第24章：模拟方法

- 模拟方法：亦称随机模拟方法、 Monte Carlo（蒙特卡罗）方法，是一种基于“随机数”的计算方法
  - 美国在第二次世界大战期间研制原子弹的“曼哈顿计划” 中， Monte Carlo方法被系统地应用于科学研究中，诞生了MANIAC (Mathematical Analyzer, Numerical Integrator and Computer)
  - Stanislaw Ulam, John von Neumann, Nicholas Metropolis, Enrico Fermi等人发明了一种基于样本统计的方法来解决关于在原子弹设计中的中子随机扩散问题和Schrodinger等式的特征值估计问题
  - 其基本思想很早以前就被人们所发现和利用
    - 17世纪：用事件发生的“频率”来决定事件的“概率”
    - 19世纪：用投针试验的方法来决定 π

Note on census-taking in Monte Carlo calculations, **E. Fermi and R.D. Richtmyer 1948**
The Monte Carlo Method, **Metropolis and S. Ulam 1949,**Journal of the American Statistical Association, 44, 335 (1949)

为什么叫Monte Carlo方法

- Monte Carlo：闻名世界的赌城——摩纳哥的一个小山城
- Monte Carlo方法：随机采样技术
- Monte Carlo方法与Monte Carlo赌博之间的联系？
  - 二者都是由随机变量驱动——掷骰子

Monte Carlo casino

主要内容

- 为什么需要Monte Carlo方法
- Monte Carlo 方法
  - 产生独立样本
    - 基本方法：概率积分变换（第一部分已讲）
    - 接受—拒绝（舍选）采样
    - 重要性采样
  - 产生相关样本：Markov Chain Monte Carlo
    - Metropolis-Hastings算法
    - Gibbs Sampler

Monte Carlo方法：一种通用的计算技术

- 模拟：从一个pdf产生 “典型”样本
  - $x \sim \pi(x)$
- 积分：在高维空间中积分
  - $I = \mathbb{E}\left[h(x)\right] = \int h(x)\pi(x)dx$
- 优化
  - $x^* = \arg\max_x \pi\left(x\right)$
- 学习
  - 有隐含变量的非监督学习（对后验分布模拟）
    - 对隐含变量积分
    - 有缺失数据的机器学习
  - MLE：$f(x;\theta)$

任务1：采样与模拟

- 对很多系统，其状态由概率模型控制。假设系统的状态 x，满足全局约束

$$x \in \Omega = \left\{ x : H_i\left(x\right) = h_i, i = 1, ..., K \right\}$$

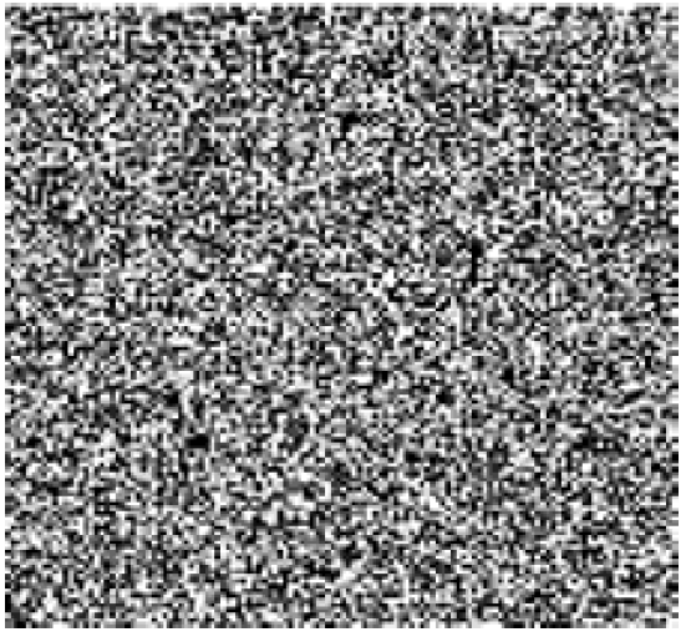
  - 这些约束可能一些逻辑约束（如8-皇后问题），也可能是一些宏观性质（如固定体积和能量的物理气体系统），也可能是统计观测（给定一些样例图像，合成有相同性质的图形）

- 任务：产生该系统的典型状态

例：噪声图像合成

- “噪声图像”模式：固定均值和方差的图像集合

$$noise = \Omega = \left\{ I_\Lambda : \lim_{\Lambda \to Z^2} \sum_{i,j \in \Lambda} I(i,j) = \mu, \lim_{\Lambda \to Z^2} \sum_{i,j \in \Lambda} \left(I(i,j) - \mu \right)^2 = \sigma^2 \right\}$$



该图像为高斯模型的一个“典型图像”

任务2：科学计算

- 科学计算中经常需要对高维空间进行积分
- Monte Carlo积分，其中 为高维空间

$$I = \int_{\Omega} h(x)\pi(x)dx = \mathbb{E}\left[h(x)\right]$$

  - 通过从分布 $\pi(x)$，产生N个样本 $X_1, ..., X_N \sim \pi(x)$
  - 根据大数定律，可以用均值去近似期望

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(X_i) \xrightarrow{p} \mathbb{E}\left[h(x)\right] = I$$

  - 该估计的标准差为

$$\widehat{se} = \frac{s}{\sqrt{N}}, \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N} h\left(X_i - \hat{I}\right)^2$$

$$\mathbb{E}\left[\bar{X}_n\right] = \mu,\ \mathbb{V}\left[\bar{X}_n\right] = \frac{\sigma^2}{N}$$

$$\mathbb{E}\left(S_n^2\right) = \sigma^2$$

Monte Carlo 积分

- 24.1例：计算 $I = \int_0^1 x^3 dx$
- 解析计算：$I = \int_0^1 x^3 dx = \frac{1}{4} x^4 \Big|_0^1 = \frac{1}{4}$
- 模拟计算：从均匀分布*Uniform*[0,1]产生*N*=10000个样本 $X_1, \ldots, X_{10000}$，则

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} h(X_i) = \frac{1}{10000} \sum_{i=1}^{10000} x_i^3 = 0.248$$

- 且标准差为0.0028。

- 注意：由于是随机采样，某次的具体结果可能与此稍有不同。
  - 参考代码见chp24_1.m

Monte Carlo 积分

- 24.2例：计算 $I=\int_{-\infty}^{x}\phi(s)ds=\Phi(x)$
- 查表可得：$I=\Phi(2)=0.9772$

- 模拟计算：
  - $I=\int_{-\infty}^{x}\phi(s)ds=\int_{-\infty}^{x}h(s)\phi(s)ds$
    - 其中
      $$h(s)=\begin{cases}1 & s<x\\ 0 & s\geq x\end{cases}$$
    - 现产生样本 $X_1,...,X_N\sim N(0,1)$
    - 得到
      $$\hat{I}=\frac{1}{N}\sum_{i=1}^{N}h(X_i)$$
    - $N=10,000,\ \hat{I}=0.9751;\qquad N=100,000,\ \hat{I}=0.9771$

参考代码见chp24_2.m

Monte Carlo 积分

- 24.3例（二项分布的贝叶斯推断）：
- 令$X \sim Binomial(n, p_1),\ Y \sim Binomial(m, p_2)$
- 要估计 $\delta = p_1 - p_2$
- 假设采用扁平先验 $f(p_1, p_2) = f(p_1) f(p_2) = 1$
- 则后验分布密度为

$$f(p_1, p_2 \mid X, Y) \propto p_1^X (1-p_1)^{n-X}\ p_2^Y (1-p_2)^{m-Y}$$

- δ 的后验均值为

$$\overline{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2) f(p_1, p_2 \mid X, Y) dp_1 dp_2$$

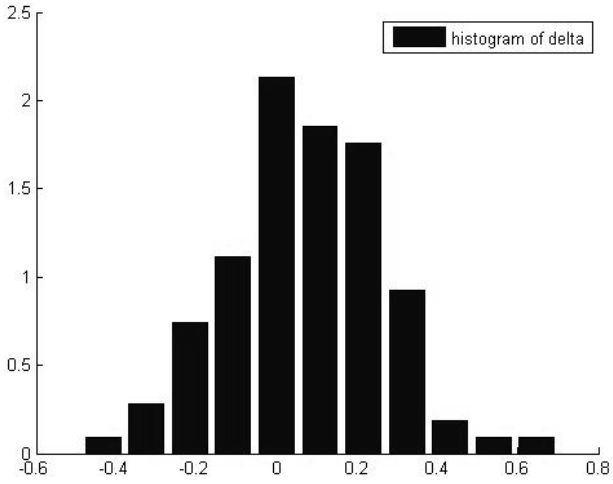
$$= \int_0^1 \int_0^1 (p_1 - p_2) p_1^X (1-p_1)^{n-X}\ p_2^Y (1-p_2)^{m-Y} d\, p_1 dp_2$$

Monte Carlo 积分

- 24.3例（续）：
- 为了避开这些积分计算，采用模拟方法
- 注意到 $f\left(p_1, p_2 \mid X, Y\right) = f\left(p_1 \mid X\right) f\left(p_2 \mid Y\right)$
- 所以 p_1, p_2 在后验分布下是独立的，且

$$p_1 \mid X \sim Beta\left(X+1, n-X+1\right),\ p_2 \mid Y \sim Beta\left(Y+1, m-Y+1\right)$$

- 因此，通过抽取
  - $P_1^{(i)} \sim Beta\left(X+1, n-Y+1\right)$
  - $P_2^{(i)} \sim Beta\left(Y+1, m-Y+1\right)$
- 得到 $\delta^{(i)} = P_1^{(i)} - P_2^{(i)}$
- 则 $\overline{\delta} = \frac{1}{N}\sum_{i=1}^{N}\delta^{(i)}$
- 当N=100时，得到估计 $\overline{\delta} = 0.07$
  - δ 的直方图如右所示
  - 参考代码见chp24_3.m



Monte Carlo 积分

- 24.4例（剂量反映的贝叶斯推断）：假设进行一个实验，给老鼠们一种药物的10种剂量中的一种，记作 $x_1 < x_2 < ... < x_{10}$ 。对每个剂量水平 x_i，用n只老鼠做实验，并且用 Y_i 表示这n只老鼠的存活数目。因此，有10个独立的二项分布 $Y_i \sim Binomial(n, p_i)$ 假设从生物学的角度知道剂量越高，死亡率越高，因此 $p_1 < p_2 < ... < p_1$
- 要估计死亡率为50%对应的剂量LD50，即

$$\delta = x_j,\ j = \min\{i : p_i \geq 0.50\}$$

- 注意到 δ 是 $p_1,..., p_{10}$ 的一个复杂的隐函数，记作 $\delta = g(p_1,..., p_{10})$
- 这意味着若已知 $(p_1,..., p_{10})$ ，则可以求出 δ 。 δ 的后验均值为

$$\overline{\delta} = \int\int ...\int_A g(p_1,..., p_{10})\, f(p_1,..., p_{10} \mid Y_1,..., Y_{10})\, dp_1 dp_2 ... dp_{10}$$

- 积分区域为 $A = \{(p_1,..., p_{10}) : p_1 < p_2 < ... < p_{10}\}$
- δ 的后验CDF为 $F(c \mid Y_1,..., Y_{10}) = \mathbb{P}(\delta \leq c \mid Y_1,..., Y_{10})$

$$= \int\int ...\int_B f(p_1,..., p_{10} \mid Y_1,..., Y_{10})\, dp_1 dp_2 ... dp_{10}$$

- 其中 $B = A \cap \{(p_1,..., p_{10}) : g(p_1,..., p_{10}) \leq c\}$

Monte Carlo 积分

- 24.4例（续）：上述10维积分可以用模拟方法近似
- 在A上取扁平先验，则除了截断，每个 是一个Beta分布：

$$P_i \mid X \sim Beta\left(Y_i + 1, n - Y_i + 1\right)$$

- 为了从后验分布中抽取样本，采取以下步骤：
  - （1）抽取 $P_i \sim Beta\left(Y_i + 1, n - Y_i + 1\right),\ i = 1,...,10$
  - （2）若 $P_1 < P_2 < ... < P_{10}$ ，继续；否则丢掉这些数据并再次抽取，直到得到一个可以保留的样本
  - （3）令 $\delta = x_j,\ j = \min\left\{i : p_i \geq 0.50\right\}$

- 重复N次，得到 $\delta^{(1)},...,\delta^{(N)}$ ，则

$$\mathbb{E}\left(\delta \mid Y_1,...,Y_{10}\right) \approx \frac{1}{N}\sum_{i=1}^{N} \delta^{(i)}$$

- 为离散变量，可以估计其概率密度函数

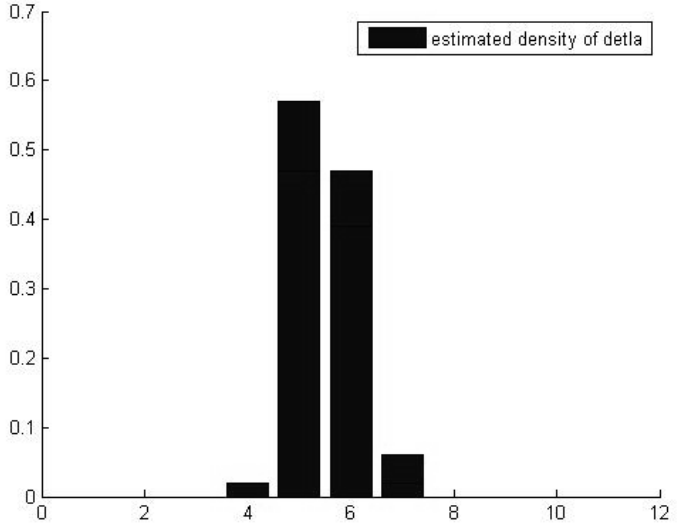
$$\mathbb{P}\left(\delta = x_j \mid Y_1,...,Y_{10}\right) \approx \frac{1}{N}\sum_{i=1}^{N} I\left(\delta^{(i)} = j\right)$$

Monte Carlo 积分

- 24.4例（续）：
- 输入：
  - 剂量：[1, 2, 3, 4, 5, 6, 7, 8, 9, 10];
  - 动物数n：[15, 15, 15, 15, 15, 15, 15, 15, 15, 15]
  - 死亡动物数Y：[0, 0, 2, 2, 8, 10, 12, 14, 15, 14]
- N=100次，得到

$$\overline{\delta} \approx \frac{1}{N}\sum_{i=1}^{N}\delta^{(i)} = 5.59$$

- 概率密度函数如右所示
  - 0 0 0 0 0.4700 0.4700 0.0600 0…
  - 参考代码见chp24_4.m



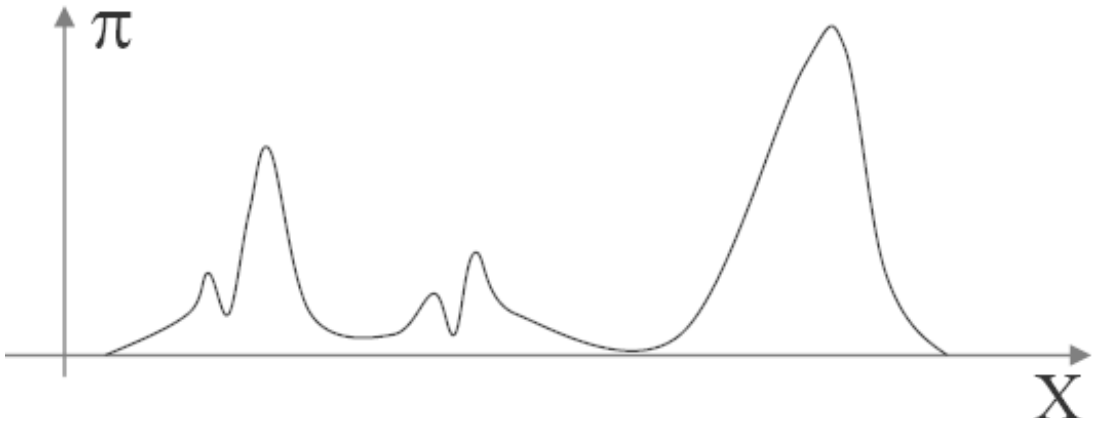
任务3：优化和Bayesian推断

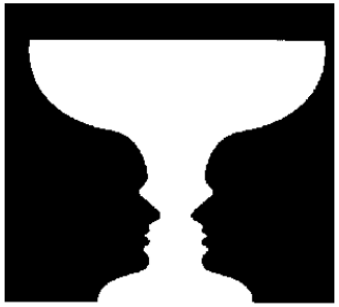
- 例：给定图像I，求图像的语义解释X

$$X^* = \arg\max_X \pi(X \mid I)$$

- 我们需要对后验进行采样，并保留多个解

$$(X_1, ..., X_k) \sim \pi(X \mid I)$$





Vase vs. face

优化

- 例：将一个求最小值的问题转化为求能量最小问题
  - 求函数 $h(x)$的最小值
  - 等价于求 $\exp\{-h(x)/T\}$
  - 可以通过对分布 $\pi(x) \propto \exp\{-h(x)/T\}$ 采样实现
  - 当T越小时，上述分布在函数h的全局最小值附近的概率越大→ 当$T \to 0$时，从上述分布的样本几乎会在h的全局最小值附近
- 这就是模拟退火的基本思想

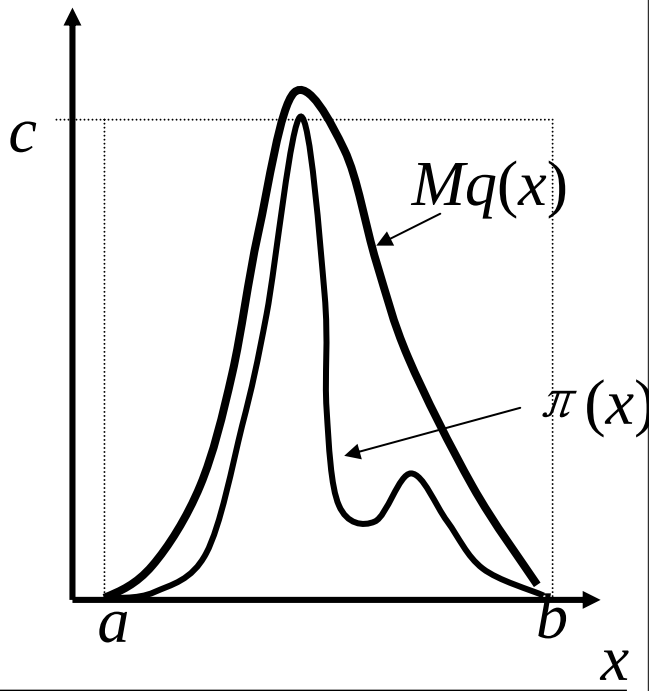
问题：怎样采样

- 直接采样：对简单模型可用

- 接受—拒绝采样（ Acceptance-Rejection sampling ）
- 重要性采样（ Importance Sampling ）
- Markov Chain Monte Carlo

接受—拒绝采样（Acceptance-Rejection Sampling

- 给定目标分布密度 $\pi(x)$
- 建议密度（Proposal Density） $q(x)$和常数M，使得
  - 对$q(x)$采样比较容易
  - $q(x)$的形状接近 $\pi(x)$, 且有
    $$\pi(x) \le Mq(x), \forall x$$
- 通过对$q(x)$的采样，
- 实现对 $\pi(x)$的采样

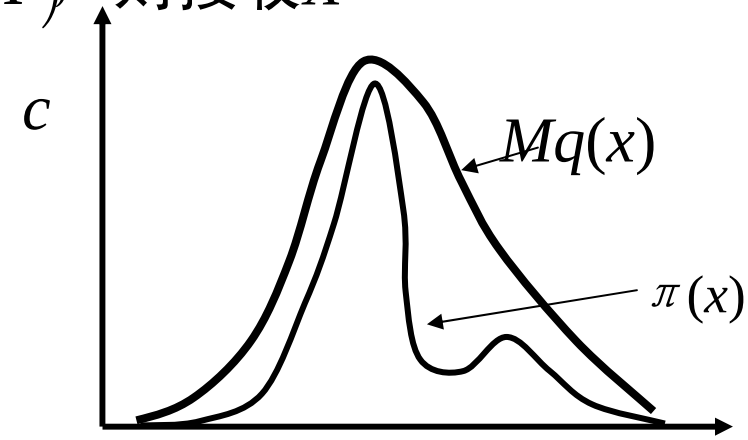


J. von Neumann, "Various techniques used in connection with random digits. Monte Carlo methods", Nat. Bureau Standards, 12 (1951), pp. 36–38.

接受—拒绝采样

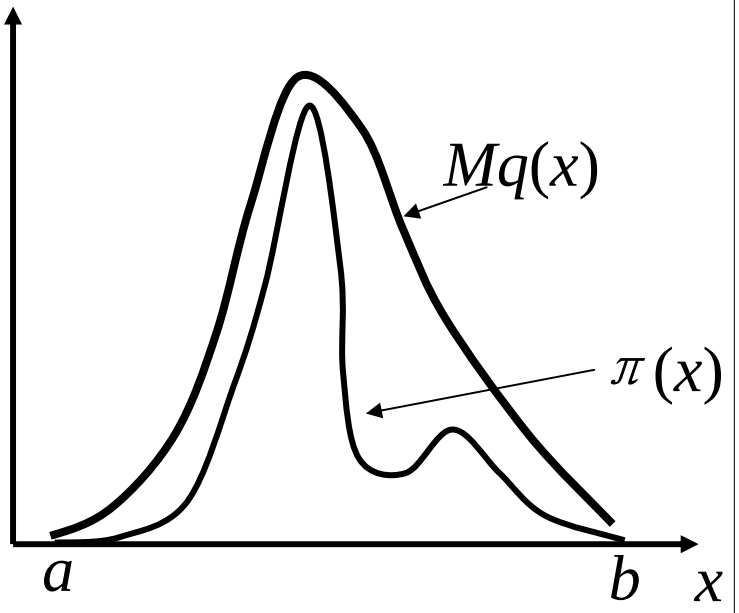
- 采样过程为：
  - 1、产生样本 $X \sim q(x)$，和 $U \sim Uniform[0,1]$
  - 2、若 $U \le \pi(X)/Mq(X)$，则接收X
- 则接收的样本服从分布 $\pi(x)$
- 等价于
  - 1、产生样本 $X \sim q(x)$，和 $U \sim Uniform[0,1]$
  - 2、$Y = Mq(X)U$，若 $Y \le \pi(X)$，则接收X

包络原则



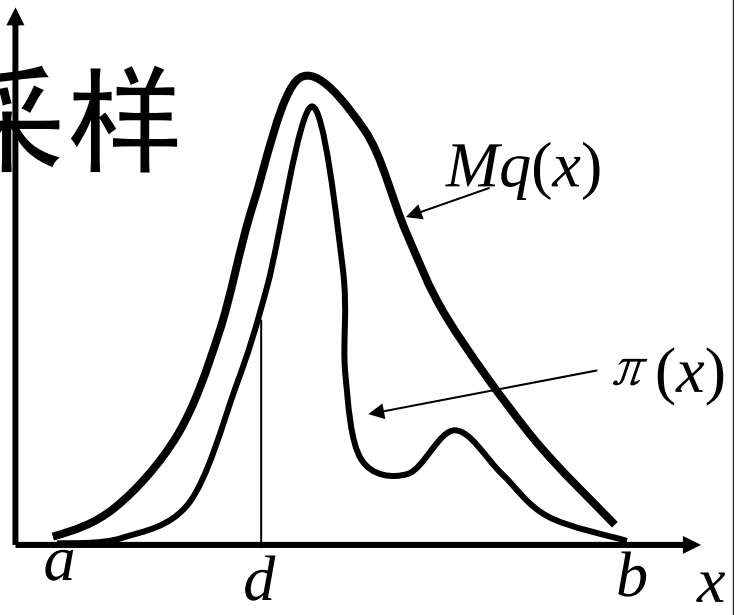
接受—拒绝采样

- 几何解释：
  - 在二维图上，随机选取位于曲线$Mq(x)$下的点$[x,y]$；
  - 选取位于曲线 $\pi(x)$下的那些点，则这些点将服从概率密度为 $\pi(x)$的分布



- 对 $X \sim \pi(x)$ 的模拟，等价于对 $(X,U) \sim Uniform[X,U], 0 \le U \le \pi(x)$
- 的模拟。

接受—拒绝采样



- 证明：
  - x和y的概率密度函数分别为

$$q_s(x) = q(x), \quad q_s(y) = \frac{1}{Mq(x)}$$

  - 联合概率密度函数为

$$q(x,y) = q_s(x) \cdot q_s(y) = \frac{q(x)}{Mq(x)} = \frac{1}{M}$$

  - 按接受—拒绝采样法抽出的随机数d的概率：

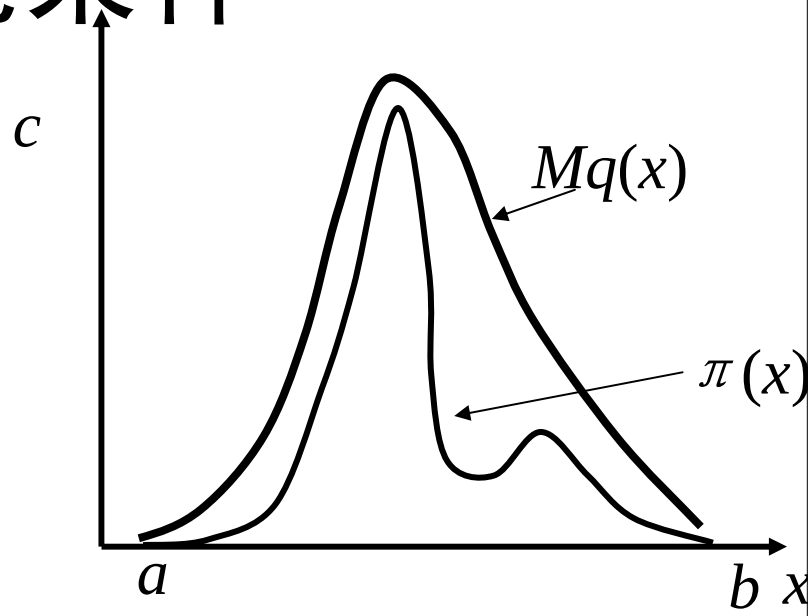
$$\mathbb{P}(X \le d \,|\, Y \le \pi(x)) = \frac{\int_a^d \int_0^{\pi(x)} q(x,y)\,dxdy}{\int_a^b \int_0^{\pi(x)} q(x,y)\,dxdy} = \int_a^d \pi(x)dx = F_X(d)$$

即d的概率函数为 $\pi(x)$

接受—拒绝采样

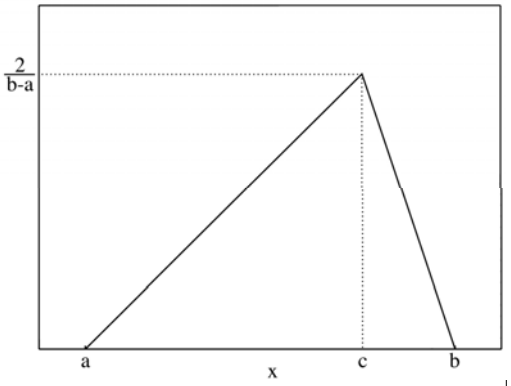
- 接受率，即采样的效率为

$$E \propto \frac{\int_a^b \pi(x)dx}{M\int_a^b q(x)dx} \propto \frac{1}{M}$$



  - 等于$q(x)$下的面积除以 $\pi(x)$下的面积
  - M的大小决定了采样的效率，q与 π相似但尾部更厚
    - 如果q、 π为概率密度函数，则 $M \geq 1$ ，$M = \max\{\pi(x)/q(x)\}$
      ➔ 当q与 π形状相似时，M越小，接受率更高，采样更有效
  - 对高维问题， M可能很大，此时拒绝率→100%

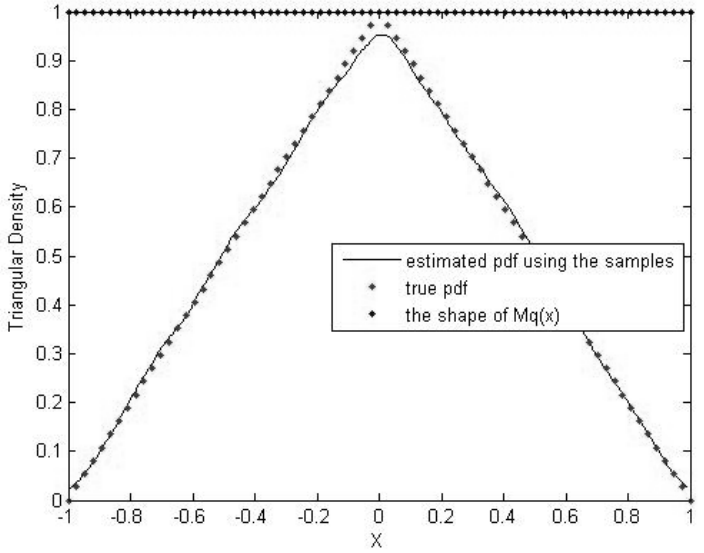
例：对三角分布采样



- 例1：对三角分布采样，取a=-1, b=1, 和c=0

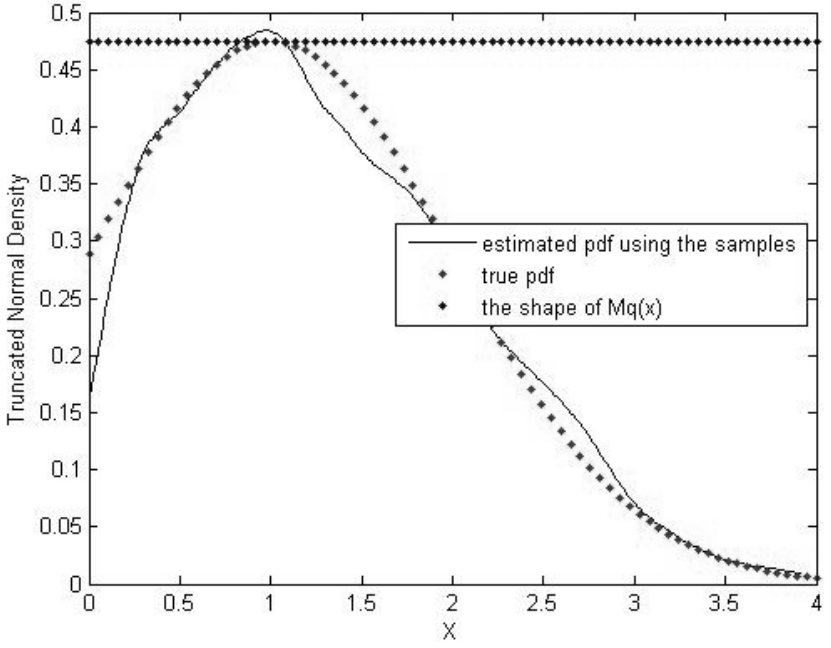
$$\pi(x\,|\,a,b,c)=\begin{cases}\dfrac{2(x-a)}{(b-a)(c-a)} & for\ a\le x\le c\\ \dfrac{2(b-x)}{(b-a)(b-c)} & for\ c\le x\le b\\ 0 & otherwise\end{cases}=\begin{cases}(x+1) & for\ -1\le x\le 0\\ (1-x) & for\ 0\le x\le 1\\ 0 & otherwise\end{cases}=\begin{cases}1-|x| & -1\le x\le 1\\ 0 & otherwise\end{cases}$$

- 其中a,b,c分别为下界、上界和众数
- 建议密度$q(x)\sim Uniform[-1,1]$
- $M=2$
- 参考代码见*acceptreject_tri.m*



例：对截断正态分布采样

- 例2：对截断正态分布采样：$\pi(x) \propto N(1,1) I(0 \le x \le 4)$
- 建议密度 $q(x) \sim Uniform[0,4]$
- $$M = \frac{4\phi(1-1)}{\Phi(4-1)-\Phi(0-1)} = \frac{4\phi(0)}{\Phi(3)-\Phi(-1)}$$
- 参考代码见
  - *acceptreject_trunnorm.m*



例：对截断正态分布采样
——不同建议分布的影响

- 例3：对截断正态分布采样 $\pi(x) \propto \phi(x) I(x > c)$
- 当c<0时，建议密度可用标准正态分布 $X \sim q(x) = \phi(x)$
  - 只接受X>c的样本
  - 最坏的情况，接受率为50%

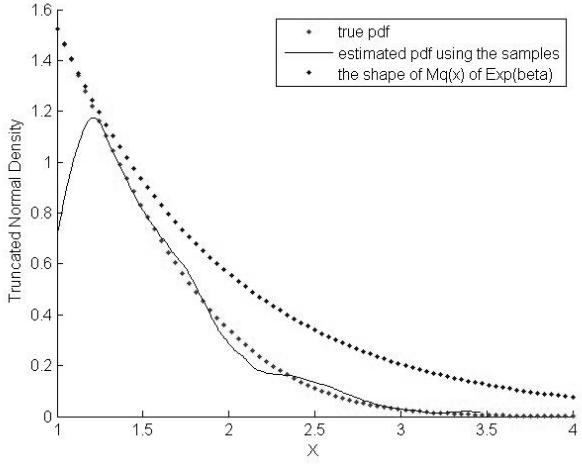
- 当c>0时，由于c较大时，上述策略效率很低（绝大多数样本被拒绝）
  - 当c=1时，接受率为 $\Phi(c)$
- 此时可采用指数分布作为包络函数
- 我们要计算最小的常数M，使得 $X \sim q(x) = \beta e^{-\beta x}$

$$\frac{\phi(x)}{1-\Phi(c)} \le M\beta e^{-\beta x}, \quad \forall x \ge 0$$

- 则最佳的M为

$$M = \frac{\exp(\beta^2/2)}{\sqrt{2\pi}\beta(1-\Phi(c))}$$

- 为了得到最小的拒绝率，我们还选择最佳的$\beta = 1$
  - 接受率为 $\frac{1}{M} = \frac{\sqrt{2\pi}(1-\Phi(c))}{\exp(1/2)}$
- 参考代码见*acceptreject_trunnorm_diff_proposals.m*

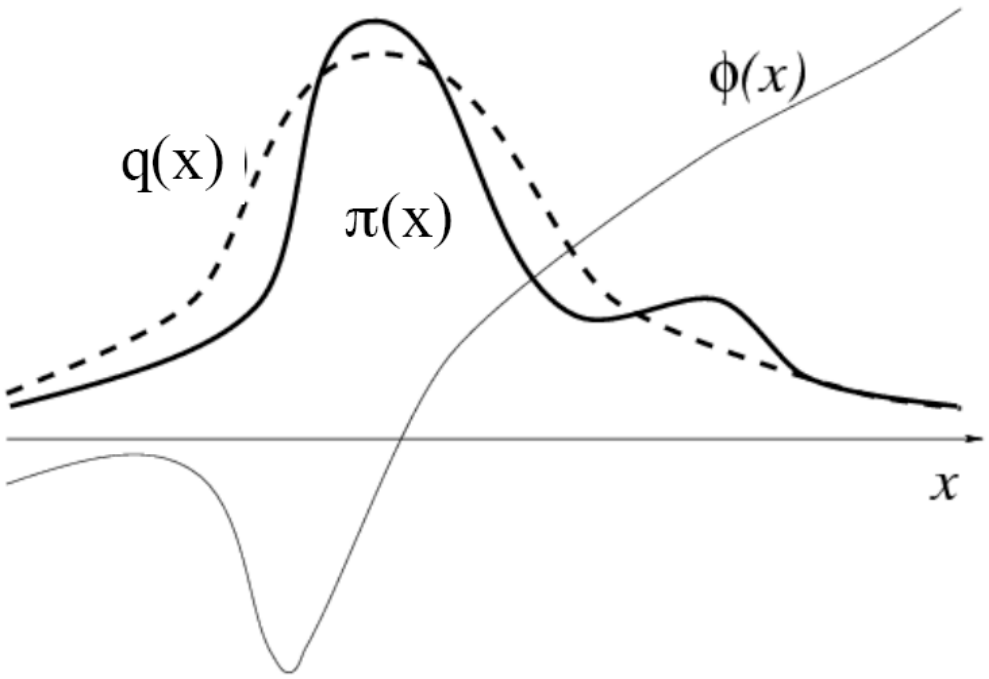


重要性采样（Importance Sampling）

- 通过从已知采样的概率 $q(x)$ 采样，近似积分

$$I = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx$$

重要性： $w(x) = \frac{\pi(x)}{q(x)}$



重要性采样

- 通过
  - 1、产生样本 $X_1,...,X_N \sim q(x)$
  - 2、计算

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} f\left(X_i\right) w\left(X_i\right), \text{ where } w\left(X_i\right) = \frac{\pi\left(X_i\right)}{q\left(X_i\right)}$$

- 该估计是一致估计：

$$\frac{1}{N}\sum_{i=1}^{N} f\left(X_i\right) w\left(X_i\right) \xrightarrow{P} \int f(x)\pi\left(x\right)dx$$

重要性采样

- $q(x)$的选择：
  - 容易采样的任意分布，如高斯分布

- 但某些分布更好一些，估计的方差更小
  - 只有下式成立，估计的方差才不为无限大

$$\mathbb{E}_q\left(\left(f(X)\frac{\pi(X)}{q(X)}\right)^2\right)=\int f(x)\pi(x)\frac{f(x)\pi(x)}{q(x)}dx<\infty$$

  - 所以$q(x)$的形状应和 $f(x)\pi(x)$的形状足够近似，且$q(x)$的尾部应比 $\pi(x)$的尾部厚
  - 当$\sup\left(\pi(x)/q(x)\right)=\infty$，权重$w(X_i)=\pi(X_i)/q(X_i)$ 变化范围很宽，只对少数样本点给与大的权重，其余样本的权重趋近于0
  - 当$\sup\left(\pi(x)/q(x)\right)<\infty$，也可以用接受—拒绝采样法直接采样

重要性采样

- 可以用下列有偏估计代替：

$$\hat{I} = \frac{\sum_{i=1}^{N} f(X_i) w(X_i)}{\sum_{i=1}^{N} w(X_i)}$$

  - 还是能收敛到 $\int f(x)\pi(x)dx$
  - 偏差很小，通常也能减小方差，从而MSE可能更小

重要性采样

- 例24.6（尾概率）：估计$I=\mathbb{P}\left(Z>3\right)=0.0013,\ Z\sim N\left(0,1\right)$
- 记 $I=\int f\left(x\right)\pi\left(x\right)dx$
- 其中 $\pi\left(x\right)=\phi\left(x\right),\ f\left(x\right)=\begin{cases}1 & x>3\\ 0 & otherwise\end{cases}$
- 方法1：基本Monte Carlo方法
  - $\hat{I}=\frac{1}{N}\sum_{i=1}^{N}f\left(X_i\right),\ X_i\sim N\left(0,1\right)$
  - 当n=100时，$\mathbb{E}\left(\hat{I}\right)=0.0015,\ \mathbb{V}\left(\hat{I}\right)=0.0039$
- 方法2：重要性采样
  - 建议分布q为标准正态分布：$\hat{I}=\frac{1}{N}\sum_{i=1}^{N}\frac{f\left(X_i\right)\pi\left(X_i\right)}{q\left(X_i\right)},\ X_i\sim N\left(0,1\right)$
  - 当n=100时，$\mathbb{E}\left(\hat{I}\right)=0.0011,\ \mathbb{V}\left(\hat{I}\right)=0.0002$
  - 方差降低了20倍
- 方法3：接受—拒绝算法（截断高斯）
- 参考代码见*importance.m*

MCMC：Markov Chain Monte Carlo

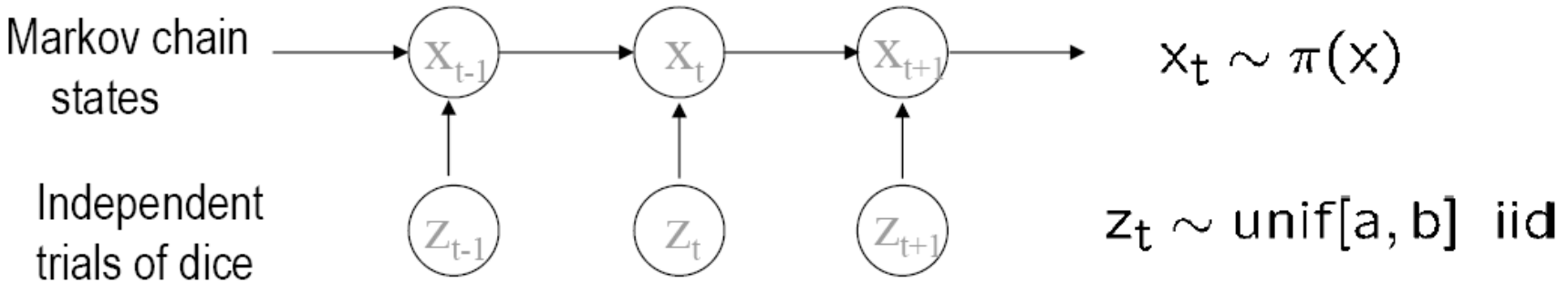
- 重要性采样和接受—拒绝采样都只$q(x)$与 $\pi(x)$很相近似时才表现很好

- 在高维空间问题中，标准的采样方法会失败：
  - 接受—拒绝采样：维数增高时，拒绝率→100%
  - 重要性采样：大多数的样本权重→ 0

- 对高维复杂问题，用马尔科夫链（Markov Chain）产生一些列相关样本，实现对分布的采样

MCMC

- MCMC：一种利用一定范围内的均匀分布的随机数，对高维空间概率进行采样的通用技术。
- 基本思想：设计一个马尔科夫链，使得其稳定概率为目标分布 $\pi(x)$



马尔科夫链

- 马尔科夫链：一种表示随机系统的数学模型，其状态变化用转换概率表示
- 马尔可夫链的当前状态至于过去最近的一些状态有关
  - 如天气变化、股票价格…
  - 如一阶马尔可夫链，只考虑最近一个状态的影响

$$\mathbb{P}(X_n = i \mid X_{n-1}, \cdots, X_0) = \mathbb{P}(X_n = x \mid X_{n-1})$$

- 条件概率 $P_{ij}\left(m, m+n\right) = \mathbb{P}(X_{m+n} = i \mid X_m = j)$ 为马氏链在时刻m处于状态i条件下，在时刻$m+n$转移到状态j的**转移概率**
- 齐次马尔可夫链：转移概率与所处的时刻n无关：

$$p_{ij} \equiv \mathbb{P}(X_n = i \mid X_{n-1} = j), \quad p_{ij}\left(m\right) \equiv \mathbb{P}(X_{n+m} = i \mid X_n = j)$$

  - 以下我们讨论的都是齐次马尔可夫链

马尔科夫链

- 概率转移矩阵：如果马尔可夫链的状态空间Ω是离散的，则用一个矩阵P表示从一个状态转移到到下一个状态的概率
  - 亦被称为转移核

    $$P = \left[p_{ij} \right],\ p_{ij} \equiv \mathbb{P}(X_n = i \mid X_{n-1} = j)$$

  - 性质：

    $$\sum_j P_{ij} = 1$$

- 令初始分布 $X_0 \sim \mu_0$
- 则 $X_1 \sim \mu_1 = \mu_0 P,\ ...,\ X_n \sim \mu_n = \mu_{n-1} P = \mu_0 P^n$
- 若 $\pi = \pi P$，则 π称为马尔可夫链的稳定状态。

马尔科夫链中的一些术语

- 可达： 经过一系列的转变，状态i可以到达状态j
  - $\exists n,\ p_{ij}(n)>0$ ，记作 $i \to j$
- 互通：$i \to j$ 且 $j \to i$
- 不可约链：所有状态之间是互通的
- 吸收态：只能出去不能进来