Notas de Aula

Introdução às Simulações de Monte Carlo em Mecânica Estatística e Teorias de Gauge na Rede

Attilio Cucchieri e Tereza Mendes

Primeiro Semestre de 2002

Conteúdo

Prefácio

Capítulo 1

Introdução ao Método de Monte Carlo

1.1 Idéia geral do método

Considere o problema de calcular a área de um círculo de raio 1 através do lançamento de pontos aleatórios uniformemente no quadrado definido por $x \in [-1, 1]$, $y \in [-1, 1]$ como na Fig. 1.1.

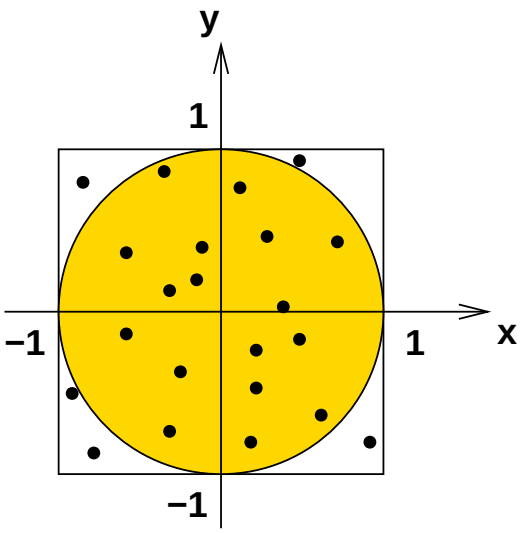


Figura 1.1: Cálculo da área de um círculo usando o método de Monte Carlo.

Uma estimativa para a área será dada pela fração destes pontos que estiver contida no círculo, no limite de grande número de pontos N. De fato, se chamarmos de a a área do círculo e de A a do quadrado, a razão entre as áreas será dada por

$$\frac{a}{A} = \frac{\pi}{4} = \frac{n}{N}, \qquad (1.1)$$

onde $n < N$ é o número de pontos contidos no círculo. Claramente, podemos escrever n como uma soma de variáveis aleatórias n_i — independentes e igualmente distribuídas — assumindo valor 1 se o ponto estiver contido no círculo e 0 se não estiver. Neste caso, o método de Monte Carlo consiste em estimar a razão de áreas desejada através da média dos N valores de n_i gerados. Esta média é por sua vez uma variável aleatória, flutuando ao redor de seu valor esperado $\pi/4$ com uma certa variância, ou desvio quadrático médio. A raiz quadrada desta variância constitui uma medida da imprecisão ou erro na determinação do valor médio $\pi/4$. Mais especificamente, sendo

$$\overline{n} \equiv \frac{1}{N} \sum_{i=1}^{N} n_i \qquad (1.2)$$

e usando o fato de que as variáveis n_i são independentes e igualmente distribuídas obtemos

$$\sigma^2_{\overline{n}} \;=\; \frac{1}{N^2}\sum_i \sigma^2_{n_i} \;=\; \frac{1}{N}\,\sigma^2_{n_i}\,, \tag{1.3}$$

onde a variância individual $\sigma^2_{n_i}$ é um número finito. Vemos portanto que o erro estatístico $\sigma_{\overline{n}}$ associado à grandeza $\overline{n}$ — produzida como estimativa para a quantidade determinística a/A, em que estamos interessados — decresce com a raiz quadrada do número N de pontos gerados

$$\sigma_{\overline{n}} \;\sim\; 1/\sqrt{N}\,. \tag{1.4}$$

Este comportamento é comum a todos os métodos de Monte Carlo (a dependência com $1/\sqrt{N}$ é dada pelo teorema central do limite) e indica que a convergência para o valor determinístico que estamos estimando é bastante **lenta**. De fato, a convergência dada por (1.4) implica que para obtermos um erro estatístico duas vezes menor é preciso um investimento computacional quatro vezes maior, ou correspondentemente para se produzir uma estimativa com um algarismo significativo a mais é necessário um esforço computacional 100 vezes maior!

Como veremos na próxima seção, o comportamento (1.4) para o erro de Monte Carlo torna o método muito pouco eficiente quando comparado aos métodos de quadratura convencionais, a não ser para aplicações envolvendo um espaço de variáveis com um número muito grande de dimensões. Mas são precisamente estes os tipos de problemas que estaremos interessados em tratar. Por exemplo, aplicações em Mecânica Estatística envolvem o cálculo de médias de observáveis na distribuição de Boltzmann para o sistema. Isto corresponde a uma integral em um espaço com número de dimensões pelo menos igual ao número de componentes do sistema, ou seja tipicamente um número no mínimo da ordem de várias centenas. (A fim de que o resultado reproduza o limite termodinâmico, este número deve aproximar-se de infinito.)

1.1.1 Exemplo

O programa em `fortran` abaixo calcula a área do círculo de raio 1 representado na Fig. 1.1 com o método de Monte Carlo descrito acima. Geradores de números aleatórios serão discutidos brevemente na Seção 1.4.1. Por ora, note que a partir da instrução `rand()`, que fornece um número aleatório entre 0 e 1, podemos facilmente construir uma variável aleatória `r` distribuída uniformemente num intervalo $[a, b]$ qualquer usando a expressão `r = (b-a) * rand() + a`.

```
! calculo da area do circulo lancando pontos no quadrado
program circle
implicit real*8 (a-h,o-z)
implicit integer (i-n)
dimension n(10000000)

read (5,*) Niter
area = 3.1415926535897 93d0
```

```
na = 0
do i = 1, Niter
  x = 2.d0 * rand() - 1.d0
  y = 2.d0 * rand() - 1.d0
  r = x**2 + y**2
  if (r.le.1.d0) then
    n(i) = 1
  else
    n(i) = 0
  endif
  na = na + n(i)
enddo

! media de n(i)
a = dfloat(na) / dfloat(Niter)
! variancia de n(i)
sigma = 0.d0
do i = 1, Niter
  sigma = sigma + ( dfloat(n(i)) - a )**2
enddo
sigma = sqrt( sigma / dfloat(Niter-1) )
! erro na media de n(i)
erro = sigma / sqrt( dfloat(Niter) )

write (6,*) Niter, " iteracoes"
write (6,*) "a =", 4.d0 * a, " +-", 4.d0 * erro
write (6,*) "MC/exact =", 4.d0 * a/area
end program
```

Note que a variável `a`, usada no cálculo da razão de áreas, é multipliacada no final pelo fator `4.d0` correspondente à área do quadrado. Como regra geral, multiplica-se por fatores constantes (como `4.d0` e `1/Niter` no exemplo acima) apenas ao final de um ciclo a fim de reduzir o número de operações, já que cada operação envolve um erro de arredondamento. Note também que a variância é "estimada" pela variável `sigma`, como explicado na Seção 1.3.1 [ver Eq. (1.18)]. Vejamos os resultados do programa para alguns valores de `Niter`:[1]

```
10 iteracoes
a = 3.2 +- 0.533333333
MC/exact = 1.01859164

1000 iteracoes
a = 3.088 +- 0.0530949628
MC/exact = 0.982940929

100000 iteracoes
```

[1] Note que evitamos o uso da variável `N` para o número de iterações no código acima pela possível confusão com o vetor `n(i)`.

```
a = 3.14376 +- 0.0051882945
MC/exact = 1.00068989

10000000 iteracoes
a = 3.142256 +- 0.000519158117
MC/exact = 1.00021115
```

Podemos notar nestes resultados o comportamento (1.4) para o erro estatístico (cada vez que N aumenta de um fator 100 o erro diminui por um fator aproximandamenmte 10). Como mencionado, este comportamento será o mesmo para todos os métodos de Monte Carlo, estando a maior eficiência de um método em relação a outro associada a um valor menor para a constante que multiplica $1/\sqrt{N}$.

1.1.2 Exercícios

1) Repita o exemplo acima para o caso da hiper-esfera de raio 1 em d dimensões (por exemplo para $d = 3, 10$), verificando que o comportamento (1.4) é independente da dimensão do espaço de integração. Note que o volume no caso geral é dado por $\pi^{d/2}/\,\Gamma(d/2+1)$. Calcule também o raio médio dos pontos gerados dentro da hiper-esfera [isto é, os pontos para os quais $n(i) = 1$], verificando o resultado exato $<r> = \int_0^1 r^d\,dr\,/\,\int_0^1 r^{d-1}\,dr\; = d/(d+1)$, que reflete o fato de a distribuição uniforme concentrar-se na superfície à medida que a dimensão d aumenta (o raio médio tende a 1).

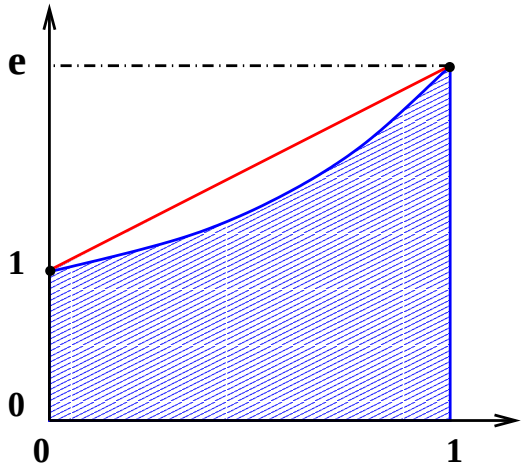


Figura 1.2: Área sob a exponencial.

2) O método descrito pode ser usado no cálculo de integrais definidas em geral. Para a integral de uma função positiva $f(x)$ no intervalo $[a,b]$ considera-se a razão entre a área sob o gráfico da função e a área do retângulo definido por $x \in [a,b]$, $y \in [0, f_{max}]$, onde f_{max} é o valor máximo de $f(x)$ no intervalo $[a,b]$. Utilize o método para calcular a integral

$$\int_0^1 e^x\,dx\,. \qquad (1.5)$$

3) Nosso método torna-se claramente ineficiente se a área do retângulo onde são gerados os pontos for grande demais comparada à área que queremos estimar. No caso acima, por exemplo, ao invés de gerar pontos uniformemente no retângulo definido por $x \in [0,1]$, $y \in [0,e]$, podem-se produzir pontos no trapézio dado pela área sob a reta $g(x) \equiv (e-1)\,x + 1$ no intervalo $[0,1]$, como na Fig. 1.2. Esta região também engloba a área desejada e possui uma área menor. Faça o cálculo da integral (1.5) gerando pontos uniformemente nesta figura. Verifique que o erro na estimativa da área é menor por um fator aproximadamente 3 quando comparado ao exercício anterior, ou seja a convergência neste

caso é quase 10 vezes mais rápida. [**Sugestão**: as coordenadas (x, y) podem ser geradas a partir de pares de números aleatórios (r_1, r_2) uniformemente distribuídos em $[0, 1]$. Suponhamos que $x = x(r_1)$. Uma vez gerado x, a distribuição para y será uniforme entre 0 e $g(x)$, ou seja $y(r_1, r_2) = r_2\, g(x(r_1))$. Para garantir que x, y sejam gerados uniformemente no trapézio devemos construir $x(r_1)$ de forma que o Jacobiano $\partial(r_1, r_2)/\partial(x, y)$ seja uma constante.]

4) Verifique graficamente as distribuições uniformes dos pontos nas áreas das figuras consideradas no Exemplo 1.1.1 e nos Exercícios 2 e 3 acima. Faça gráficos de (x, y) respectivamente para o quadrado, o retângulo e o trapézio, usando todos os pontos gerados. Considere agora somente os pontos dentro da área desejada, isto é os pontos com $n(i) = 1$, gerando gráficos para o círculo e para a região sob a exponencial.

1.2 Comparação com métodos determinísticos

Nesta seção faremos uma pequena revisão dos métodos mais simples empregados no cálculo de integrais definidas de funções em espaços de baixa dimensão, as chamadas quadraturas.

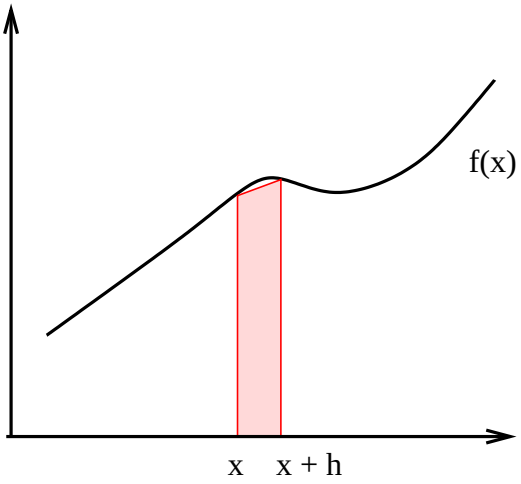


Figura 1.3: Regra do trapezóide.

Como primeiro método vamos considerar a **regra do trapezóide** — ou aproximação de dois pontos — que consiste em estimar a área compreendida entre as abscissas x e $x+h$ como a área do trapezóide definido pela aproximação **linear** da função entre estes dois pontos (ver Fig. 1.3). Para obter este resultado, suponhamos que a função seja conhecida apenas nos pontos extremos deste intervalo: $f_1 \equiv f(x)$ e $f_2 \equiv f(x+h)$. Podemos então escrever a expansão de Taylor (linear) para $f(x')$ em torno do ponto médio do intervalo $x_0 \equiv x + h/2$, estimando o valor de $f(x_0)$ e da derivada $f'(x_0)$ em termos de f_1 e f_2

$$f(x') \;=\; \frac{f_1 + f_2}{2} \;+\; \frac{f_2 - f_1}{h}\,(x' - x_0) \;+\; \mathcal{O}[(x' - x_0)^2 f''] \tag{1.6}$$

onde f'' é a derivada segunda de $f(x)$ calculada em algum ponto desconhecido do intervalo $[x, x + h]$. Claramente, se a função $f(x)$ for linear a expansão acima será exata. Podemos agora estimar o valor da integral no intervalo $[x, x + h]$ usando (1.6) e escrever

$$\int_x^{x+h} f(x')\,dx' \;=\; \frac{h}{2}\,[f(x) \;+\; f(x+h)] \;+\; \mathcal{O}(h^3 f'')\,. \tag{1.7}$$

Vemos que a integral neste intervalo é dada pela área do trapezóide na Fig. 1.3, com um erro de ordem h^3. [Note que o termo linear na expansão (1.6) não contribui

para a integral e portanto o erro é o mesmo que teríamos obtido se ao invés da aproximação linear tivéssemos empregado a aproximação constante, correspondente a um "histograma", com contribuição $h\,f(x_0)$ para a integral.] Quando considerada sucessivamente para todos os intervalos h no domínio de integração $[a, b]$, ou seja para os intervalos $[a, a+h], [a+h, a+2h], \cdots, [b-h, b]$, esta regra fornece o resultado

$$\int_a^b f(x)\,dx \;=\; h\,[\,\frac{1}{2}f(a) \,+\, f(a+h) \,+\, f(a+2h) \\ +\,\cdots\,+\, f(b-h) \,+\, \frac{1}{2}f(b)\,] \,+\, \mathcal{O}(h^2)\,, \qquad (1.8)$$

de onde se vê que o erro no cálculo da integral é proporcional ao quadrado do intervalo tomado entre as abscissas para as quais o integrando é calculado.[2]

Uma precisão bem melhor é obtida pela **regra de Simpson**, baseada em uma aproximação de três pontos para a integral em pequenos intervalos e correspondente a uma aproximação quadrática da função. Consideremos, neste caso, o intervalo $[x-h,\, x+h]$ e suponhamos que a função seja conhecida nos extremos deste intervalo e no ponto médio x. Uma estimativa da expansão (quadrática) de Taylor para a função $f(x')$ ao redor do ponto x é dada por

$$f(x') \;=\; f(x) \,+\, \frac{f(x+h)-f(x-h)}{2h}\,(x'-x) \\ +\,\frac{f(x+h)-2f(x)+f(x-h)}{h^2}\,\frac{(x'-x)^2}{2} \,+\, \mathcal{O}[(x'-x)^3 f''']\,, \qquad (1.9)$$

onde f''' é a derivada terceira de $f(x)$ calculada em algum ponto desconhecido do intervalo $[x-h,\, x+h]$. Neste caso pode-se verificar que se $f(x)$ for quadrática a expansão acima será exata. Usando (1.9) podemos estimar a integral da função no intervalo $[x-h,\, x+h]$ obtendo

$$\int_{x-h}^{x+h} f(x')\,dx' \;=\; \frac{h}{3}\,[f(x-h) \,+\, 4\,f(x) \,+\, f(x+h)] \,+\, \mathcal{O}(h^5 f^{(4)})\,, \qquad (1.10)$$

onde $f^{(4)}$ é a derivada quarta de $f(x)$ calculada em algum ponto desconhecido do intervalo $[x-h,\, x+h]$. Note que neste caso consideramos um intervalo de comprimento $2h$. Note também que o erro é menor do que o esperado, ou seja da ordem de h^5 ao invés de h^4, pois o termo cúbico da expansão em série de Taylor da função $f(x')$ não contribui para a integral. Quando considerada sucessivamente para os intervalos $[a, a+2h], [a+2h, a+4h], \cdots, [b-2h, b]$, esta regra fornece o resultado

$$\int_a^b f(x)\,dx \;=\; \frac{h}{3}\,[\,f(a) \,+\, 4f(a+h) \,+\, 2f(a+2h) \,+\, 4f(a+3h) \,+\, 2f(a+4h) \\ +\,\cdots\,+\, 4f(b-h) \,+\, f(b)\,] \,+\, \mathcal{O}(h^4)\,, \qquad (1.11)$$

[2] Note que o erro original, da ordem de h^3, é somado $(b-a)/h$ vezes, correspondentes ao número de intervalos em que foi aplicado o resultado (1.7), e portanto o erro final e de ordem h^2.

de onde vemos que o erro no cálculo da integral é proporcional à quarta potência do intervalo tomado entre as abscissas para as quais o integrando é calculado, ou seja temos um resultado com erro duas ordens de magnitude menor do que no caso da regra do trapezóide.

Os dois exemplos acima são casos especiais de fórmulas fechadas de Newton-Cotes, baseadas no cálculo da função em abscissas a intervalos igualmente espaçados. Em geral há métodos mais eficientes, como as quadraturas Gaussianas, em que o integrando é calculado para abscissas escolhidas convenientemente com espaçamentos variáveis. De qualquer forma, apesar de sua simplicidade, podemos notar imediatamente que estes métodos possuem precisão **muito** maior do que a dos métodos de Monte Carlo. De fato, seja N o número de pontos considerados nos métodos acima, correspondente ao esforço computacional para se chegar a uma dada precisão. Este número é dado pela largura do intervalo h por $N = (b-a)/h$. Temos portanto que a precisão, ou o erro a ser comparado com o caso de Monte Carlo é $\mathcal{O}(N^{-2})$ para a regra do trapezóide e $\mathcal{O}(N^{-4})$ para a regra de Simpson. Ou seja — lembrando que o erro de Monte Carlo é de $\mathcal{O}(N^{-1/2})$ — é necessário um N quatro vezes maior para se reduzir à metade o erro com Monte Carlo, enquanto que utilizando-se $2N$ no método do trapezóide reduzimos o erro por um fator quatro, e no método de Simpson por um fator 16, o que necessitaria 256 vezes mais esforço com o método de Monte Carlo.

Apesar disso, um fato importante a lembrar é que, ao contrário do que acontece com os métodos de quadraturas determinísticos, o erro discutido para os métodos de Monte Carlo é independente da dimensão do espaço de integração (como vimos na seção anterior, por exemplo no Exercício 1.1.2.1). Para os casos determinísticos, como os de pontos igualmente espaçados discutidos acima, um número de pontos N utilizados para uma integral em d dimensões corresponde a intervalos $h \sim 1/N^{1/d}$ em cada dimensão, e portanto a precisão obtida é $\mathcal{O}(N^{-2/d})$ para a regra do trapezóide e $\mathcal{O}(N^{-4/d})$ para a regra de Simpson. O mesmo não acontece para Monte Carlo porque os pontos são gerados uniformemente no espaço multi-dimensional de variáveis. Isto determina que à medida que consideremos integrais em espaços de dimensões mais altas o método de Monte Carlo tornar-se-á progressivamente mais competitivo. Por exemplo, em dimensão 4 sua precisão será comparável à do método do trapezóide e em dimensão 8 à do método de Simpson. Para termos uma idéia do número de dimensões envolvido em aplicações usuais em física estatística ou teorias de gauge na rede consideremos o seguinte. Em um sistema físico cada variável (por exemplo spins em sítios de uma rede, ou campos quânticos em pontos do espaço) representa ao menos uma dimensão do espaço de integração, e em geral várias dimensões se se tratar de variáveis com mais graus de liberdade. Desta forma, mesmo para um sistema extremamente simples como o modelo de Ising em 3 dimensões, utilizando-se para cálculo uma rede com apenas 10 sítios de lado teremos que considerar uma integral (ou melhor, neste caso, uma soma) num espaço de 1000 dimensões e não há nenhuma esperança de podermos usar os métodos determinísticos convencionais.[3] Para um

[3]Para o modelo de Ising temos apenas dois estados possíveis para cada variável, e portanto as médias na distribuição de Boltzmann envolvem somas com $2^d = 2^{1000} \sim 10^{300}$ termos. Mesmo considerando-se uma máquina na região de Teraflops (1 Teraflop = 10^{12} operações de ponto flutuante

modelo de spins contínuos de Heisenberg teremos 2 dimensões contínuas por sítio, ou seja $d = 2000$, e para uma teoria de gauge pura (sem efeitos de férmions dinâmicos) no caso simples de matrizes $SU(2)$, tomadas em elos de uma rede quadridimensional, teremos uma integral com $d = 10^4 \times 4 \times 4 = 160.000$ dimensões!

1.2.1 Exercícios

1) Refaça a integral da função exponencial dos Exercícios 1.1.2.2–3 usando o método do trapezóide e compare a precisão à obtida pelos métodos de Monte Carlo. Note que o mesmo pode ser feito para a integral do círculo, considerando-se duas vezes a integral do semi-círculo definido pela função $f(x) = \sqrt{1-x^2}$ para $x \in [-1, 1]$.

2) Idem para o método de Simpson.

1.3 Método de Monte Carlo estático

1.3.1 Amostragem simples

Seguindo a idéia do método de Monte Carlo introduzida na Seção 1.1 — baseada em estimar-se uma integral (determinística) a partir de uma média de variáveis aleatórias — podemos calcular de maneira geral a integral[4]

$$I = \int_a^b \frac{f(x)}{b-a}\, dx \tag{1.12}$$

simplesmente gerando pontos x_i uniformemente no intervalo $[a, b]$ e tomando a média dos valores $f(x_i)$. Este procedimento corresponde a interpretar a integral I como a média estatística da variável $f(x)$ na medida uniforme

$$u(x)\, dx \;\equiv\; \frac{dx}{b-a}\,, \tag{1.13}$$

normalizada de maneira que $\int_a^b u(x)\, dx = 1$. Escrevemos portanto a estimativa para o valor de I

$$\overline{f} \;\equiv\; \frac{1}{N} \sum_{i=1}^{N} f(x_i), \tag{1.14}$$

a qual convergirá, no limite de N tendendo a infinito, para a integral desejada. O erro estatístico associado à estimativa $\overline{f}$, dado pela raiz quadrada de sua variância, terá o comportamento familiar (1.4) dos métodos de Monte Carlo. De fato, como na Seção 1.1, as variáveis $f(x_i)$ são **independentes** — valendo portanto que a variância da soma é igual à soma das variâncias — e igualmente distribuídas, ou seja a variância de $f(x_i)$ é a mesma para todo i. Portanto, de maneira análoga à eq. (1.3), temos o erro

$$\sigma_{\overline{f}} \;=\; \frac{\sigma_f}{\sqrt{N}}\,, \tag{1.15}$$

por segundo) esta soma levaria da ordem de 10^{280} anos, algo como 10^{270} vezes a idade do universo!

[4] O fator $1/(b-a)$, inserido sem perda de generalidade, simplifica a nossa discussão.

onde a variância individual σ_f^2 é dada por

$$\sigma_f^2 \;=\; \int_a^b [\, f(x) \;-\; I\,]^2\, u(x)\, dx \;= <(\, f - <f>)^2> = <f^2> - <f>^2 \;. \qquad (1.16)$$

Note que indicamos com $<\cdot>$ as médias estatísticas na distribuição $u(x)$.

O método acima, chamado de **amostragem simples**, é claramente mais geral do que o procedimento geométrico introduzido na Seção 1.1, já que nem sempre será possível encontrar uma figura apropriada onde saibamos lançar pontos uniformemente. De fato, podemos pensar no método geométrico (também chamado de "hit or miss") como uma amostragem simples em que uma função degrau $f(x,y)$ assumindo valor 1 ou zero — respectivamente nos casos em que os pontos (x,y) estejam dentro ou fora da área desejada — é mediada para um número N de pontos (x_i, y_i) gerados uniformemente no quadrado (ou no trapézio, no caso do Exercício 1.1.2.3), multiplicando-se no final pelo volume do espaço de integração. Estas duas versões de métodos de Monte Carlo, assim como a versão mais eficiente discutida na próxima seção, constituem exemplos de **métodos de Monte Carlo estático**, por serem baseadas na geração dos N pontos aleatórios de maneira independente. Em outras palavras, não há a noção de tempo associada aos dados produzidos e as médias calculadas podem ser tomadas a partir dos dados em qualquer ordem.

Nota: $\overline{f}$ constitui um **estimador** para a integral I, ou seja uma variável aleatória que converge para I no limite $N \to \infty$. Além disso, $\overline{f}$ é um estimador **não-viciado**, pois sua média é igual a I para qualquer valor de N

$$<\overline{f}> = \int_a^b \overline{f}\; u(x)\; dx \;=\; \frac{1}{N} \sum_{i=1}^{N} <f> = \; I\,, \qquad (1.17)$$

onde usamos o fato de que as variáveis x_i — e portanto as variáveis $f_i \equiv f(x_i)$ — são igualmente distribuídas. Analogamente, podemos construir um estimador não-viciado para a variância σ_f^2 da seguinte forma

$$\begin{aligned} \sigma_{f,N}^2 \;&\equiv\; \frac{N}{N-1}\left(\overline{f^2} \;-\; \overline{f}^2\right) \;\equiv\; \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N} f_i^2 \;-\; \overline{f}^2\right) \\ &=\; \frac{1}{N-1}\sum_{i=1}^{N} (\, f_i \;-\; \overline{f}\,)^2\,, \end{aligned} \qquad (1.18)$$

já que

$$\lim_{N\to\infty} \sigma_{f,N}^2 \;=\; \sigma_f^2 \qquad (1.19)$$

e

$$\begin{aligned} <\sigma_{f,N}^2> \;&=\; \frac{1}{N-1}\sum_i <f_i^2> \;-\; \frac{N}{N-1}\;\frac{1}{N^2}\;\sum_i\sum_j <f_i\, f_j> \\ &=\; <f^2> \;-\; \frac{1}{N\,(N-1)}\sum_{i\neq j} <f_i\, f_j> \;=\; \sigma_f^2\,. \end{aligned} \qquad (1.20)$$

Com base em (1.15), podemos agora escrever o seguinte estimador não-viciado para o erro $\sigma_{\overline{f}}$

$$\sigma_{\overline{f},N} = \frac{\sigma_{f,N}}{\sqrt{N}} . \tag{1.21}$$

É esta a forma que usamos para estimar o erro nos resultados de Monte Carlo no caso estático, como no exemplo de programa 1.1.1. Note que usamos várias vezes o fato de as variáveis f_i serem igualmente distribuídas. Além disso, ao escrever (1.15) e na última passagem de (1.20), usamos o fato de as variáveis serem independentes, ou seja $< f_i\, f_j > = < f_i >< f_j >$ para $i \neq j$. Para o caso de Monte Carlo dinâmico (discutido na Seção 1.5, e usado no restante do curso) os valores de f_i estarão correlacionados. Neste caso, como veremos, teremos ainda o comportamento (1.4) para o erro de Monte Carlo, mas a constante de proporcionalidade não será mais dada por $\sigma_{f,N}$.

Como exemplo de amostragem simples, consideremos o fragmento de programa abaixo, escrito para uma função $f(x)$ geral. (Compare ao modelo de programa para o caso geométrico fornecido da Seção 1.1.)

```
fmed = 0.d0
do i = 1, N
  x(i) = (b - a) * rand() + a
  f(i) = ffunc(x(i))
  fmed = fmed + f(i)
enddo
fmed = fmed/dfloat(N)
```

Os valores `ffunc(x(i))` serão obtidos a partir da definição para a função $f(x)$ explicitamente ou usando-se uma sub-rotina, e serão armazenados no vetor `f(i)`. Por exemplo para a função exponencial a instrução explícita seria `f(i) = exp(x(i))`, ou alternativamente a sub-rotina a ser adicionada ao final do programa seria

```
function ffunc(x)
implicit real*8 (a-h,o-z)

ffunc = exp(x)

return
end function
```

O erro para a estimativa `fmed` é dado por (1.18) e (1.21) como

```
sigma = 0.d0
do i = 1, N
  sigma = sigma + (f(i) - fmed)**2
enddo
sigma = sigma / dfloat(N-1)
erro = sqrt( sigma / dfloat(N) )
```

Note que de acordo com (1.18) podemos calcular `sigma` de duas maneiras diferentes: 1) calculando primeiro $\overline{f}$ e depois mediando $(f - \overline{f})^2$ ou 2) calculando simultaneamente $\overline{f}$ e $\overline{f^2}$ e tomando a diferença $\overline{f^2} - \overline{f}^2$. Embora a segunda maneira seja implementada usando-se apenas um ciclo do tipo "`do i = 1, N`" e a primeira requeira dois ciclos, costuma-se utilizar a primeira maneira, pois para um número grande de iterações a subtração de quantidades mediadas estará mais sujeita a erros de arredondamento acumulados. De fato, escrevamos cada variável f_i em termos de sua diferença em relação a $\overline{f}$

$$f_i \;=\; \overline{f} \,+\, \delta f_i \,. \tag{1.22}$$

Vemos que o estimador $\sigma^2_{N,f}$ é dado simplesmente como

$$\sigma^2_{N,f} \;=\; \frac{1}{N-1} \sum_i {\delta f_i}^2 \,. \tag{1.23}$$

No primeiro caso a soma acima é calculada diretamente, enquanto que no segundo caso ela deve resultar de

$$\frac{N}{N-1} \left[\frac{1}{N} \sum_i (\overline{f}^2 \,+\, 2\,\overline{f}\,\delta f_i \,+\, {\delta f_i}^2) \;-\; \overline{f}^2 \right] , \tag{1.24}$$

após o cancelamento de $\overline{f}^2$ e de $\sum_i \delta f_i = 0$. (A soma que escrevemos em parênteses é apenas o número f_i^2, mas desta forma podemos visualizar as várias contribuições individuais.) Claramente teremos mais erros de arredondamento envolvidos no segundo caso, especialmente considerando que em geral as variáveis δf_i serão pequenas comparadas a $\overline{f}$. Neste caso, para valores grandes de i, o termo ${\delta f_i}^2$ a ser adicionado durante o cálculo da somatória pode ser da ordem do erro de arredondamento acumulado e portanto sua contribuição será perdida.

1.3.2 Amostragem por importância

O método de amostragem simples discutido na seção anterior gera pontos uniformemente no espaço de integração. Claramente, se a função $f(x)$ for fortemente concentrada em alguma região deste espaço o algoritmo perderá grande parte do tempo adicionando termos $f(x_i)$ à média para valores x_i "pouco importantes", ou seja para pontos em que a função f possui valor desprezível. Isto acontece por exemplo para cálculos de médias de observáveis com a distribuição de Boltzmann para um sistema, a qual se concentra exponencialmente nas configurações energeticamente mais favoráveis. Em casos como este gostaríamos de considerar médias em que os valores de x_i gerados possuíssem contribuição significativa para a soma. Isso é possível se redefinirmos a função a ser integrada, como descrito a seguir.

Escrevamos a integral (1.12) da forma

$$I \;=\; \int_a^b \frac{f(x)}{b-a}\; dx \;=\; \int_a^b \frac{f(x)}{w(x)}\, \frac{w(x)}{b-a}\, dx \tag{1.25}$$

onde $w(x)$ é uma função positiva em $[a,b]$ satisfazendo

$$\int_a^b w(x)\,dx \;=\; b-a\;. \tag{1.26}$$

Neste caso estamos calculando a média da função $f(x)\,/\,w(x)$ em $[a,b]$ com x distribuído neste intervalo segundo a medida normalizada $w(x)/(b-a)$. Note que estamos integrando uma função diferente, e a variável x não é mais distribuída uniformemente no intervalo $[a,b]$. Consideremos agora a mudança de variáveis

$$y(x) \;=\; a \,+\, \int_a^x w(x')\,dx'\;, \tag{1.27}$$

a qual satisfaz as condições $y(b) = b$ e $y(a) = a$. Se a relação $y = y(x)$ puder ser invertida — ou seja se pudermos determinar a função $x = x(y)$ — e já que $dy/dx = w(x)$, podemos finalmente escrever a integral I como

$$I = \int_a^b \frac{f[x(y)]}{w[x(y)]}\,\frac{dy}{(b-a)}\;. \tag{1.28}$$

Esta é a média de $f[x(y)]/w[x(y)]$ em $[a,b]$ com y distribuído uniformemente neste intervalo. Claramente, se a função $w(x)$ for convenientemente escolhida, o cálculo da média acima será uma maneira mais eficiente de se determinar a integral I. De fato, se $w(x) \approx f(x)$ no intervalo $[a,b]$ então a razão $f(x)\,/\,w(x)$ — que é a nova função a ser integrada — é quase constante e as flutuações de $f(x_i)$ em torno do valor médio I serão reduzidas. Esta redução na largura dos valores que podem ser assumidos pela função de integração é exatamente o que buscávamos, ou seja, os termos somados para o cálculo da média serão de importância equivalente, já que para a nova função as abscissas x (geradas a partir da distribuição uniforme para y) serão pontos onde a função assume valores significativos.

O programa para cálculo de I será modificado da seguinte maneira[5]

```
fmed = 0.d0
do i = 1, N
  y(i) = (b - a) * rand() + a
  x(i) = xfunc(y(i))
  f(i) = ffunc(x(i))/wfunc(x(i))
  fmed = fmed + f(i)
enddo
fmed = fmed/dfloat(N-1)
```

enquanto que o programa para cálculo de `sigma` não muda. Como veremos nos exercícios, o método modificado será bem mais eficiente que o original.

Em geral porém será difícil encontrar uma mudança de variáveis como a introduzida acima, pois ela implica que se saiba integrar $w(x)$ e inverter $y(x)$. O que devemos notar no caso geral é que a amostragem por importância envolve o cálculo

[5]Devem ser fornecidas definições para `xfunc`, `ffunc` e `wfunc`.

de uma função suave — um observável — para valores da variável x — ou correspondentemente valores para a configuração de um sistema físico — dados por uma distribuição normalizada não uniforme e em geral fortemente concentrada em regiões do espaço de estados, de forma que uma amostragem uniforme — ou "simples" — seria totalmente ineficiente.

Tipicamente tem-se

$$< A > = \int A(x)\, w(x)\, dx \,, \tag{1.29}$$

onde $w(x)$ é uma distribuição normalizada não uniforme. Por exemplo, considera-se a distribuição Gaussiana

$$w(x) \;=\; \frac{e^{-x^2}}{\sqrt{2\pi}} \,, \tag{1.30}$$

ou a distribuição de Boltzmann para um sistema físico

$$w(x) \;=\; \frac{e^{-\beta\, \mathcal{H}(x)}}{\mathcal{Z}} \,, \tag{1.31}$$

onde $\mathcal{H}$ e $\mathcal{Z}$ são respectivamente a Hamiltoniana e a função de partição do sistema, β é a temperatura inversa, e x representa as possíveis configurações de valores para as componentes do sistema. Neste caso os observáveis $A(x)$ serão geralmente a energia, magnetização, etc. [Em nossa discussão acima $A(x) = f(x)/w(x)$ e a distribuição normalizada é $w(x)/(b-a)$.]

O cálculo da média de $A(x)\, w(x)$ com x distribuído uniformemente é claramente inviável, e daí a necessidade da amostragem por importância, ou seja, da amostragem de x de acordo com $w(x)$. O problema da amostragem da distribuição $w(x)$ é central para os métodos de Monte Carlo, seja no caso estático discutido até aqui que no caso mais geral de amostragem de $w(x)$ de forma dinâmica, discutido na Seção 1.5. Na próxima seção tratamos o problema de amostragem de algumas distribuições de maneira exata e introduzimos o método da rejeição, o qual permite a amostragem em casos mais complicados.

1.3.3 Exercícios

1) Repita a integral do círculo e da função exponencial [veja o comentário sobre $f(x)$ para o círculo no Exercício 1.2.1.1] usando o método de amostragem simples descrito na Seção 1.3.1, e compare a eficiência à do método geométrico dos Exercícios 1.1.2.

2) Considerando mais uma vez a integral da função exponencial em $[0, 1]$, efetue uma mudança de variáveis como discutido na Seção 1.3.2, tomando neste caso

$$w(x) \;=\; 2\,[\,(e-1)\,x \,+\, 1\,]\,/\,(e+1)\,. \tag{1.32}$$

Verifique que a convergência do método é cerca de 60 vezes mais rápida do que no caso de amostragem simples do exercício anterior. Faça um gráfico da função dada pela razão $f(x)/w(x)$, verificando que ela é suave no intervalo $[0, 1]$, e portanto apropriada para a amostragem por importância da variável x. Note que a função

$w(x)$ definida acima é proporcional à reta da Fig. 1.2 e propocional (ou, neste caso, igual) à probabilidade de produzir um dado valor x. Portanto pontos distribuídos uniformemente na área sob a função $w(x)$ terão abscissas dadas pela distribuição $w(x)$. Ou seja, os valores $x(i)$ que buscamos são as abscissas dos pontos gerados uniformemente no trapézio do Exercício 1.1.2.2.

3) Em geral não será necessário (ou desejável) armazenar os valores `x(i)` e `y(i)` em vetores[6] durante o cálculo de I e portanto nos programas das Seções 1.3.1 e 1.3.2 podemos usar variáveis simples `x` e `y` e armazenar apenas `f(i)`. Estes pontos foram armazenados a fim de verificarmos que eles tenham a distribuição esperada. Faça histogramas e calcule o valor médio para `x(i)` e `y(i)` nos casos acima, verificando a distribuição uniforme para `x(i)` no Exercício 1 e para `y(i)` no Exercício 2, e a distribuição linear $w(x)/(b-a)$ para `x(i)` no Exercício 2. Note que para uma distribuição de probabilidades normalizada $w(x)$ teremos $w(x)\,dx = n(x)/N$ no limite de grande número de iterações N, onde $n(x)$ é o número de vezes que o valor gerado na simulação está entre x e $x+dx$. (O incremento dx será tomado finito e o menor possível, dependendo de N.) Portanto o histograma ou gráfico de $n(x)/(N\,dx)$ deve aproximar a função $w(x)$.

1.4 Amostragem de variáveis aleatórias

Como vimos, a amostragem por importância é uma necessidade para a implementação eficiente de algoritmos de Monte Carlo. Nela estará envolvida uma consideração cuidadosa da distribuição de probabilidades, a fim de produzir sequências de valores para as variáveis do problema — ou amostras — de acordo com a distribuição desejada. Seja para aplicações em que se deseja apenas amostrar uma distribuição (por exemplo a distribuição Gaussiana) ou como parte de métodos para amostragem de distribuições complicadas com muitas variáveis — como nos métodos de Monte Carlo dinâmico discutidos na próxima seção — o emprego de técnicas de amostragem eficientes é crucial para o desempenho da simulação. Nesta seção discutimos vários métodos de amostragens exatas, que serão depois usados em aplicações físicas como parte dos algoritmos de Monte Carlo dinâmico (por exemplo como sub-rotinas de programas). O estudo destes métodos é independente dos algoritmos para aplicações físicas, e a seção pode ser usada como referência à medida que estes métodos forem introduzidos. (Nas aplicações específicas será feita referência aos ítens desta seção.) Iniciamos com uma descrição dos métodos para geração de variáveis uniformes. A seguir discutimos como o uso destas variáveis permite a amostragem de distribuições diretamente ou, quando isto não for possível, pelo método de rejeição.

[6] Em aplicações físicas cada valor destes poderá corresponder por exemplo a uma configuração para um sistema com milhares de variáveis de spins. Armazenam-se então apenas os observáveis calculados para estas configurações, a não ser em casos em que as configurações sejam muito difícies de produzir e seja vantajoso guardá-las.

1.4.1 Geradores de números aleatórios

A função `rand()` que usamos até agora constitui um exemplo de gerador de números aleatórios. Cada execução da função produz um número aleatório r uniformemente distribuído em [0,1], ou seja com distribuição de probabilidades $u(r)\,dr = dr$. Em nossas aplicações usamos N chamadas à função, produzindo uma sequência de N números aleatórios r_i $(i = 1, \ldots, N)$ independentes e igualmente distribuídos com distribuição uniforme em [0,1]. Na verdade, porém, esta sequência é *determinística*: sempre que esta operação for repetida a partir do mesmo ponto inicial a sequência obtida será a mesma. O gerador é uma prescrição algébrica que, dado um número inicial, produz uma sequência fixa de números r_i com a distribuição desejada. Os números gerados desta maneira para aplicações numéricas estocásticas são chamados de **pseudo-aleatórios**. O termo refere-se ao fato de eles serem aleatórios mas reprodutíveis. O ponto inicial da sequência pode ser escolhido inicializando-se a função `rand()` com uso de um número inteiro — a semente inicial — da seguinte maneira

```
! escolha da semente (inteira), que chamamos de iseed
! e chamada da funcao que faz a inicializacao
iseed = 2002
call srand(iseed)
```

A cada chamada da função `rand()` a semente é usada para produzir a variável real r_i e atualizada para uso na próxima chamada. Note que em nossas aplicações a função `rand()` não foi inicializada, ou seja, a semente inicial não foi escolhida, sendo tomada como o valor de "default", igual a 1.

Geradores diferentes produzirão sequências diferentes a partir de uma dada semente inicial, mas as aplicações numéricas em que tais sequências são usadas não podem depender do gerador escolhido, ao menos para geradores de boa qualidade (testes de qualidade serão discutidos abaixo). Geradores são em geral construídos de maneira que a semente seja um número inteiro (ou vários números inteiros) e a cada passo da sequência um novo número inteiro é produzido e usado como semente para o passo sucessivo. Estes números inteiros são então convertidos em números reais entre 0 e 1. O objetivo de um gerador assim construído é reproduzir tão proximamente quanto possível a distribuição uniforme.

Historicamente os primeiros geradores para aplicações numéricas produziam sequências de números manualmente, com o auxílio de roletas elétricas, e armazenados em listas para uso posterior. Outras sugestões de procedimentos para produção sistemática de números aleatórios foram **... completar...**

Os métodos mais simples comumente usados são os chamados métodos congruenciais lineares. Veremos a seguir exemplos de geradores de números aleatórios com o método congruencial linear, e testes para estabelecer a sua eficiência.

Um gerador de números aleatórios — ou melhor pseudo-aleatórios, como discutido acima — é uma função $r_{n+1} = F(r_n)$ com $F(r) : [0,1] \to [0,1]$ possuindo as seguintes características:

- A distribuição dos pontos r_i deve ser uniforme, ou seja a sequência de números pseudo-aleatórios deve ser indistinguível de uma sequência de números alea-

tórios. Mais especificamente, todos os testes satisfeitos por uma sequência de números aleatórios devem ser satisfeitos também pela sequência de números pseudo-aleatórios.

- Todos os números pseudo-aleatórios possuem um período T. Isto é, após ter gerado T números r_i a sequência se repete. Claramente o período T deve ser muito maior do que o comprimento da sequência necessária para uma simulação a fim de evitar correlações nos dados produzidos.
- Deve ser possível armazenar a cada momento a semente associada a um número da sequência, de forma que possa ser repetida qualquer parte da sequência.
- A sequência de números produzida a partir de uma certa semente deve ser a mesma em computadores diferentes.
- O tempo empregado na geração de cada número deve ser o menor possível.

testes...

O **método congruencial linear** para geradores de números aleatórios baseia-se numa função $F(r_n)$ dada por

$$\begin{aligned} i_{n+1} &= (a\, i_n + c) \mod m \\ r_{n+1} &= \frac{\mathrm{dble}(i_{n+1})}{\mathrm{dble}(m)} \end{aligned}$$

onde a, c e m são números inteiros fixos. O método congruencial linear gera portanto uma sequência de números inteiros i_n à qual é associada uma sequência de números reais r_n. Para inicializar a sequência é necessário portanto fixar o valor de i_0. Claramente i_{n+1} é menor do que m e maior ou igual a zero, ou seja i_{n+1} pode assumir no total m valores diferentes. Isto implica que $r_{n+1} \in [0, 1)$ e que o período máximo deste gerador seja igual a m. (Dependendo da escolha de a, c e m e da semente inicial i_0 o período da sequência gerada pode ser menor do que m.)

Como vimos, a fim de termos um gerador com um período longo é necessário usar um valor de m muito grande. Neste caso pode acontecer que m seja maior do que o maior inteiro que pode ser armazenado no computador (veja o Exercício 2). Nestes casos pode-se resolver o problema fatorizando m. Por exemplo, se $m = m_1\, m_2$ podemos calcular o módulo de um número inteiro z usando a seguinte relação

$$s = z \mod m = z \mod m_1 + m_1(\lfloor \frac{z}{m_1} \rfloor \mod m_2) \tag{1.33}$$

onde $\lfloor z/m_1 \rfloor$ indica a parte inteira de z/m_1. Para demonstrar esta relação observe que $s = z \mod m$ implica $z = p\, m + s$ con $p = \lfloor z/m \rfloor$. Podemos portanto escrever

$$\begin{aligned} z &= p_1\, m_1 + s_1 \\ s_1 &= z \mod m_1 \\ p_1 &= \lfloor \frac{z}{m_1} \rfloor \end{aligned}$$

e

$$\begin{aligned} p_1 &= p_2\, m_2 + s_2 \\ s_2 &= p_1 \bmod m_2 \\ p_2 &= \lfloor \frac{p_1}{m_2} \rfloor \end{aligned}$$

Obtemos que

$$z = p_1\, m_1 + s_1 = (p_2\, m_2 + s_2)\, m_1 + s_1 = p_2\, m + (s_2\, m_1 + s_1)\ .$$

Usando agora o fato de que $s_1 \leq m_1 - 1$ e $s_2 \leq m_2 - 1$ é imediato verificar que $s_2\, m_1 + s_1$ é menor do que m, ou seja

$$s = z \bmod m = s_1 + s_2\, m_1$$

e esta é exatamente a fórmula (1.33).

Chiaramente o número $a\, i_n + c$ também pode ser grande demais. Neste casos se a é maior do que, por exemplo, m_2 é conveniente também escrever $a = a_2\, m_2 + b$ con a_2 e b interi. Isto implica que

$$i_{n+1} = (a_2\, m_2\, i_n + b\, i_n + c) \bmod m\ .$$

Allo stesso tempo si puo' scrivere $i_n = l_n\, m_1 + k_n$ con l_n e k_n interi. Ne segue che

$$\begin{aligned} i_{n+1} &= (a_2\, m\, l_n + a_2\, m_2\, k_n + b\, l_n\, m_1 + b\, k_n + c) \bmod m \\ &= (a_2\, m_2\, k_n + b\, l_n\, m_1 + b\, k_n + c) \bmod m \\ &= (a\, k_n + c) \bmod m_1 + (b\, l_n + \lfloor \frac{a\, k_n + c}{m_1} \rfloor) \bmod m_2 \end{aligned}$$

ovvero

$$\begin{aligned} k_{n+1} &= (a\, k_n + c) \bmod m_1 \\ l_{n+1} &= (b\, l_n + \lfloor \frac{a\, k_n + c}{m_1} \rfloor) \bmod m_2 \end{aligned}$$

Infine se $m_2 = e\, m_1$ com e intero abbiamo che

$$k_{n+1} = (b\, k_n + c) \bmod m_1 \tag{1.34}$$

$$l_{n+1} = (b\, l_n + e\, a_2\, k_n + \lfloor \frac{b\, k_n + c}{m_1} \rfloor) \bmod m_2 \tag{1.35}$$

Abbiamo cosi' due sequenze di numeri interi $0 \leq k_n \leq m_1 - 1$ e $0 \leq l_n \leq m_2 - 1$. Infine possiamo calcolare

$$r_{n+1} = \frac{\mathrm{dble}(i_{n+1})}{\mathrm{dble}(m)} = \frac{\mathrm{dble}(l_{n+1}\, m_1 + k_{n+1})}{\mathrm{dble}(m_1\, m_2)} = \frac{\mathrm{dble}(l_{n+1}) + \frac{\mathrm{dble}(k_{n+1})}{\mathrm{dble}(m_1)}}{\mathrm{dble}(m_2)}\ . \tag{1.36}$$

Usando le disuguaglianze per l_n e k_n possiamo verificare che $0 \leq i_n \leq m - 1$ e $r_n \in [0, 1)$. Nota che nelle formule (1.34)–(1.36) non dobbiamo mai calcolare i_n ne' usare m, ovvero tutte le operazioni involvono "numeri piccoli". Ovviamente in questo caso dobbiamo fornire i valori di l_0 e k_0 al fine di inizializzare la sequenza di numeri pseudo-aleatori r_n. (Per una applicazione di queste formule vedi Exercícios 3 e 4.)

1.4.2 Amostragens diretas

Apresentamos a seguir alguns exemplos em que é possível uma amostragem exata de distribuições a partir da distribuição uniforme discutida acima. Sejam r e x duas variáveis aleatórias com distribuições de probabilidade respectivamente $u(r)$ e $w(x)$. Se as variáveis estiverem relacionadas por $x = x(r)$ teremos a correspondente transformação para as probabilidades

$$u(r)\,|dr| \;=\; w(x)\,|dx|\,. \tag{1.37}$$

Tomemos agora r uniforme em $[0,1]$ — ou seja $u(r)\,dr \;=\; dr$ — e tomemos x em $[a,b]$ dada pela distribuição normalizada $w(x)$

$$\int_a^b w(x)\,dx \;=\; 1\,. \tag{1.38}$$

De acordo com (1.37) podemos escrever a relação entre a variável uniforme r e a variável geral $x(r)$ como

$$\left|\frac{dr(x)}{dx}\right| dx \;=\; w(x)\,dx \tag{1.39}$$

e portanto, tomando os incrementos dx e dr de mesmo sinal, teremos a seguinte expressão para $r = r(x)$

$$r(x) \;=\; \int_a^x w(x')\,dx'\,. \tag{1.40}$$

Em geral, seja uma variável aleatória x com distribuição $w(x)\,dx$ em $[a,b]$. Se pudermos resolver (1.39) para $r(x)$ calculando a integral em (1.40) e pudermos inverter esta solução determinando $x(r)$, será possível gerar valores de x de acordo com sua distribuição de probabilidades $w(x)\,dx$ — ou seja **amostrar** x — simplesmente gerando valores de r em [0,1] como descrito na seção anterior, já que x será dado por $x(r)$. Note que a solução (1.40) corresponde à mudança de variáveis introduzida na Seção 1.3.2 em (1.27) no caso de $w(x)$ normalizada a $(b-a)$ e y distribuída uniformemente em $[a,b]$. Note também que no caso geral de n variáveis aleatórias x_i com $i = 1,\ldots,n$ a derivada em (1.39) será substituída pelo Jacobiano $\partial(r_1,\ldots,r_n)/\partial(x_1,\ldots,x_n)$ para as n variáveis uniformes r_i.

• **Distribuição linear:** Consideremos a distribuição de probabilidades normalizada $w(x)\,dx$ com densidade $w(x)$ dada por

$$w(x) \;=\; c_1 \,+\, c_2\,x, \tag{1.41}$$

onde os coeficientes c_1 e c_2 estão relacionados pela condição de normalização (1.38). Neste caso a integral (1.40) pode ser feita exatamente, fornecendo o resultado

$$r(x) \;=\; c_1(x-a) \,+\, \frac{c_2}{2}(x^2-a^2)\,. \tag{1.42}$$

Invertendo e tomando a solução com sinal positivo (que corresponde à solução com $x = a$ quando $r = 0$) temos

$$x(r) \;=\; -\frac{c_1}{c_2} \,+\, \sqrt{\left(\frac{c_1}{c_2}+a\right)^2 + \frac{2r}{c_2}}\,. \tag{1.43}$$

A equação acima fornece os valores de x com a distribuição linear desejada a partir de valores para r uniformemente distribuidos em [0,1]. Fazemos a seguir algumas verificações.

Claramente se $r = 0$ obtemos $x = a$. Impondo agora a condição de normalização (1.38) obtemos

$$c_1(b-a) + \frac{c_2}{2}(b^2 - a^2) = 1\,. \tag{1.44}$$

Usando este resultado podemos escrever $(c_1 + c_2 a)\,(b-a) = 2 - (c_1 + c_2 b)(b-a)$ e $2c_2(b-a)^2 = 4((c_1 + c_2 b)(b-a) - 1)$, ou seja $(c_1/c_2 + a)^2 + 2/c_2 = (c_1/c_2 + b)^2$ e claramente $x = b$ se $r = 1$. Finalmente, se r for uniformemente distribuída no intervalo [0, 1] teremos

$$\left|\frac{dr}{dx}\right| dx = (c_1 + c_2\, x)\, dx \tag{1.45}$$

ou seja x será distribuída com densidade $w(x)$ neste mesmo intervalo.

• **Distribuição exponencial:** Consideremos agora a distribuição de probabilidades $w(x)\,dx$ onde a densidade $w(x)$ é dada por

$$w(x) = e^{-x}\,. \tag{1.46}$$

Esta densidade é normalizada a 1 no intervalo $[0, \infty]$. Neste caso obtemos

$$r(x) = \int_0^x e^{-x'}\, dx' = 1 - e^{-x}\,, \tag{1.47}$$

ou seja

$$x(r) = -\log(1-r) \tag{1.48}$$

com r uniformemente distribuída in $[0, 1]$. Claramente temos $x(0) = 0$, $x(1) = \infty$ e

$$\left|\frac{dr}{dx}\right| dx = e^{-x} dx\,, \tag{1.49}$$

ou seja x é distribuída com densidade $w(x)$. Note que a relação

$$x(r) = -\log(r) \tag{1.50}$$

também nos permite obter a distribuição $e^{-x}dx$ e corresponde à solução de (1.40) com incrementos dx e dr de sinais opostos, ou seja, neste caso

$$r(x) = 1 - \int_a^x w(x')\, dx'\,, \tag{1.51}$$

satisfazendo $r(b) = 0$ e $r(a) = 1$.

• **Distribuição gaussiana:** O método de Box-Muller nos permite gerar duas variáveis aleatórias independentes com distribuição gaussiana [com média zero, variância um e normalizadas a 1 no intervalo $(-\infty, +\infty)$]

$$w(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,. \tag{1.52}$$

Para isso consideremos duas variáveis aleatórias independentes r_1 e r_2, uniformemente distribuídas em $[0,1]$ e definamos

$$\begin{aligned} x_1 &= \sqrt{-2\,\log r_1}\,\cos(2\,\pi\,r_2) \\ x_2 &= \sqrt{-2\,\log r_1}\,\sin(2\,\pi\,r_2) \end{aligned}$$

Estas relações podem ser invertidas, fornecendo r_1 e r_2 como funções de x_1 e x_2:

$$\begin{aligned} r_1 &= \exp\left(-\frac{x_1^2+x_2^2}{2}\right) \\ r_2 &= \frac{1}{2\pi}\arctan\left(\frac{x_2}{x_1}\right) \end{aligned}$$

Claramente quando $r_1,\, r_2 \in [0,1]$ temos que $x_1,\, x_2 \in (-\infty,\, +\infty)$. Além disso

$$dr_1\,dr_2 \;=\; |J|\,dx_1\,dx_2 \;=\; \frac{1}{2\pi}\,e^{-(x_1^2+x_2^2)/2}\,dx_1\,dx_2\;, \tag{1.53}$$

onde J é o Jacobiano $\partial(r_1,r_2)/\partial(x_1,(x_2)$. Note que para obter uma variável aleatória y com distribuição gaussiana (normalizada) com média a e variância σ é suficiente considerar

$$y \;=\; \sigma x \,+\, a \tag{1.54}$$

onde x é uma variável aleatória com distribuição gaussiana (normalizada) com média 0 e variância 1.

• **Distribuição uniforme sobre o círculo unitário:** Neste caso é suficiente gerar um ângulo ϕ uniformemente distribuído em $[0,2\pi]$, ou seja

$$\phi \;=\; 2\pi r \tag{1.55}$$

com r uniformemente distribuída em $[0,1]$. O ponto sobre o círculo é dado então por

$$\begin{aligned} x &= \cos\phi \\ y &= \sin\phi \end{aligned}$$

Nota: este método aplica-se também à amostragem uniforme de variáveis do grupo $U(1)$, ou seja números complexos $x \,+\, i\,y$ de módulo 1.

• **Matrizes** $SU(2)$**:** Uma matriz $g \in SU(2)$ pode ser parametrizada do seguinte modo

$$g \;\equiv\; g_0\,\mathbb{1} + i\vec{g}\cdot\vec{\sigma} \;=\; \begin{pmatrix} g_0+ig_3 & g_2+ig_1 \\ -g_2+ig_1 & g_0-ig_3 \end{pmatrix}\;, \tag{1.56}$$

onde $\mathbb{1}$ é a matriz identidade 2×2, g_i com $i=0,1,2,3$ são números reais e as componentes do vetor $\vec{\sigma} \equiv (\sigma^1,\sigma^2,\sigma^3)$ são as três matrizes de Pauli. É fácil verificar que

a matriz escrita assim satisfaz a relação $g\,g^\dagger = \mathbb{1}$, onde $g^\dagger$ indica a matriz complexa conjugada de g. Além disso, a condição $\det g = 1$ é satisfeita se $g_0^2+g_1^2+g_2^2+g_3^2 = 1$. Isto implica que gerar uma matriz $SU(2)$ é equivalente a gerar um vetor real com quatro componentes e módulo unitário, ou seja um vetor $(x_1,\, x_2\, x_3\, x_4)$ na esfera unitária quadri-dimensional. Claramente, este vetor pode ser obtido extendendo a quatro dimensões o método ilustrado acima no caso bi-dimensional, ou seja criar pontos uniformemente distribuídos em um cubo quadri-dimensional de lado 2 e aceitar somente pontos para os quais a condição $r^2 \leq 1$ seja satisfeita. Mais exatamente

1) sejam $r_1,\, r_2\, r_3,\, r_4$ uniformemente distribuídas em $[-1,\, 1]$ e seja $r^2 = r_1^2+r_2^2+r_3^2+r_4^2$;

2) o vettore $(r_1,\, r_2\, r_3\, r_4)$ é aceito se $r^2 \leq 1$ e recusado em caso contrário; neste segundo caso volta-se ao passo anterior;

3) toma-se como punto na sfera unitária quadri-dimensionale

$$\begin{aligned} x_1 &= r_1/r \\ x_2 &= r_2/r \\ x_3 &= r_3/r \\ x_4 &= r_4/r \end{aligned}$$

In questo caso la probabilidade de aceitação é igual a $\pi^2/32 \approx 0.308$. Una probabilita' di accettazione due volte maggiore, vale a dire uguale a $\pi^2/16 \approx 0.616$ puo' essere ottenuta generando due coppie di punti, (x_1, y_1) e (x_2, y_2) nel círculo unitário usando il metodo sopra descritto e prendendo come punto na sfera unitária quadri-dimensional

$$\begin{aligned} x_1 &= x_1 \\ x_2 &= y_1 \\ x_3 &= x_2\sqrt{(1-r_1^2)\,/\,r_2^2} \\ x_4 &= y_2\sqrt{(1-r_1^2)\,/\,r_2^2} \end{aligned}$$

dove $r_1^2 = x_1^2 + y_1^2$ e $r_2^2 = x_2^2 + y_2^2$. Chiaramente $x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1$ e uno puo' verificare che

1.4.3 Método da rejeição

Seja a varíavel aleatória x em $[a, b]$ com distribuição de probabilidades normalizada $f(x)$. Mesmo quando não for possível a amostragem direta de x por métodos como os descritos na seção anterior, será ainda possível a sua amostragem exata se pudermos encontrar uma distribuição $g(x)$ satisfazendo $g(x) \geq f(x)$ para todo x em $[a, b]$ a qual saibamos amostrar diretamente. (Veja a Fig. 1.4.) Neste caso o método da rejeição consiste em

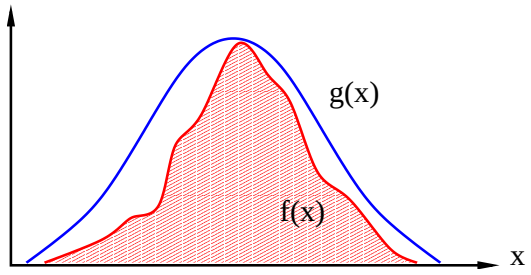


Figura 1.4: Método da rejeição para amostragem de distribuições.

1) gerar valores de x com distribuição proporcional a $g(x)$

2) aceitar cada valor com probabilidade $f(x)/g(x)$

Claramente os valores aceitos para x terão distribuição proporcional a $f(x)$, como desejado. Note que a distribuição $g(x)$ não pode ser normalizada, já que apresenta valor sempre maior do que $f(x)$ a qual é normalizada a 1 no intervalo. O importante é manter a distribuição de probabilidade relativa para diferentes valores de x. Note também que $g(x)$ deve ser escolhida tão próxima quanto possível de $f(x)$ pois a aceitação média será dada por

$$A = \frac{\int_{-\infty}^{\infty} f(x)\,dx}{\int_{-\infty}^{\infty} g(x)\,dx}\,. \tag{1.57}$$

Apresentamos a seguir alguns exemplos do uso deste método.

• **Distribuição uniforme sobre o círculo unitário:** Este caso foi discutido na seção anterior. Em geral porém o cálculo das funções trigonométricas $\cos\phi$ e $\sin\phi$ é bastante lento e em alguns casos pode ser mais rápido utilizar o seguinte método:

1) sejam r_1, r_2 uniformemente distribuídas em $[-1,\,1]$ e seja $r^2 = r_1^2 + r_2^2$;

2) o par $(r_1,\,r_2)$ é aceito se $r^2 \leq 1$ e recusado em caso contrário; neste segundo caso volta-se ao passo anterior;

3) toma-se como ponto no círculo unitário

$$\begin{aligned} x &= r_1/r \\ y &= r_2/r \end{aligned}$$

Claramente os pontos

$$\begin{aligned} x' &= r_1 = r\,\cos\phi \\ y' &= r_2 = r\,\sin\phi \end{aligned}$$

estão contidos no círculo de raio 1 e são uniformemente distribuídos; além disso

$$dr_1\,dr_2 = r\,dr\,d\phi\,, \tag{1.58}$$

ou seja a distribuição em $d\phi$ é uniforme. Note por outro lado que a distribuição em dr não é uniforme, mas a distribuição $r\,dr$ é o que precisamos para ter pontos distribuídos uniformemente no círculo de raio 1.

Note também que a probabilidade de aceitação do ponto 2) é bastante elevada, ou seja igual a $\pi/4 \approx 0.785$.

1.4.4 Exercícios

1) Escrever um gerandor de números aleatórios usando o método congruencial linear com $a = 137, c = 187$ and $m = 256$ e verificar que o período deste gerador é igual a m e que este gerador não passa no teste espectral em duas dimensões, ou seja os pontos (r_n, r_{n+1}) não são uniformemente distribuídos no quadrado $[0,1] \times [0,1]$ (fazer um gráfico).

2) Escrever um programa que verifique para qual valor de n o computador fornece um valor negativo para o número 2^n.

3) Escrever um gerador de números aleatórios usando o método congruencial linear com $a = 65539, c = 0$ and $m = 2^{31}$. Este gerador era muito usado pelas máquinas IBM nos anos sessenta. Note que podemos escrever $a = 3 + 2^{16}$.

4) Verificare analiticamente che l'integrale

$$I \;=\; 2 \int_0^1 dx \int_0^1 dy \int_0^1 dz \, \sin^2(2\,\pi\,(9x - 6y + z)) \tag{1.59}$$

e' uguale a 1. Calcolare numericamente lo stesso integrale usando il gerador de números aleatórios do exercício precedente con $x = r_n, y = r_{n+1}$ e $z = r_{n+2}$ e verifcare che il risultato e' zero! Al fine di capire questo risultato verificare que a combinação $9x - 6y + z$ produz números inteiros (positivos ou negativos), ovvero este gerador não passa no teste espectral em três dimensões.

5) Dimostrare che punti prodotti nel modo seguente sono distribuiti uniformemente nella sfera (tri-dimensionale) di raggio unitario:

- 1) siano r_1, r_2 uniformemente distribuite in $[-1,\, 1]$ e sia $r^2 = r_1^2 + r_2^2$;
- 2) la coppia $(r_1,\, r_2)$ e' accettata se $r^2 \leq 1$ e rifiutata altrimenti; in questo secondo caso si ritorna al passo precedente;
- 3) si prende come punto nella sfera di raggio unitario

$$\begin{aligned} x &= 2\,r_1\,\sqrt{1 - r^2} \\ y &= 2\,r_2\,\sqrt{1 - r^2} \\ z &= 1 - 2\,r^2 \end{aligned}$$

6) Generare matrici $SU(2)$ usando $U(1)$-embedding. Nota: la matrice

$$g = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & 1 \end{pmatrix}$$

non e' una matrice $SU(2)$.

1.5 Método de Monte Carlo dinâmico

Passaremos a considerar agora problemas físicos em que a distribuição de probabilidade a ser amostrada depende de um grande número de variáveis e é dada por um peso estatístico fortemente concentrado em uma região do espaço de estados. Como exemplo consideremos a distribuição de Boltzmann (1.31). Como mencionado anteriormente, a amostragem simples neste caso é inviável, e devemos encontrar métodos para amostragem por importância do peso estatístico (1.31). Infelizmente este peso é complicado demais para permitir uma amostragem independente como fizemos até aqui, mas será possível uma amostragem em que as configurações para o sistema (correspondendo às variáveis x_i para $i = 1, \ldots, N$) sejam igualmente distribuídas mas *correlacionadas*. Neste caso a média será tomada como de costume, mas o erro terá que ser tratado com cuidado. Como veremos na Seção 1.6, seu comportamento será ainda com $1/\sqrt{N}$, mas a constante de proporcionalidade estará ligada à correlação entre as amostras.

O método para produção de amostras com a correta distribuição de probabilidades baseia-se na teoria de cadeias de Markov. Define-se como cadeia de Markov um processo estocástico que é totalmente determinado por uma probabilidade fixa de transição entre seus estados (a matriz de transição). Cada transição é chamada de um passo, e a evolução de tais passos é chamada de cadeia. Discutimos estes processos em mais detalhe abaixo. A propriedade importante é que, dadas certas condições para a probabilidade de transição, a distribuição de probabilidades para os estados da cadeia evolui para uma distribuição estacionária, independente do tempo de observação da cadeia e da distribuição de probabilidades inical. Esta distribuição é única. Portanto, no regime estacionário, médias no tempo correspondem a médias estatísticas na distribuição estacionária. A estratégia dos métodos de Monte Carlo dinâmicos é portanto a criação de uma cadeia de Markov tal que a distribuição estacionária seja a distribuição desejada (por exemplo a distribuição de Boltzmann).

1.5.1 Algoritmo de Metropolis

1.5.2 Aplicação ao modelo de Ising

1.5.3 Algoritmos de banho térmico

1.5.4 Exercícios

1.6 Tratamento de erros

1.6.1 Método de binning

1.6.2 Método da janela auto-consistente

1.6.3 Métodos de jack-knife e bootstrap

1.6.4 Exercícios

1.7 O fenômeno de freiamento crítico

1.7.1 Métodos de sobre-relaxação

1.7.2 Algoritmos de aglomerados

Discutimos a seguir o algoritmo de aglomerados (clusters) de Swendsen-Wang para modelos de spins discretos de Potts sem campo magnético. A idéia central deste algoritmo foi desenvolvida por Fortuin e Kasteleyn em 1969, em conexão com o modelo de aglomerados aleatórios, e mais tarde adaptada para simulações de Monte Carlo por Swendsen e Wang, em 1987.

O modelo de Potts é uma generalização do modelo de Ising em que spins em sítios vizinhos tendem a apresentar o mesmo valor, com Hamiltoniana dada — no caso sem campo magnético e para uma configuração de spins $S \equiv (S_1, S_2, \ldots, S_V)$ em uma rede com V sítios — por

$$\mathcal{H}(S) \;=\; \sum_{<ij>} J_{ij}\,(1 \,-\, \delta_{S_i S_j})\,, \tag{1.60}$$

onde os spins S_i assumem valores discretos de 1 a q, a soma em $< ij >$ é sobre primeiros vizinhos e fazemos uso da função delta de Kronecker

$$\delta_{nm} \;=\; \begin{cases} 1 & \text{se } n = m \\ 0 & \text{se } n \neq m \end{cases} \tag{1.61}$$

Note que o modelo de Ising corresponde ao caso $q = 2$ e acoplamento $J = J_{ij}/2$ em analogia com nossa definição para a Hamiltoniana de Ising na Seção 1.5.2. (Para variáveis de Ising pode-se escrever $\delta_{S_i S_j} = (1 + S_i S_j)/2$.)

Consideremos primeiramente a função de partição do sistema

$$\begin{aligned}\mathcal{Z} &= \sum_{\{S\}} e^{-\beta\,\mathcal{H}(S)} = \sum_{\{S\}} e^{-\beta\,\sum_{<ij>}\,J_{ij}\,(1\,-\,\delta_{S_iS_j})} = \sum_{\{S\}}\prod_{<ij>} e^{-\beta\,J_{ij}}\,e^{\beta\,J_{ij}\,\delta_{S_iS_j}} \\ &= \sum_{\{S\}}\prod_{<ij>} e^{-\beta\,J_{ij}}\,\left(\delta_{S_iS_j}\,e^{\beta\,J_{ij}}\,+\,1\,-\,\delta_{S_iS_j}\right) \\ &= \sum_{\{S\}}\prod_{<ij>}\left[p_{ij}\,\delta_{S_iS_j}\,+\,(1\,-\,p_{ij})\right], \qquad (1.62)\end{aligned}$$

onde definimos $p_{ij} \equiv 1 - e^{-\beta\,J_{ij}}$. Note que representamos as somas nas configurações S por $\sum_{\{S\}} \equiv \sum_{S_1}\,\sum_{S_2}\ldots\sum_{S_V}$. Aplicando agora a identidade trivial

$$a\,+\,b\,=\,\sum_{n\,=\,0,1}\,(a\,\delta_{n1}\,+\,b\,\delta_{n0}) \qquad (1.63)$$

a cada elo $<ij>$ — ou seja introduzindo para cada elo uma nova variável $n_{ij} = 0,\,1$ — obtemos

$$\begin{aligned}\mathcal{Z} &= \sum_{\{S\}}\sum_{\{n\}}\prod_{<ij>}\left[p_{ij}\,\delta_{n_{ij}1}\,\delta_{S_iS_j}\,+\,(1\,-\,p_{ij})\,\delta_{n_{ij}0}\right] \\ &= \sum_{\{S\}}\sum_{\{n\}}\left[\prod_{<ij>:\,n_{ij}=1} p_{ij}\,\delta_{S_iS_j}\right]\left[\prod_{<ij>:\,n_{ij}=0}(1\,-\,p_{ij})\right]. \qquad (1.64)\end{aligned}$$

Podemos observar que após fazer algumas manipulações à função de partição original obtivemos uma descrição do modelo em termos da soma de um peso estatístico não-normalizado — dado pelos termos entre colchetes — para todas as configurações das variáveis de spins, assim como para todas as configurações das variáveis n_{ij} que introduzimos para os elos. Evidentemente as variáveis n_{ij} não são físicas, e o peso para uma configuração de spins S para o modelo de Potts não deve depender delas. Este peso normalizado é dado por

$$\mu_{Potts}(S)\,=\,\frac{e^{-\beta\,\mathcal{H}(S)}}{\mathcal{Z}}\,\equiv\,\sum_{\{n\}}\mu_{FKSW}(S,n)\,, \qquad (1.65)$$

onde definimos o modelo de Fortuin-Kasteleyn-Swendsen-Wang (FKSW) para as variáveis de sítios e elos através da distribuição de probabilidades

$$\mu_{FKSW}(S,n)\,\equiv\,\frac{1}{\mathcal{Z}}\left[\prod_{<ij>:\,n_{ij}=1} p_{ij}\,\delta_{S_iS_j}\right]\left[\prod_{<ij>:\,n_{ij}=0}(1\,-\,p_{ij})\right]. \qquad (1.66)$$

Podemos notar duas características importantes na distribuição acima

- As configurações deste modelo possuindo o peso acima diferente de zero são formadas por elos n_{ij} "ocupados" (corresondendo a $n_{ij} = 1$) apenas entre sítios i, j com spins de mesmo valor. Isto é, os elos entre sítios de spins diferentes serão

sempre vazios ($n_{ij} = 0$), dando uma contribuição fixa igual a $(1 - p_{ij})$ para a função de partição. Para um elo $<ij>$ tal que $S_i = S_j$ há a possibilidade de ele estar ocupado ou vazio com contribuições respectivamente p_{ij} e $(1 - p_{ij})$ para a função de partição. Dizemos portanto que um elo deste tipo está ocupado com probabilidade p_{ij} e vazio com probabilidade $(1 - p_{ij})$.

- A dependência nos spins é trivial: dada uma configuração para os elos n_{ij} os spins devem possuir valor igual dentro de cada aglomerado (cluster) definido pelos pontos ligados por elos ocupados. Já que o peso não depende do valor específico dos spins em cada aglomerado, este valor é distribuído com igual probabilidade entre os q valores possíveis ou seja com probabilidade $1/q$, independentemente dos outros aglomerados. Portanto, o número total de configurações de spins com mesmo peso para uma configuração fixa de elos é $q^{\mathcal{N}(n)}$ onde $\mathcal{N}(n)$ é o número de aglomerados para a configuração de elos n.

Baseado na segundo observação acima podemos inverter a ordem das somas em (1.64) entre sítios e elos e realizar a soma nos spins trivialmente, definindo o modelo de aglomerados aleatórios de Fortuin-Kasteleyn (ou RC: "random cluster model")

$$\begin{aligned} \mu_{RC}(n) &= \sum_{\{S\}} \mu_{FKSW}(S,n) \\ &= \frac{1}{\mathcal{Z}} \left[\prod_{<ij>:n_{ij}=1} p_{ij} \right] \left[\prod_{<ij>:n_{ij}=0} (1 - p_{ij}) \right] q^{\mathcal{N}(n)} . \end{aligned} \tag{1.67}$$

Note que a função de partição $\mathcal{Z}$ é a mesma para os três modelos. A diferença entre os modelos é dada pela maneira de interpretarmos $\mathcal{Z}$ como soma de um peso estatístico nas configurações de variáveis. Considerando as variáveis como spins, elos ou elos e spins teremos respectivamente os modelos de Potts, RC e FKSW, com as respectivas distribuições de probabilidades dadas acima. Ou seja, obtivemos dois novos modelos definidos a partir do modelo de Potts. Notemos primeiramente que o modelo de RC possui uma distribuição simples para o caso $q = 1$: elos independentes, ocupados com probabilidade p_{ij} e vazios com probabilidade $(1 - p_{ij})$. Este é o modelo de percolação simples de elos. Notemos agora que o modelo FKSW, que pode ser simulado de maneira eficiente como descreveremos abaixo, é na verdade equivalente ao modelo de Potts em que estávamos originalmente interessados, com a condição de que sejam estudados apenas observáveis de spins. De fato, as médias para um observável A que só dependa das variáveis de spins serão idênticas se tomadas na distribuição $\mu_{Potts}(S)$ ou $\mu_{FKSW}(S,n)$

$$\begin{aligned} < A(S) >_{Potts} &= \sum_{S} A(S)\, \mu_{Potts}(S) \\ &= \sum_{S} \sum_{n} A(S)\, \mu_{FKSW}(S,n) \;\; = < A(S) >_{FKSW} \end{aligned} \tag{1.68}$$

pela própria definição de $\mu_{FKSW}(S,n)$, Eq. (1.65).

Descrevemos a seguir o algoritmo de Swendsen-Wang para simulação da distribuição $\mu_{FKSW}(S,n)$, já que médias dos observáveis usuais serão idênticas às tomadas para o modelo de Potts. O algoritmo se baseia no método de re-amostragem

parcial, que já encontramos para os algoritmos locais de Metropolis e banho térmico para modelos de spins. Este método afirma que uma forma válida de produzir os estados da cadeia de Markov com a distribuição estacionária desejada é considerar passos em que as variáveis sejam atualizadas por partes: são mantidas fixas parte das variáveis e são escolhidos novos valores para as variáveis restantes de forma a preservar a distribuição estacionária, alternando-se os grupos de variáveis fixas. Em nosso caso, vimos que apesar de spins e elos não serem independentes as distribuições condicionais (isto é, as distribuições obtidas mantendo-se parte das variáveis fixas) nos dois casos são bastante simples. Podemos então claramente definir passos formados mantendo primeiramente fixos os spins e gerando novos elos e em seguida mantendo fixos os elos e gerando novos spins.

Portanto, usando as distribuições condicionais discutidas acima podemos escrever da seguinte maneira o algoritmo de aglomerados de Swendsen-Wang para simulação da distribuição de probabilidades μ_{FKSW} dada em Eq. (1.66)

1) mantendo fixas as variáveis S escolher as variáveis n da seguinte maneira: se $S_i \neq S_j$ tomamos $n_{ij} = 0$ e se $S_i = S_j$ tomamos $n_{ij} = 1$ com probabilidade p_{ij} e $n_{ij} = 0$ com probabilidade $(1 - p_{ij})$

2) em seguida, mantendo fixas as variáveis n escolhemos as variáveis S trivialmente: todos os S_i em um aglomerado tomam o mesmo valor entre os q valores possíveis, independentemente para cada aglomerado. Para o caso de Ising isto consiste em refletir ou não todos os spins em cada aglomerado com probabilidade $1/2$.

O passo número 1) é sem dúvida mais difícil de ser implementado que o número 2) já que é nele que serão identificados e armazenados os aglomerados. Uma vez que isto esteja feito o passo 2) é trivial. A identificação dos aglomerados é também a parte mais lenta da simulação, e portanto deve ser feita da maneira mais eficiente possível. Descrevemos no Apêndice A alguns algoritmos para identificação de aglomerados.

Definimos portanto um algoritmo válido para a atualização de configurações de spins através da introdução de variáveis adicionais para os elos da rede. O papel das variáveis de elo é o de determinar estruturas formadas por muitos sítios — os aglomerados — que serão atualizadas conjuntamente, ou seja de maneira *global*. Isto é muito diferente do que observamos nos casos de algoritmos locais, em que as atualizações eram feitas beseadas em critérios locais, como por exemplo a amostragem exata da probabilidade condicional para cada sítio, no algoritmo de banho térmico. A grande vantagem da atualização global está na possibilidade de grandes deslocamentos no espaço de configurações em um único passo, de forma que a distribuição estacionária seja preservada e de forma que o passo dado resulte em uma configuração fisicamente relevante para o sistema. A escolha dos modos globais a serem atualizados é a grande dificuldade de algoritmos deste tipo, chamados de algoritmos globais. Ao contrário de algoritmos baseados em atualizações por modos globais fixos — por exemplo blocos de spins com lado variável (como no método de multi-grid) ou valores fixos do momento em uma descrição do sistema no espaço de momentos (como no método de

aceleração de Fourier) — o algoritmo de aglomerados baseia-se em modos globais de certa forma "escolhidos" pelo próprio sistema.

Embora não seja aplicável para uma classe geral de sistemas — por exemplo teorias de gauge na rede ou qualquer modelo tomando valores em um grupo — o desempenho do algoritmo de aglomerados para modelos de spins é extraordinário: o fenômeno de freiamento crítico é praticamente eliminado, encontrando-se expoente crítico z próximo a zero.

O algoritmo pode ser implementado também para o caso com campo magnético, de duas maneiras diversas: escolhendo o novo valor para os spins de um cluster com probabilidade dependente do campo, ou através da introdução de um "spin fantasma", representando o campo magnético e ligado a todos os spins com a interação J apropriada. Este spin é incluído ou nãso em aglomerados e atualizado da mesma forma que os spins reais, e portanto um aglomerado neste caso podem ser formado por componentes disjuntas na rede, se estiverem todas ligadas pelo spin fantasma.

Outra generalização possível é para spins contínuoscom simetria $O(n)$, através da técnica de "embedding": escolhe-se uma direção aleatória e os spins são temporariamente vistos como variáveis de Ising dadas pelo sinal da projeção de cada spin ao longo da direção escolhida, ao passo que a interação entre dois spins é dada pelo produto dos módulos destas projeções para os dois sítios. São então atualizadas as variáveis de Ising com o método de aglomerados usual, e é incorporada a atualização , ou seja a inversão ou não da projeção de cada spin na direção escolhida. Para cada passo é escolhida uma nova direção de forma a manter a ergodicidade do algoritmo.

Finalmente, notamos que ao invés de atualizar todos os aglomerados a cada passo do algoritmo pode-se considerar apenas um aglomerado, tomado por exemplo a partir da escolha aleatória de um sítio da rede. Desta forma são priorizados aglomerados com número maior de sítios, e os deslocamentos no espaço de configurações são maiores. Embora a análise de eficiência seja mais difícil neste caso, pois deve-se considerar que cada passo do algoritmo não visita todos os sítios mas apenas aqueles que estiverem no aglomerado selecionado, obtém-se que para esta versão do método — introduzida por Wolff — o expoente z é ainda menor do que no algoritmo original.

1.8 Projetos

1.9 Bibliografia

- Sokal, A. 1996. *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms.* Escola de verão "Functional Integration: Basics and Applications" em Cargèse, `http://citeseer.nj.nec.com/sokal96monte.html`. (Versão atualizada do "Cours de Troisième Cycle de la Physique en Suisse Romande", de 1989).

- Koonin, S.E. & Meredith, D. 1990. *Computational Physics, Fortran Version.* Addison-Wesley, Redwood City.

- Press, W.H.; Flannery, B.P.; Teukolsky, S.A. & Vetterling, W.T. 1988–1992. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge. (Várias versões. Veja também `http://www.nr.com/`.)

Capítulo 2

Física dos Fenômenos Críticos

2.1 Introdução

São chamadas de fenômenos críticos as propriedades de sistemas na vizinhança de uma transição de fase de segunda ordem, ou **ponto crítico**. Como primeiro exemplo de transições de fase, consideremos a transição líquido/gás em fluidos, cujo diagrama de fases em função da pressão e da temperatura está representado na Fig. 2.1.

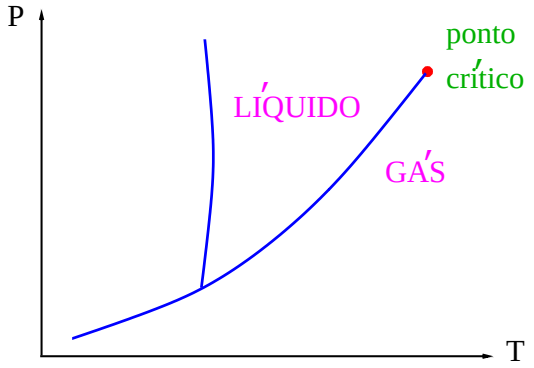


Figura 2.1: Diagrama de fases para o sistema líquido-gás.

Observamos que partindo da fase líquida e considerando um valor fixo para a pressão podemos representar o aumento de temperatura como um movimento da esquerda para a direita começando na fase líquida para temperaturas mais baixas e passando para a fase gasosa ao cruzar a linha de transição líquido/gás. Ao longo da linha de transição há diferença entre as densidades da fase líquida e gasosa, e a transição é descontínua ou de primeira ordem, com coexistência de fases e calor latente. Esta diferença de densidades diminui até chegar a zero no ponto crítico, e é tomada como **parâmetro de ordem** para o sistema.

Este comportamento é análogo ao observado na vizinhança do ponto correspondente para um material ferromagnético, cujos diagramas de fase estão representados na Fig. 2.2. Neste caso o parâmetro de ordem, que no fluido é a diferença de densidade entre as fases, vai ser dado pela magnetização. As duas fases são dadas por magnetizações opostas, alinhadas ao campo magnético externo. Note que abaixo da temperatura crítica T_c há magnetização mesmo quando o campo H tende a zero. Nos dois casos o ponto que chamamos de crítico possuirá propriedades muito especiais, ligadas ao comportamento singular de várias grandezas físicas. De fato, observa-se que grandezas termodinâmicas divergem no ponto crítico como potências da temperatura

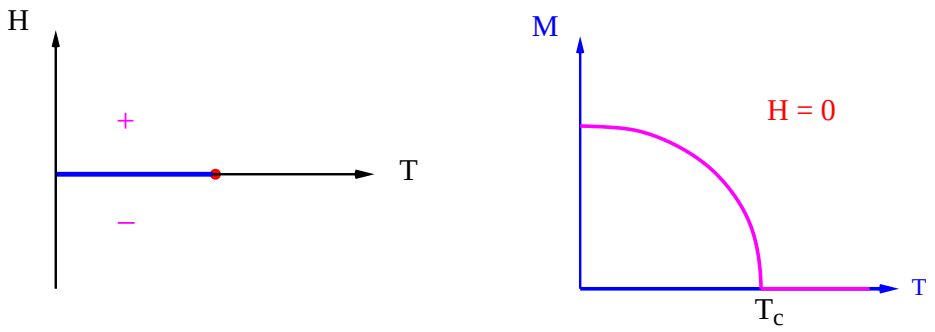


Figura 2.2: Diagrama de fases para o magneto.

reduzida t, que mede a distância da temperatura T ao seu valor crítico T_c

$$t \equiv (T - T_c)/T_c \,. \tag{2.1}$$

O ponto crítico corresponde a t tendendo a zero, t negativo corresponde a estar abaixo e t positivo a estar acima de T_c. Vemos abaixo exemplos de expoentes críticos no comportamento para $t \to 0$:

Calor específico	$C \sim \|t\|^{-\alpha}$
Parâmetro de ordem	$M \sim \|t\|^{\beta}$
Suscetibilidade	$\chi \sim \|t\|^{-\gamma}$
Comprimento de correlação	$\xi \sim \|t\|^{-\nu}$

Grandezas com expoentes negativos divergem, ao passo que o parâmetro de ordem é zero para t positivo e vai a zero com expoente β para t negativo. A suscetibilidade está associada à flutuação do parâmetro de ordem, e diverge à medida que o comprimento de correlação ξ diverge. **Obs:** este comportamento com potências de t pode ser 1) observado experimentalmente 2) obtido da hipótese de escala (que afirma que ξ é o comprimento de escala relevante e diverge à temperatura crítica) usando-se análise dimensional; obtêm-se assim também as relações de hiper-escala entre os expoentes e 3) dado pelo do grupo de renormalização. A propriedade de universalidade pode ser vista na Fig. 2.3, dos anos 40, contendo oito sistemas de fluidos diversos. Os valores de T_c e n_c são muito diferentes, mas após serem apropriadamente rescalados, todos os pontos caem sobre a mesma curva. Podemos concluir que um programa de estudo de transições de fase de segunda ordem consiste em construir um modelo para a interação envolvida, dado pela Hamiltoniana $\mathcal{H}$, e obter expressões para os observáveis (e.g. M, χ) a partir da função de partição $\mathcal{Z}$ e energia livre F definidas por $\mathcal{H}$. Espera-se que um modelo simples, que incorpore as simetrias relevantes, seja suficiente para a descrição do comportamento na região crítica.

$$\mathcal{Z} \;=\; \int e^{-\beta\mathcal{H}}\,; \quad F \;=\; -\frac{1}{\beta}\log\mathcal{Z}$$

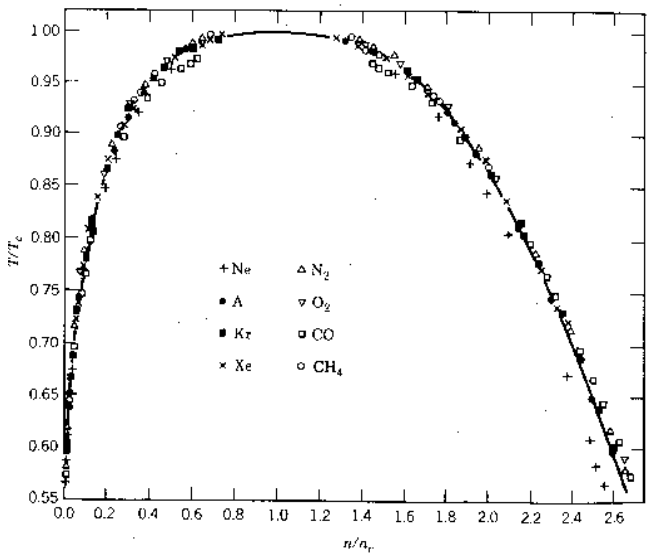
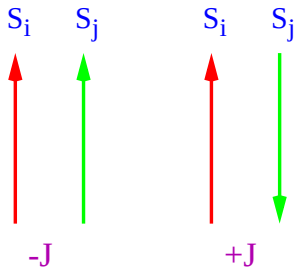


Figura 2.3: Dados experimentais para sistemas de fluidos.

2.2 Exemplos de fenômenos críticos

2.2.1 Modelos de spins



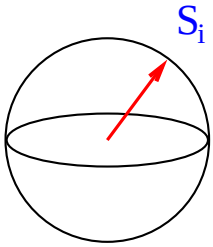
O modelo de ferromagnetismo mais simples é o famoso **modelo de Ising**, que introduzimos na Seção 1.5.2, definido pela Hamiltoniana

$$\mathcal{H} = -J \sum_{<i,j>} S_i S_j + H \sum_i S_i \, . \qquad (2.2)$$

Ele incorpora o fato de spins vizinhos quererem alinhar-se uns com os outros e com o campo, deixando aos spins apenas a liberdade entre dois estados: apontando para cima (direção do campo) ou para baixo. Apesar de ser tão simples o modelo apresenta transição de fase de segunda ordem já em dimensão d igual a 2, caso em que pode ser resolvido exatamente (a campo nulo). Esta é a famosa solução de Onsager, dos anos 40, que contribuíu muito para a mecânica estatística: na época não se acreditava que a descrição das propriedades da transição de fase, como por exemplo os expoentes críticos, estivesse contida na Hamiltoniana ou na função de partição do sistema. (Faziam-se tentativas de adicionar termos auxiliares à Hamiltoniana que contivessem informação sobre a transição.) Em dimensão 3, resultados para os expoentes obtidos por teoria de perturbação ou simulações de Monte Carlo estão de acordo com observações experimentais de diversos sistemas de fluidos. O modelo reproduz o comportamento esperado para o parâmetro de ordem — a magnetização — a campo nulo: tendendo a zero continuamente ao aproximar-se da temperatura crítica por baixo, e zero acima de T_c. Note que a Hamiltoniana sem campo é simétrica por reflexão simultânea de

todos os spins. Esta simetria é quebrada explicitamente pela introdução do campo magnético, e não é respeitada pelo parâmetro de ordem a baixas temperaturas, mesmo a campo nulo. Portanto, abaixo de T_c tem-se uma **quebra espontânea de simetria**, tratando-se neste caso de uma simetria discreta.

Uma outra classe de modelos com simetria discreta são os **modelos de Potts**, que introduzimos na Seção 1.7.2, os quais representam uma generalização do modelo de Ising para um número q de estados discretos. **COMPLETAR...**



Os **modelos** $O(n)$, ou n-vetoriais correspondem a uma generalização do modelo de Ising para o caso da simetria contínua de rotação. As variáveis de spin são tomadas como vetores numa esfera unitária em um espaço de n dimensões ($n \geq 2$) e a Hamiltoniana

$$\mathcal{H} = -J \sum_{<i,j>} \mathbf{S}_i \cdot \mathbf{S}_j + H \sum_i \mathbf{S}_i \qquad (2.3)$$

é definida em termos do produto escalar de spins em sítios vizinhos. A principal diferença em relação ao modelo de Ising é que são agora possíveis configurações em que os spins se encontrem localmente aproximadamente alinhados mas para grandes distâncias o alinhamento é perdido, resultando em uma média nula para a magnetização. Tais configurações — chamadas ondas de spins — possuem energia arbitrariamente baixa, e tenderão a destruir a ordem do sistema mesmo a baixas temperaturas (veja a Fig. 2.4).

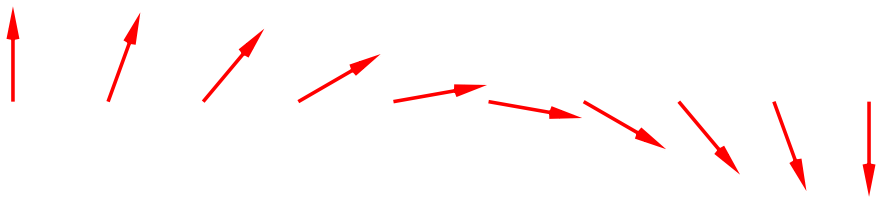


Figura 2.4: Onda de spins.

De fato, ao contrário do modelo de Ising, estes modelos não apresentam transição de fase com magnetização espontânea em duas dimensões, como demonstrado pelo teorema de Mermim-Wagner.[1] Em três dimensões há transição de fase e há a presença de magnetização espontânea abaixo de uma temperatura crítica. Neste caso a quebra da simetria contínua de rotação a baixas temperaturas (dada pela magnetização espontânea) está associada a modos de Goldstone, as ondas de spin, que causam divergência da suscetibilidade a campo zero não só ao redor da temperatura crítica, mas também para toda a fase de baixas temperaturas.

Os modelos $O(n)$ são de interesse geral para a mecânica estatística: o caso $n = 2$, ou modelo XY, descreve a transição de fase para o hélio super-fluido e o caso $n = 3$ corresponde à versão clássica do modelo de Heisenberg quântico para magnetos. Além

[1] Para o caso $n = 2$ há transição de tipo Kosterlitz-Thouless, sem magnetização espontânea.

disso, acredita-se que o caso $O(4)$ esteja diretamente relacionado à transição de fase de desconfinamento de quarks na cromodinâmica quântica, que discutiremos na Seção 2.2.3.

2.2.2 Percolação

A teoria de percolação permite um estudo das propriedades geométricas associadas a fenômenos críticos em numerosas áreas da física, como por exemplo o comportamento de fluidos em passagem por meios porosos, a propagação de incêndios em florestas ou as propriedades de condutividade elétrica de ligas metálico-isolantes. Modelos para percolação são definidos através de uma prescrição para a formação de aglomerados. Diz-se que há "percolação" quando um destes aglomerados extender-se de uma parte a outra do sistema, considerado no limite de volume infinito. Por exemplo o modelo de percolação de sítios de uma rede é definido dando-se uma probabilidade p de ocupação para cada sítio, e consideram-se como aglomerados os grupos de pontos contíguos ocupados. O modelo é então estudado em função do parâmetro p, e são considerados os comportamentos de observáveis como por exemplo a probabilidade de percolação $\mathcal{P}(p)$ e o tamanho médio dos aglomerados não-percolantes $\mathcal{S}(p)$. Apesar de sua simplicidade, modelos de percolação exibem transições de fase com comportamento crítico análogo ao de sistemas mais complicados, sendo por exemplo associados expoentes críticos aos observáveis geométricos, como vemos a seguir. **completar...**

Consider

- p: density of occupied sites (or bonds)
- p_c: lowest p such that origin belongs to the percolating cluster

then

$$\begin{aligned} P &\sim (1 - p/p_c)^{\beta} \quad p \to p_{c_+} \\ S &\sim (p_c - p)^{-\gamma} \quad p \to p_c \end{aligned}$$

simple site (or bond) percolation fails to reproduce Ising-model exponents

$$\text{geom. cluster} \geq \text{droplet}$$

Solution: site percolation + T-dependent bond probability

$$p_b = 1 - e^{-2\,J/KT}$$

then get agreement for exponents, and percolation point corresponds to T_c.

2.2.3 Transição de desconfinamento em QCD

Um tipo de transição de fase bem diferente é o da transição de fase de desconfinamento de quarks na Cromodinâmica Quântica (QCD). A QCD é a teoria que explica a interação forte entre hádrons — por exemplo prótons e nêutrons — através de um modelo de partículas elementares chamadas quarks, dotadas de carga de cor (em três tipos: azul, vermelha e amarela, sendo que a combinação das três cores é neutra)

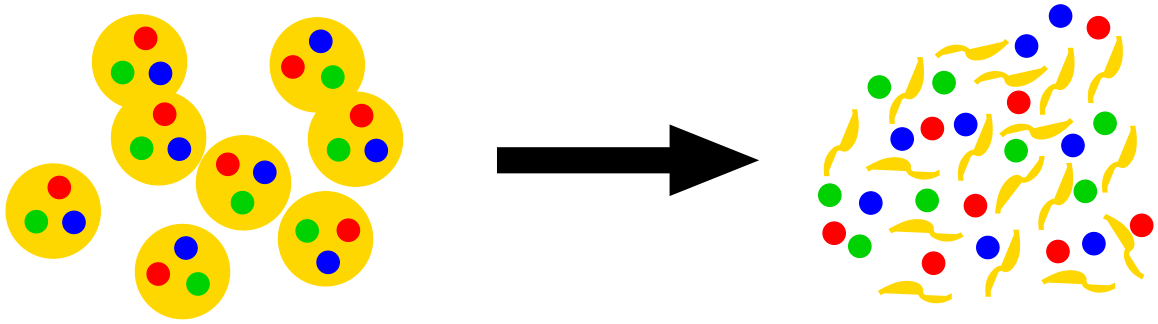


Figura 2.5: Transição de fase de um gás de hádrons para um plasma de quarks e glúons.

e interagindo por troca de campos de gauge, os chamados glúons. Acredita-se que devido à natureza da interação forte quarks e glúons estejam permanentemente confinados dentro de hádrons. Esta é uma consequência de a interação ser desprezível a pequenas distâncias (a propriedade de liberdade assintótica) e aumentar com a distância: como consequência a energia para separar dois quarks torna-se infinita, gerando **confinamento**. Portanto os quarks são partículas livres assintoticamente (a grandes momentos ou pequenas distâncias), mas que se encontram confinados em objetos de cor neutra (os hádrons) pela interação forte. A pergunta que se coloca naturalmente é se ocorre desconfinamento uma vez que a temperatura (ou a densidade) seja suficientemente alta. Tal transição teria ocorrido de forma inversa a partir do Big Bang, quando um plasma de quarks e glúons quente e denso teria dado origem aos hádrons. Neste caso quer-se saber também qual seria a natureza desta transição: qual a ordem, etc. Estas perguntas envolvem uma grande dificuldade experimental, dada a dificuldade de obtenção de energias suficientemente altas, mas podem ser estudadas teoricamente. Em princípio a teoria de QCD das interações fortes permite a descrição das duas fases (hádrons e plasma de quarks e glúons) e da transição entre elas. A temperatura zero estudos de QCD usando teoria de perturbação fornecem resultados para situações de alta energia ou momento, mas os fenômenos de baixa energia (incluindo o confinamento) devem ser estudados de maneira **não-perturbativa** (a expansão perturbativa, baseada em acoplamentos fracos, deixa de ser válida a baixas energias). Este fato é ainda mais acentuado a altas temperaturas, pois a expansão perturbativa apresenta problemas infra-vermelhos mesmo para energias altas. Tanto a zero como a altas temperaturas, o meio de estudo não-perturbativo mais eficiente é a formulação de rede, onde o espaço-tempo é discretizado, e a teoria pode ser estudada através de simulações numéricas de Monte Carlo.

Além da transição discutida acima, o próprio limite do contínuo de uma teoria de gauge na rede é um fenômeno crítico, como veremos no Cap. ???

2.3 Comportamento de escala

2.3.1 Leis de escala

$$\begin{aligned} F_s(t,H) &= b^{-d}\, F_s(b^{y_t}t, b^{y_H}H) \\ \xi(t,H) &= b\, \xi(b^{y_t}t, b^{y_H}H) \end{aligned}$$

- $F_s(t,H) = t^{d/y_t}\Phi(H/t^{y_H/y_t}) \rightarrow \alpha,\beta,\gamma,\nu$ in terms of y_t, y_H ($y_t = 1/\nu$)
- $F_s(t,H) = H^{d/y_H}\Psi(t/H^{y_t/y_H})$ get δ ($M_{t=0} \sim H^{1/\delta}$) and universal scaling function

  $$M/H^{1/\delta} = f(t/H^{1/\Delta})$$

  where $\Delta = \beta\delta = \nu\, y_H$

2.3.2 Escala de tamanho finito

$$F_s(t,H,u_j,L) = L^{-d}Q(L^{1/\nu}t, L^{\Delta/\nu}H, L^{y_j}u_j)$$

Q is universal, u_j are irrelevant fields: $y_j < 0$

$$M = L^{-\beta/\nu}g(L^{1/\nu}t, L^{\Delta/\nu}H, L^{-\omega})$$

2.4 Cálculo de grandezas críticas em simulações

2.4.1 O parâmetro de ordem

2.4.2 Localização do ponto crítico

- Binder Cumulant: at $H = 0$

  $$U_L \equiv -\frac{1}{3}\frac{\partial^4 F_s}{\partial H^4}\chi^{-2}L^{-d} = 1 - \frac{1}{3}\frac{<M^4>}{<M>^2}$$

  has a constant (universal) value at T_c (neglecting ω)

- χ^2 Method: expand FSS form of χ around $t = 0$

  $$\chi = L^{\gamma/\nu}[c_0 + (c_1 + c_2L^{-\omega})tL^{1/\nu} + c_3L^{-\omega}]$$

2.4.3 Expoentes críticos

2.4.4 Exercícios

2.5 Extrapolação a volume infinito

2.5.1 Exercícios

2.6 Projetos

2.7 Bibliografia

- Huang, K. 1987. *Statistical Mechanics*. Wiley & Sons, New York.
- Binney, J.J.; Dowrick, N.J.; Fisher, A.J. & Newman, M.E.J. 1992. *The Theory of Critical Phenomena*. Claredon Press, Oxford.
- Stauffer, D. & Aharony, Amnon 1992. *Introduction to Percolation Theory*. Taylor & Francis, London.
- Binder, K. & Heermann, D.W. 1992. *Monte Carlo Simulation in Statistical Physics*. Springer-Verlag, Berlin.

Apêndice A

Algoritmos para identificação de aglomerados

FAZER...