Notas Incompletas de Aula para o curso de MAP214

Nestor Caticha IF-USP*

6th May 2003

1 Preliminares:

Neste curso serão abordados os seguintes tópicos: Monte Carlo. O objetivo é estudar alguns métodos de um ponto de vista teórico mas com énfase mais numa descrição física do que na prova de teoremas. Começaremos antes por uma breve exposição de idéias sobre probabilidade de um ponto de vista Bayesiano. Métodos Monte Carlo são descritos a seguir. Em primeiro lugar olhamos para alguns métodos de geração de variáveis aleatórias distribuidas de acordo a uma distribuição dada. Estes não são de aplicabilidade geral o quê nos leva a métodos de rejeição. Métodos estáticos e dinâmicos sao abordados com énfase no algoritmo de Metropolis e nos casos mais simples de distribuição gaussiana multivariada. Algumas aplicações serão discutidas: Segmentação de mercados, geração de cenários e cálculo de risco. Métodos de quasi Monte carlo serão mencionados.

2 Probabilidades, informação e o teorema de Bayes.

2.1 Teoremas de Reparametrização de Cox

Assumirei que já sabem teoria de probabilidade, mas apresentarei um enfoque ou interpretação um pouco diferente que é útil para lidar com inferência e tomadas de decisão. A lógica Aristotélica lida com situações onde há informação completa no sentido que é possível atribuir com certeza as classificações VERDADEIRA ou FALSA às diferentes asserções. No entanto em situações onde a informação é incompleta isso não é possível. Cox (1948) fez a seguinte pergunta:[1]

Qual é a forma de lidar de forma consistente com essas situações?

Cox evitou o uso de palavra probabilidade e insistiu em tentar descrever a plausibilidade de uma asserção a partir da informação (numérica ou linguística)

*nestor@if.usp.br

[1]ver o livro de E. Jaynes em http://bayes.wustl.edu : Probability Theory. Nos dois primeiros capítulos são apresentados e provados os teoremas de reparametrização de Cox. Nesta introdução serei extremamente conciso pois dar mais detalhes seria repetir material encontrado em Jaynes.

disponível. De certa forma isso é subjetivo pois depende do que *eu* sei. Por outro é objetivo, pois procuramos regras para atribuir plaussibilidade de forma que, independentemente de quem tem a informação, a atribuição será a mesma.

Para sua surpresa Cox descobriu que a atribuição de forma consistente das plausibilidades leva a uma teoria de plausibilidade que por uma reparametrização pode ser identificada exatamente com a teoria probabilidade -satisfazendo os axiomas de Kolmogorov - e ainda fornece como condição necessária para a consistencia da teoria, que o teorema de Bayes seja satisfeito. Dois comentários são necessarios. Em primeiro lugar todas as probabilidades são condicionais, i.e. há necessidade de descrever de que forma uma asserção tem atribuida uma probabilidade, qual é o contexto de informação. Em segundo lugar, neste caso consistencia tem o seguinte significado: Se há duas maneiras de calcular a probabilidade da conjunção de duas asserções, então é necessário que o resultado obtido pelos dois caminhos seja igual, se a informação disponível for a mesma. Para isto Cox necessitou postular essencialmente tres condições que a sua teoria deveria satisfazer.

1. As plausibilidades deveriam ser números reais.

2. uma condição (técnica) de monotonicidade (que é equivalente a sentido comum)

3. Consistencia.

A primeira parece óbvia, mas a estrutura matemática que aparece se os números reais forem substituidos por complexos, é a da Mecânica Quântica.

2.2 Teorema de Bayes

Considere asserções A, B, C ..., que tanto podem representar frases como ¨ *o valor da variável X está entre x e $x + dx$* ¨˙ou ¨ *hoje vai chover* ¨¸ou ainda ¨ o preço de y vai subir ¨. A notação $P(A|C)$ deve ser lido como a ¨ *a probabilidade de que A seja verdadeira dado que C é verdadeira* ¨. A conjunção de duas asserções AB, deve ser entendida como ¨ A e B ¨. As duas maneiras de atribuir probabilidades a essa conjunção-de forma consistente- leva ao teorema de Bayes. A primeira é

$$P(AB|C) = P(A|C)P(B|AC)$$

e a segunda, naturalmente é

$$P(AB|C) = P(B|C)P(A|BC),$$

temos então que

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}. \tag{1}$$

Esta fórmula é central no desenvolvimento que segue nestas notas. Os teoremas de Cox ou equivalentemente os axiomas de Kolmogorov nos dão as regras de manipulação das probabilidades, mas não fornecem os números que devemos atribuir à probabilidade de uma asserção. Esta deve ser feita com base na informação disponível. Dependendo da natureza dessa informação diferentes métodos devevem ser usados. Por exemplo se a informação é dada na forma de pares (entrada, saída) o teorema de Bayes é muito útil. Em outros casos informação adicional sobre o grupo de simetrias pode levar a uma atribuição de probabilidades adequada. Nos casos onde a informação é dada na forma de valores médios, o método de máxima entropia é a forma mais honesta de atribuir números às probabilidades das asserções. ¨ *Honesta* ¨ significa aqui que nenhuma hipótese será feita ou usada, além daquela contida na asserção C ou nos dados.

3 Infêrencia.

Suponhamos que temos informação na forma de dados sobre um experimento realizado em um sistema físico. Por algum motivo propomos um modelo funcional para descrever o sistema. A função depende de um conjunto de parâmetros λ e devemos por exemplo encontrar o valor dos parâmetros que melhor descreve os dados disponíveis. Os dados podem ser na forma de n pares ordenados $D_n = \{(x_i,\, y_i)\}_{i=1,\ldots n}$, que representam os n valores medidos y_i da variável Y, quando a medida foi feita nas condições experimentais dadas por x_i. A possibilidade de ruído η nos leva a procurar λ tal que

$$y_i = f_\lambda(x_i) + \eta \tag{2}$$

seja o melhor ajuste. Mais o que significa melhor? O teorema de Bayes nos dá a resposta: substituiremos

- a asserção A, por ¨ *O vetor de parâmetros tem o valor* λ ¨ que representaremos por ¨ λ ¨
- a asserção B por ¨$D_n = \{(x_i,\, y_i)\}_{i=1,\ldots n}$ *representa a informação na forma de dados* ¨ que representaremos por ¨ D_n ¨
- a asserção C por ¨ *condições em que foi feita a experiência* ¨

Assim o teorema de Bayes será reescrito como

$$P(\lambda|D_nC) = \frac{P(\lambda|C)P(D_n|\lambda C)}{P(D_n|C)}. \tag{3}$$

Os seguintes nomes são dados as distribuições de probabilidades que aparecem na fórmula anterior:

- $P(\lambda|C)$ distribuição de probs. *a priori* . Codifica o que sabiamos *antes* (não tem necessariamente significado temporal) de ter acesso aos dados.
- $P(D_n|\lambda C)$: a verossimilhança, é a probabilidade que atribuimos, com base no modelo, de que os dados teriam sido observados se o vetor de parâmetros tivesse efetivamente o valor λ

- $P(\lambda|D_n C)$ distribuição de probs. *POSTERIOR*, aquela que atribuimos levanod em conta tudo: os dados e a distribuição *a priori*.
- $P(D_n|C) = \int P(\lambda|C)P(D_n|\lambda C)d\lambda$ (devido à normalização da pdf) é a *evidência* (serve para decidir entre diferentes modelos)

A previsão bayesiana para o valor do vetor λ é :

$$\lambda_B = \int \lambda \, P(\lambda|D_n C) d\lambda.$$

É comum lidar com casos em que a dimensão de λ é grande. Quanto é grande? A fronteira entre grande (e impossível) e pequeno (e possível) é difusa. Em integração acima de 10 esaremos certamente na região de grande dimensionalidade. Não é possível usar técnicas numéricas tipo trapezio ou Simpson, Gauss ou seus derivados e é necessário ou fazer aproximações analíticas ou integrar numericamente usando técnicas de Monte Carlo.

4 Integração Numérica

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

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

por

$$I_T = \frac{1}{N}(\frac{1}{2}f(x_1) + \sum_{i=2}^{N-1} f(x_i) + \frac{1}{2}f(x_N)), \tag{4}$$

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. Note que esta pode ser a dimensão de um portfolio. Neste caso é necessário lidar com integrais do tipo

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

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 $10^{19}s$, 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.

5 Monte Carlo

5.1 Teorema Central do Limite.

Daremos um teorema sem enunciado e sem prova. Considere uma variável X com valores x em um intervalo dado e distribuição $P(x)$. Assumimos que os valores médios $\overline{x}$ e $\overline{x^2}$existem e são finitos.[2]A variância σ_x é definida por

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

que também é finita.

Considere ainda uma sequência de N amostragens independentes de X: $\{x_i\}_{i=1,\ldots N}$, 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

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

[2]Definimos os momentos $\overline{x^n} = \int x^n P(x)dx$

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

$$\overline{y} = \overline{x} \;\; e \;\; \sigma_y = \frac{\sigma_x}{\sqrt{N}}$$

Notem que se o objetivo for encontrar o valor esperado de x, que é $\overline{x}$, e não for possível realizar a integral, podemos estimar $\overline{x}$ a partir de y (isso pode ser generalizado para o cálculo de $\overline{f} = \int fP(x)dx$.) Qual é vantagem sobre simplesmente fazer uma medida (amostragem) de x? É que neste último caso o erro seria da ordem de σ_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 diminuido não só aumentando N mas também se mudarmos σ_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 $\sigma_x^2 = \overline{x^2} - \overline{x}^2$é finito. Dado $y = \frac{1}{N}\sum_{i=1}^{N} x_i$ mostre, a partir de

$$P(y) = \int \cdots \int dx_i \cdots dx_N \delta \left(y - \frac{1}{N}\sum_{i=1}^{N} x_i \right) \prod_{i=1}^{N} P(x_i)$$

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

Exercício:Distribuição de Cauchy Considere o problema acima, exceto que $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ 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 caclular pelo método dos resiudos.

5.2 Monte Carlo

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

$$I = \int_a^b f(x)\,dx \simeq I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{5}$$

onde os $\{x_i\}$ são escolhidos alaeatoriamente 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 σ_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. (5) 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[3].

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

$$y(x) = \int_0^x w(z)dz \tag{6}$$

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

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

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

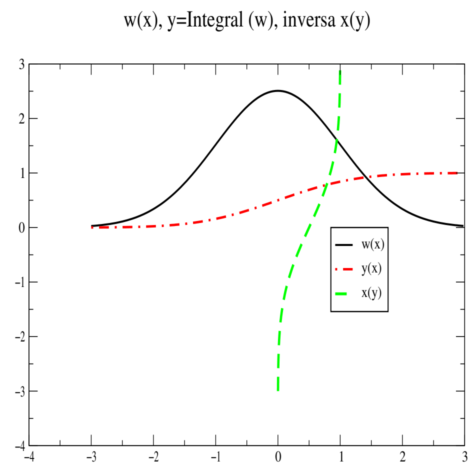
e a aproximação Monte Carlo é

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(x(y_i)) \tag{7}$$

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. [6]). A função inversa permite calcular o valor de x onde deverá ser calculada a função $f(x)$.

[3] Uma forma trivial de conseguir a redução de σ_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

Este método assume que saibamos fazer a integral da equação 6, mas não é em geral possível fazé-lo de forma analítica.



5.3 Exemplos analíticos.

Ao realizar um cálculo MC teremos, tipicamente, acesso a um gerador de números aleatórios distribuidos uniformemente em $(0,1)$. O objetivo é, aqui de forma analítica e posteriormente, de forma numérica, mostrar como gerar números aleatórios distribuidos 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

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

$$P_Y(y)dy = P_X(x)\left|\frac{\partial x}{\partial y}\right| dy \tag{8}$$

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

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

e zero fora.

5.3.1 Distribuição Exponencial

Suponha que queremos gerar amostras de uma distribuição exponencial. i.e $P_X(x) = exp(-x)$. Integrando a eq. (8) obtemos

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

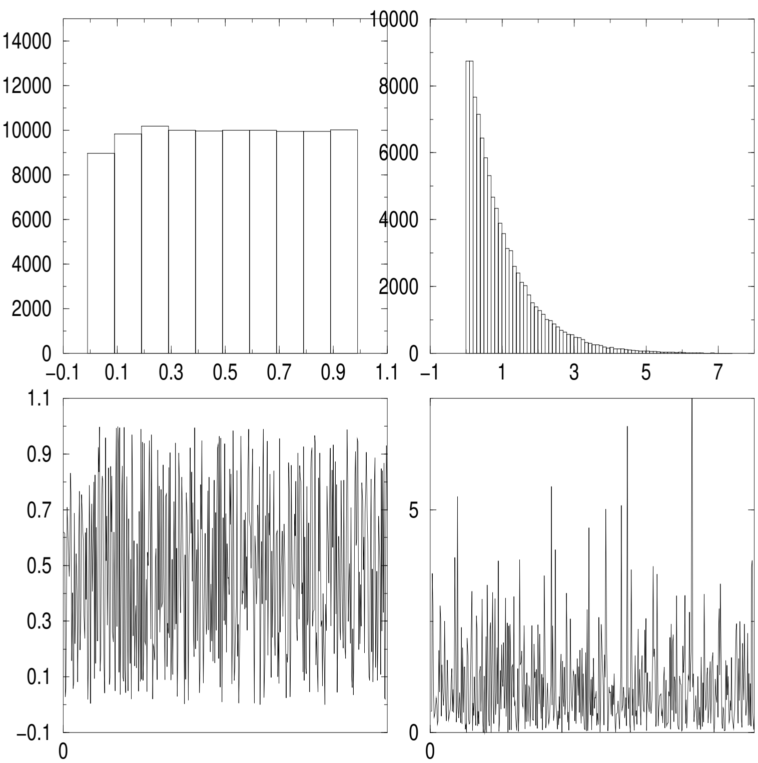
$$y(x) = \int_0^x P_X(x) dx = \int_0^x e^{-z} dz$$

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

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 distribuidos 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)



5.3.2 Distribuição Normal

Para gerar números distribuidos 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 (approximadamente). 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 (8) . 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_1e 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.(8) temos:

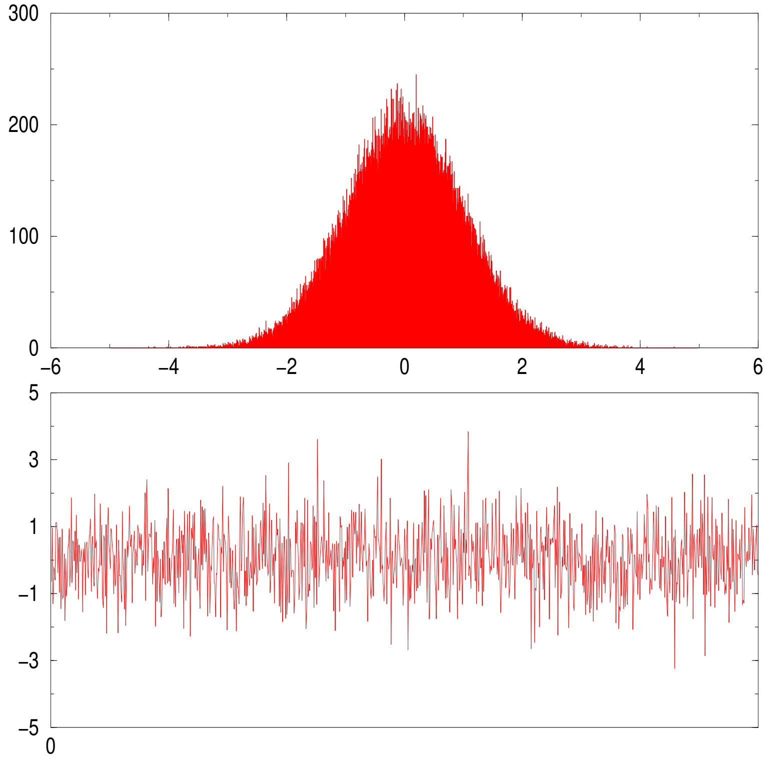
$$\int\int P_Y(y'(x_1, x_2)) \left|\frac{\partial y}{\partial x}\right| dy_1' dy_2' = \int\int P_X(x')dx_1' dx_2'$$

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.



5.4 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[4] , enquanto que na segunda classe, construiremos um processo dinâmico que usara informação anterior para gerar o próximo número.

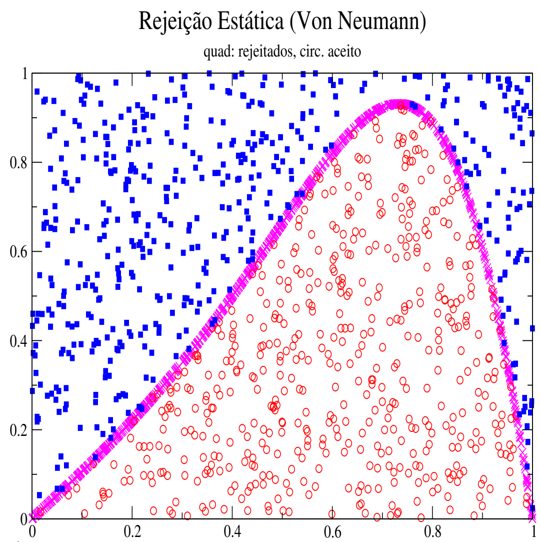
Suponhamos que a região onde $w(x) \neq 0$ está contida em (a, b) e que ela é limitada, tal que $w(x) < c$. No método de rejeição estático geramos dois NAU ξ e η e definimos

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

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

.

[4]Tão independentemente quanto o gerador de números pseudo-alealtórios o permitir.



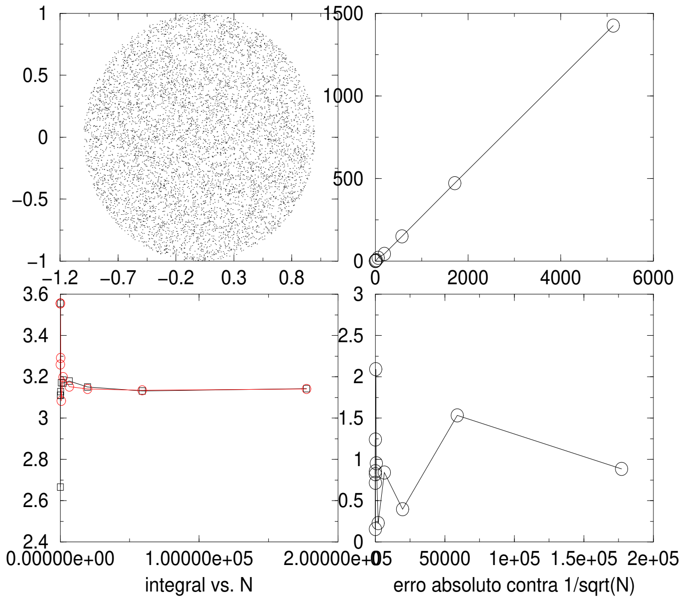
A sequência de números aceitos x são as abicissas dos circulos na figura acima.

5.5 Círculo

Exemplo: calcule π

A figura mostra os resultados de algumas simulações para estimar π. Foram gerados m pares de numeros aleatórios (x, y). Se $z = x^2 + y^2 \leq 1$ então o ponto é aceito, de outra forma é rejeitado. O resultado (figura abaixo esq. acima) mostra os pares aceitos. Continuando no sentido horário, temos os resultado para $m = 1, 3, 9....3^{12}$, respectivamente :

- número de pares rejeitados contra aceitos
- resultado de π=(numero aceito/numero total) contra m
- erro absoluto vezes $\sqrt{m}$ contra m



5.6 Métodos Dinâmicos

A idéia por trás dos processos de Monte Carlo dinâmicos é a de um processo estocástico em tempo discreto. Um processo determinístico, em oposição, é tal que dado um certo conjunto de informações, é possível -em principio- determinar a evolução futura. Um processo estocástico serve para modelar o caso em que a informação é incompleta e às várias possibilidades de evolução são atribuidas probabilidades. O objetivo é construir um processo estocástico com distribuição de equilíbrio associada igual ao $w(x)$ dado. Note que o processo estocástico é uma caminhada aleatória. Consideremos um grande número de caminhadas independentes. O processo deve ser tal que a fração das caminhadas na vizinhança de x seja proporcional a $w(x)$, pelo menos se aproxime dela assintoticamente no tempo, e chamaremos de $P(x,t)$ à distribuição no instante t.

O conceito principal para entender o processo de MC dinâmico é a probabilidade de transição, $\Gamma(x|x_n, x_{n-1}, ...x_0, ...)$, que em princípio pode depender de toda a história da evolução. Um processo é chamado Markoviano (de 1 passo) se só depende da estado atual[5]

$$\Gamma(x_{n+1}|x_n, x_{n-1}, ...x_0) = \Gamma(x_{n+1}|x_n),$$

ou de forma vaga, para onde o processo vai (o futuro), depende somente de onde está agora (o presente) e não do passado. Chamaremos a sequência $\{x_0, x_1, ..., x_n, ...\}$ de cadeia de Markov[6]. . Para os nossos objetivos estas cadeias

[5] outra notação comum é $\Gamma(x_n \to x_{n+1})$

[6] A cadeia de Markov é caracterizada pelas probabilidades de transição e pela distribuição inicial de probabilidades de x

são ferramenta suficiente.

Um ingrediente necessário que o processo deverá satisfazer é convergência para o equilíbrio. A distribuição de equilíbrio ou invariante ou estacionária deve satisfazer a condição de estacionaridade

$$w(x) = \int w(z)\Gamma(x|z)dz, \tag{9}$$

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

$$P(x,t+1) = \int P(z,t)\Gamma(x|z)dz,$$

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

$$\Delta P(x,t) = P(x,t+1) - P(x,t) = \int P(z,t)\Gamma(x|z)dz - P(x,t)\int \Gamma(z|x)dz,$$

$$\Delta P(x,t) = P(x,t+1) - P(x,t) = \int \left[P(z,t)\Gamma(x|z) - P(x,t)\Gamma(z|x)\right] dz \tag{10}$$

A interpretação é imediata, a variação da probabilidade, de um instante para o outro, tem duas contribuições, de entrada e saída. O primeiro termo $[P(z,t)\Gamma(x|z)]\,dz$ representa o número de caminhadas em um volume dz em torno de z no instante t, que fizeram a sua transição para x no instante $t+1$. O segundo termo representa a saida, isto é os que estavam em x e escapam para z. A integral leva em conta todas as contribuições do espaço. É óbvio a partir das eqs. [9, 10]

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

o que sugere uma condição

$$w(z)\Gamma(x|z) = w(x)\Gamma(z|x) \tag{11}$$

que se a matriz de probabilidade de transições satisfizer então $w(x)$ será estacionaria. Esta condição, chamada de balanceamento detalhado, não é necessária, mas só suficiente. Além de haver motivações físicas para impó-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. [11] $w(x)$ é um ponto fixo da dinâmica. Mas a pergunta que resta é sobre a estabilidade. É razoavel esperar a estabilidade dado que se em t , $P(x,t) > w(x)$, 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 t , $P(x,t) < w(x)$ então o número será menor.

Há várias maneiras de satisfazer a equação [11]. Embora todas levem a algoritmos corretos, no sentido que

$$I = \int_a^b f(x)w(x)dx \simeq I_{MC} = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{12}$$

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

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

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:

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

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:

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

τ é tempo de correlação exponencial e mede a eficiência do processo em gerar números aleatórios independentes distribuidos de acordo com $w(x)$. Agora podemos escrever

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

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

Outro tempo importante é τ_R, o tempo de relaxação para o equilíbrio. Este mede quanto tempo demora para que o processo estocástico perca memória das condições iniciais e os x sejam efetivamente reprsentativos de $w(x)$. Do ponto de vista de eficiência é razoavel não considerar e.g. os primeiros $10\tau_R$ passos gerados pelo processo. Se $C(k)$ efetivamente decair exponencialmente esses dois tempos são iguais, mas há casos em que não, e.g. perto de transições de fase críticas.

5.7 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 $x_n \to x_{n+1} = x_T$, ou seja x_T é aceito ou se não. Neste caso de rejeição fazemos a transição trivial $x_n \to x_{n+1} = x_n$, de forma que x_n é incluido 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 familia 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

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

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 d 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 β que em física é intepretado 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 medio $< E(x) >$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 aceito, mas se há uma tentativa de escolha de um lugar de energia mais alta, a tentativa não é autoamaticamente rejeitada. Se o aumentode 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.