x)T(z|x)w(x)$$

que é satisfeita por uma família de escolhas possíveis, em particular se definirmos

$$A(x|z) = F\left(\frac{w(x)T(z|x)}{w(z)T(x|z)}\right)$$

e  $F$  tal que

$$\frac{F(a)}{F(1/a)} = a \text{ para todo } a$$

A escolha mais comum, para a probabilidade de tentativa de mudança é tomar

$$T(z|x) = \text{Const dentro de uma bola centrada em } x$$

isso leva a uma taxa de tentativas simétricas ( $T(z|x) = T(x|z)$ ), e portanto basta tomar

$$\frac{A(x|z)}{A(z|x)} = \frac{w(x)}{w(z)}$$

A escolha associada ao nome de Metropolis ( ) é

$$F(a) = \min(1, z)$$

o que leva ao seguinte algoritmo:

1. escolha o valor inicial  $x_0$
2. dado  $x_n$  determinaremos  $x_{n+1}$ : escolha um valor de tentativa  $x_T$  (uniformemente dentro de uma bola de raio  $d$  em torno de  $x_n$ )
3. verifique se  $w(x_T)$  é maior ou menor que  $w(x_n)$ .
  - Se  $w(x_T) \geq w(x_n)$  então **aceita** :  $x_{n+1} = x_T$
  - Se  $w(x_T) \leq w(x_n)$  então escolhe um número aleatório uniforme  $0 \leq \xi < 1$  e

**aceita** :  $x_{n+1} = x_T$  se  $w(x_T) \geq w(x_n)\xi$

**rejeita** :  $x_{n+1} = x_n$  se  $w(x_T) \leq w(x_n)\xi$

volta ao item 2

Imagine o caso em que a função  $w(x)$  pode ser parametrizada da forma

$$w(x) = \frac{e^{-\beta E(x)}}{Z}$$

esse é um dos casos mais interessantes (distribuição de Boltzmann-Gibbs) e a função  $E(x)$  é interpretada como a energia de um sistema no estado  $x$  ou a função custo de um processo.  $Z$  é uma constante em relação a  $x$  mas depende do parâmetro  $\beta$  que em física é interpretado como o inverso da temperatura. Este tipo de função ocorre quando a probabilidade que devemos atribuir a uma dada configuração é baseada na informação que temos sobre o valor médio  $\langle E(x) \rangle$  e é o resultado de encontrar a distribuição com a máxima entropia consistente com a informação dada.

O algoritmo de Metropolis pode ser redescrito da seguinte forma:

1. escolha o valor inicial  $x_0$
2. dado  $x_n$  determinaremos  $x_{n+1}$ : escolha um valor de tentativa  $x_T$  (uniformemente dentro de uma bola de raio  $d$  em torno de  $x_n$ )
3. verifique se  $E(x_T)$  é maior ou menor que  $E(x_n)$ .
  - Se  $E(x_T) \leq E(x_n)$  então **aceita** :  $x_{n+1} = x_T$
  - Se  $E(x_T) \geq E(x_n)$  então escolhe um número aleatório uniforme  $0 \leq \xi < 1$  e

**aceita** :  $x_{n+1} = x_T$  se  $\exp(-\beta(E(x_T) - E(x_n))) \geq \xi$

**rejeita** :  $x_{n+1} = x_n$  se  $\exp(-\beta(E(x_T) - E(x_n))) \leq \xi$

volta ao item 2

A processo realiza a caminhada aleatória de forma que uma diminuição na energia é sempre aceita, mas se há uma tentativa de escolha de um lugar de energia mais alta, a tentativa não é automaticamente rejeitada. Se o aumento de energia for muito grande então sim é rejeitada, mas se não for, então é aceita. A escala de grande ou pequeno é determinada pela razão dos fatores de Boltzmann de cada configuração.

enddocument