

机器学习

Machine Learning

北京航空航天大学计算机学院智能识别与图像处理实验室
IRIP Lab, School of Computer Science and Engineering, Beihang University

黄 迪 刘庆杰

2020年秋季学期
Fall 2020

课前回顾

概述

● Vladimir N. Vapnik

1936年 出生于苏联

1958年 乌兹别克国立大学 硕士

1964年 莫斯科控制科学学院 博士

1964-1990年 莫斯科控制科学学院
曾担任计算机科学与研究系主任

1991-2001年 美国AT&T贝尔实验室
发明支持向量机理论

2002-2014年 NEC实验室(美国)
从事机器学习研究

2014年至今 美国Facebook公司
从事人工智能研究

1995年和**2003年**，他分别被伦敦大学皇家霍洛威学院和美国哥伦比亚大学聘为计算机专业的教授。**2006年**，他成为美国国家工程院院士。



概述

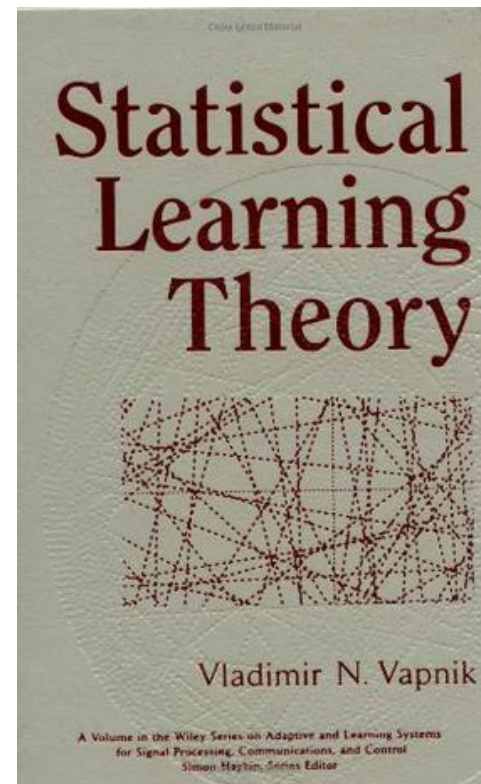
● V. Vapnik对于统计机器学习的贡献

1968年，Vapnik和Chervonenkis提出了VC熵和VC维的概念，这些是统计学习理论的核心概念。同时，他们发现了泛函空间的大数定理，得到了关于收敛速度的非渐进界的主要结论。

1974年，Vapnik和Chervonenkis提出了结构风险最小化归纳原则。

1989年，Vapnik和Chervonenkis发现了经验风险最小化归纳原则和最大似然方法一致性的充分必要条件，完成了对经验风险最小化归纳推理的分析。

90年代中期，有限样本情况下的机器学习理论研究逐渐成熟起来，形成了较完善的理论体系——统计学习理论。

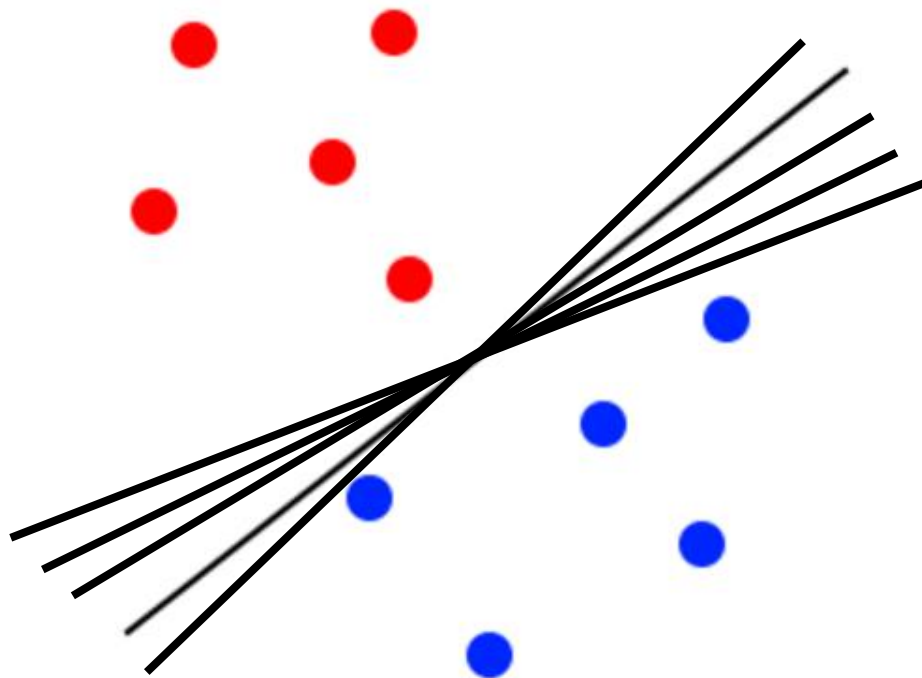


线性分类模型

- 两类样本的线性分类问题

$$y(x) = w^T x + b$$

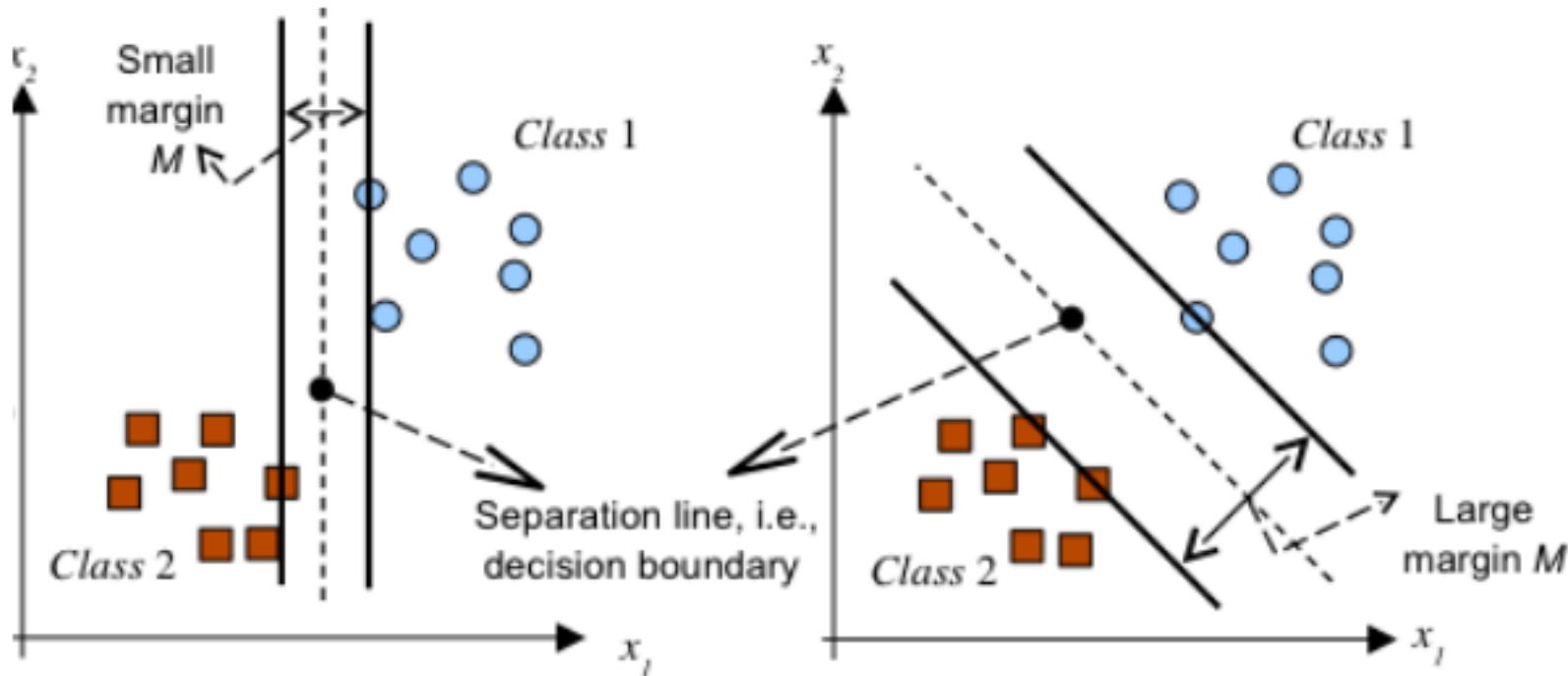
$$y(x, w) = f\left(\sum_{j=1}^M w_j x_j\right)$$



支持向量机

- SVM从线性可分情况下的**最优分类面**发展而来。

最优分类面就是要求分类线**不但能将两类正确分开**(训练错误率为0), 且使**分类间隔**最大。SVM考虑寻找一个满足分类要求的超平面, 并且**使训练集中的点距离分类面尽可能的远**, 也就是寻找一个分类面**使它两侧的空白区域(Margin)最大**。



支持向量机

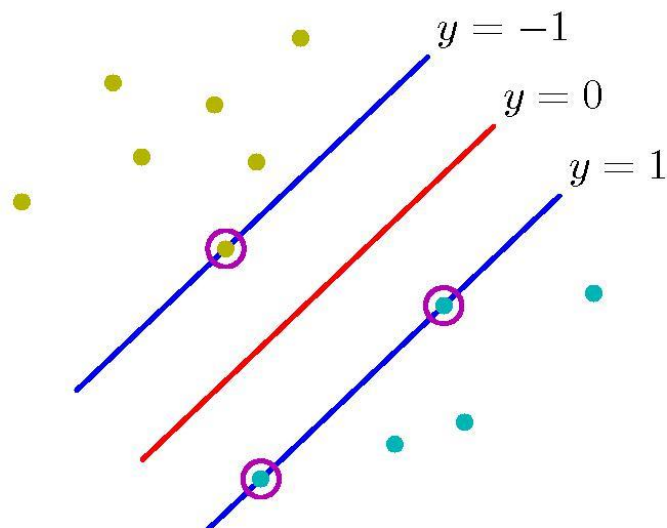
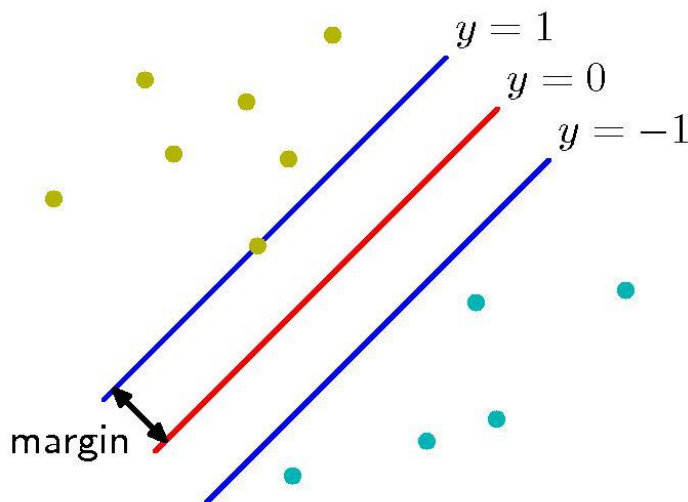
● 线性支持向量机

样本集 $\{x_n, t_n\}, n = 1, 2, \dots, N, x_n \in \mathcal{R}^d; t_n \in \{-1, 1\}$

分类器 $y(x) = w^T x + b$

$$t_n = \begin{cases} 1, y(x_n) > 0 & \text{if } x_i \in w_1 \\ -1, y(x_n) < 0 & \text{if } x_i \in w_2 \end{cases}$$

→ $t_n y(x_n) > 0$



支持向量机

● 线性支持向量机

样本集任意一点 x_n 到分类面(满足 $t_n y(x_n) > 0$)的距离

$$\frac{t_n y(x_n)}{\|w\|} = \frac{t_n (w^T x_n + b)}{\|w\|}$$

优化 w 和 b 使 Margin 最大

$$\arg \max_{w, b} \left\{ \frac{1}{\|w\|} \min_n [t_n (w^T x_n + b)] \right\}$$

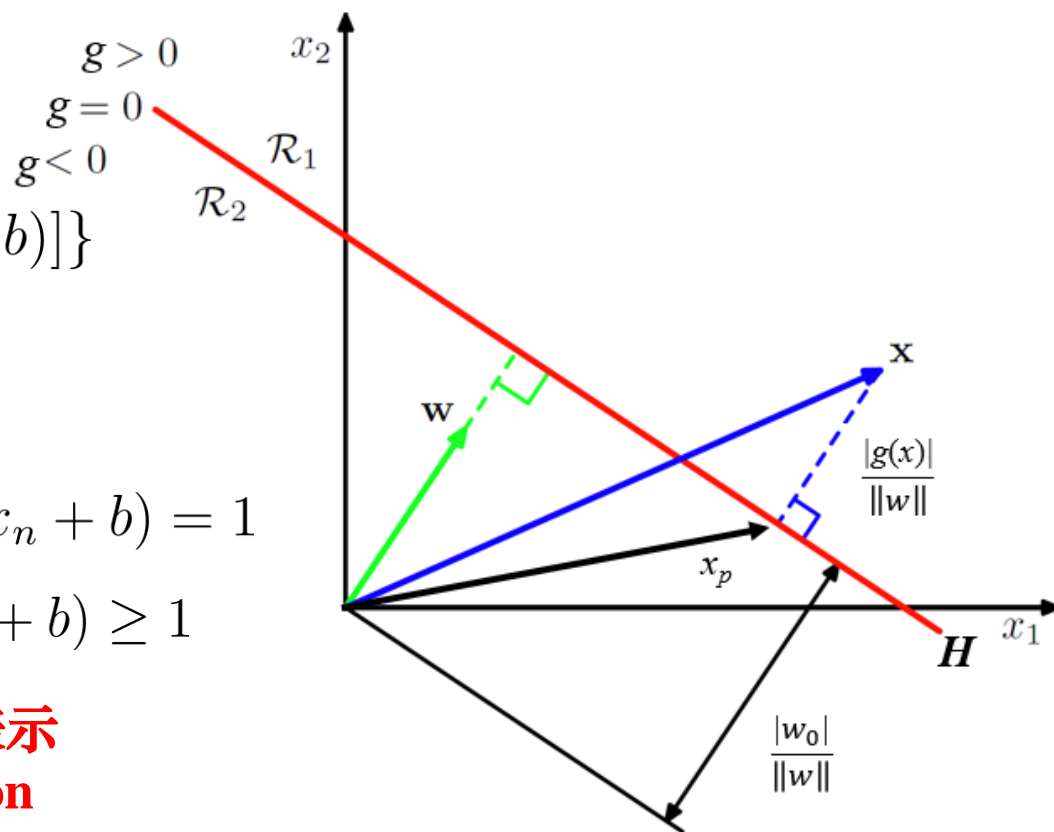
求解复杂

$$w \rightarrow kw, b \rightarrow kb$$

对于离超平面最近的点 $t_n (w^T x_n + b) = 1$

那么对于所有点满足 $t_n (w^T x_n + b) \geq 1$

对于决策超平面的标准表示
Canonical Representation



支持向量机

● 线性支持向量机

问题转化为最大化 $\|w\|^{-1}$, 等价于 $\arg \min_{w,b} \frac{1}{2} \|w\|^2$

二次规划问题

$$s.t. \ t_n(w^T x_n + b) \geq 1$$

拉格朗日乘子法 $L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{n=1}^N a_n \{t_n(w^T x_n + b) - 1\}, a_n \geq 0$

分别对变量求导 $\frac{\partial L(w,b,a)}{\partial w} = w - \sum_{n=1}^N a_n t_n x_n = 0$

$$\frac{\partial L(w,b,a)}{\partial b} = \sum_{n=1}^N a_n t_n = 0$$

代入 L 得到对偶形式:

$$\tilde{L}(a) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m x_n^T x_m$$

二次规划问题

$$w.r.t. \ a_n \geq 0, n = 1, \dots, N, \sum_{n=1}^N a_n t_n = 0$$

KKT条件

● KKT(Karush-Kuhn-Tucker)条件

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, n$$

$$\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, l$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \dots, k$$

KKT对偶互补条件

$$g_i(w^*) \leq 0, \quad i = 1, \dots, k$$

$$\alpha^* \geq 0, \quad i = 1, \dots, k$$

如果 w^* , α^* , β^* 满足KKT条件, 那么它们就是原问题和对偶问题的解。

补充条件隐含如果 $\alpha^* > 0$, 那么 $g_i(w^*) = 0$, 即 w 处于可行域的边界上, 是起作用的(Active)约束, 而位于可行域内部的点都是不起作用的约束, 其 $\alpha^* = 0$ 。

支持向量机

● 线性支持向量机

KKT条件:

$$a_n \geq 0$$

$$t_n y(x_n) - 1 \geq 0$$

$$a_n \{t_n y(x_n) - 1\} = 0$$

支持向量:

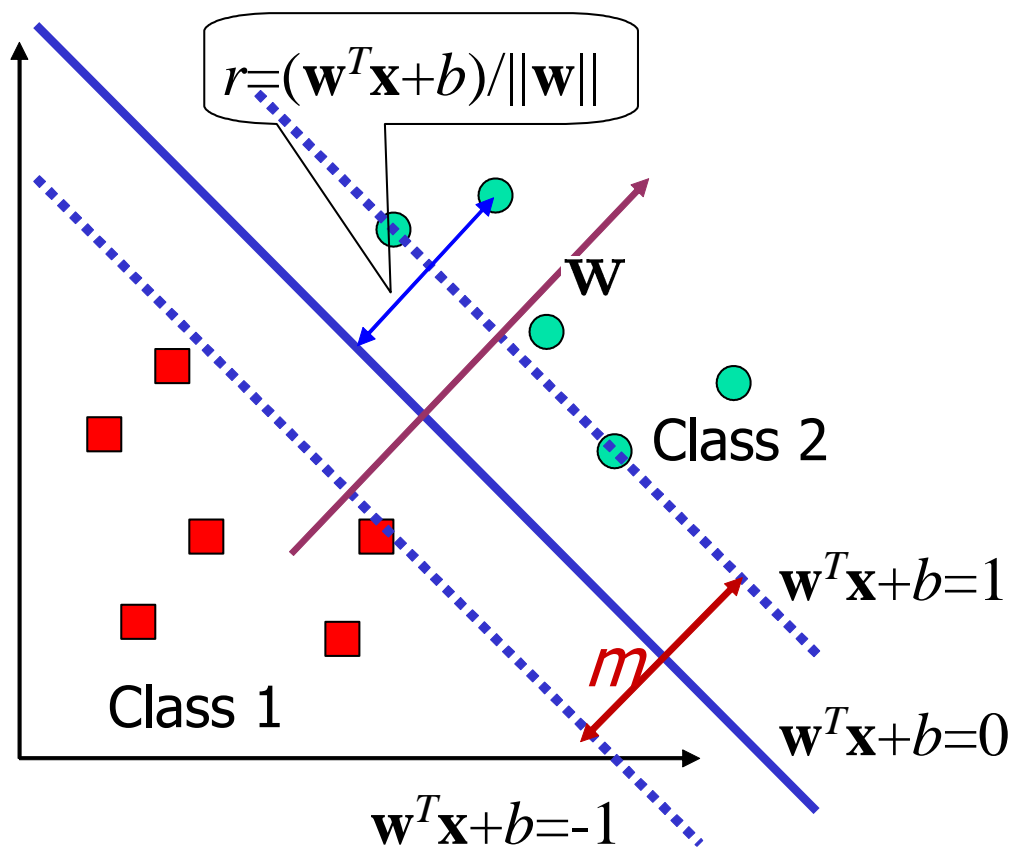
$$t_n (w^T x + b) = 1, a_n > 0$$

非支持向量:

$$t_n (w^T x + b) > 1, a_n = 0$$

$$y(x) = \sum_{n=1}^N a_n t_n x^T x_n + b$$

$$b = \frac{1}{N_S} \sum_{n \in S} (t_n - \sum_{m \in S} a_m t_m x_n^T x_m)$$



超平面法向量是支持向量的线性组合

支持向量机

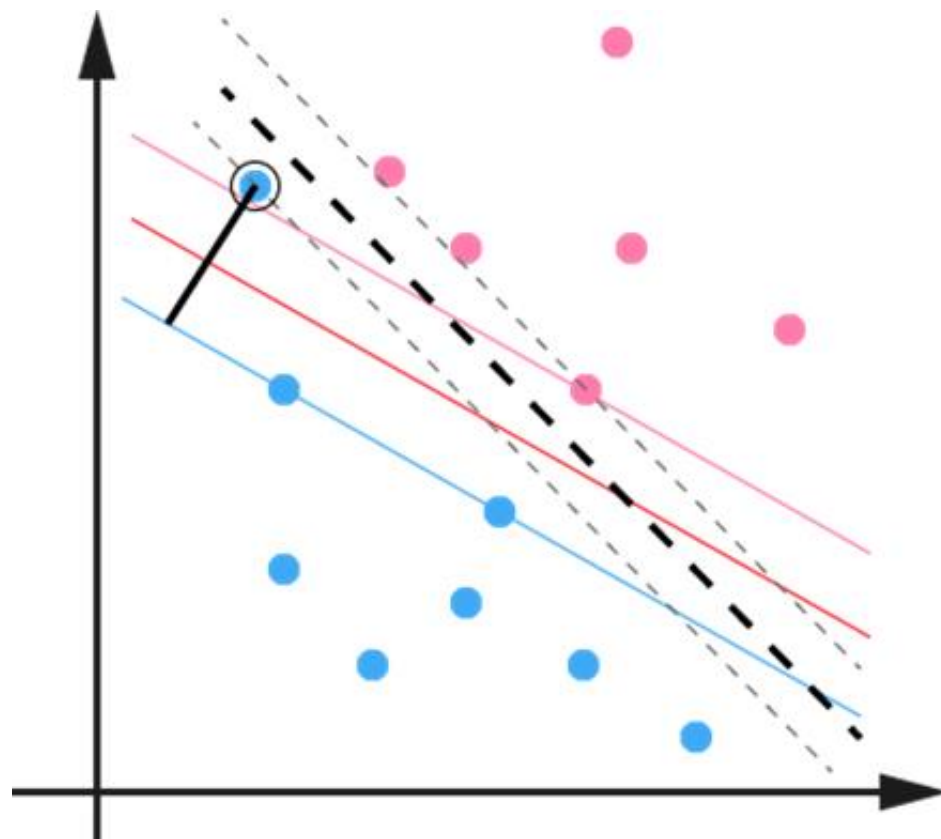
- 处理噪声和离群点

求解最优分类面的时间代价大还可能导致泛化性能差。因此，对于分布有交集的数据需要有一定范围内的“错分”，又有较大分界区域的**广义最优分类面**。

准确性



泛化性



支持向量机

● 处理噪声和离群点

引入松弛变量 $\xi_n \geq 0$

$$\xi_n = 0$$

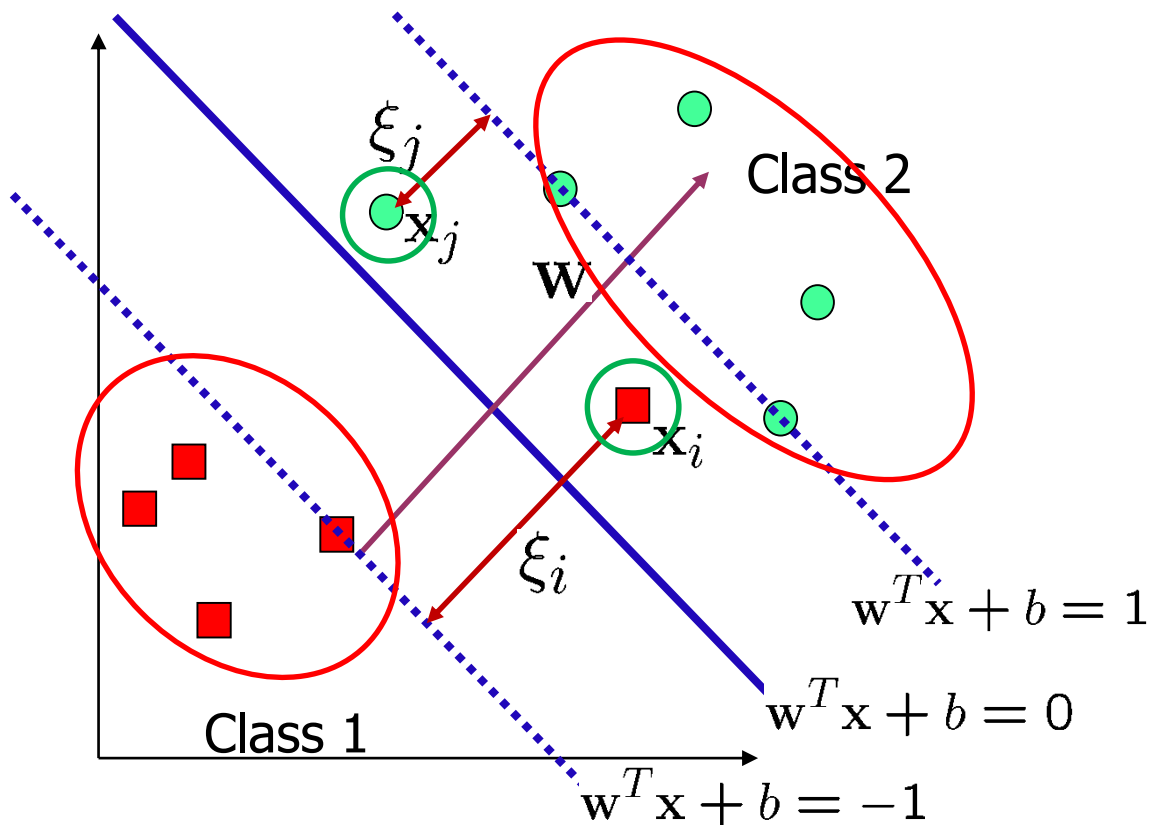
$$\xi_n = |t_n - y(x_n)|$$



$$\xi_n = 0$$

$$0 < \xi_n \leq 1$$

$$\xi_n > 1$$



原有约束 $t_n y(x_n) \geq 1 \rightarrow t_n y(x_n) \geq 1 - \xi_n$

支持向量机

● 处理噪声和离群点

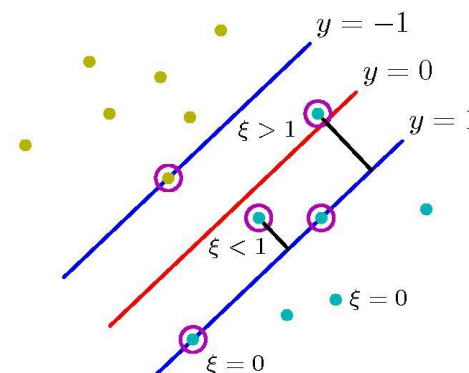
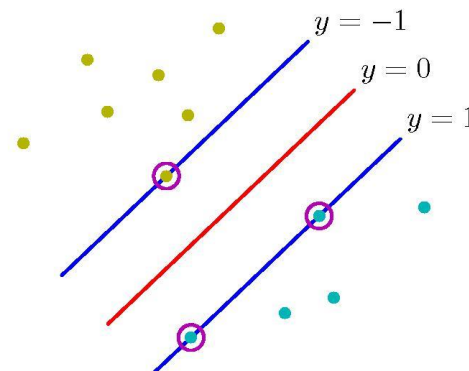
这种处理方式也被视为是从硬间隔(Hard Margin)向软间隔(Soft Margin)的转变。

硬间隔

$$\min_{w,b} \frac{1}{\|w\|^2}$$
$$s.t. \quad t_n(w^T x_n + b) \geq 1, \quad n = 1, \dots, N$$

软间隔

$$\min_{w,b,\xi} \frac{1}{\|w\|^2} + C \sum_{n=1}^N \xi_n$$
$$s.t. \quad t_n(w^T x_n + b) \geq 1 - \xi_n, \quad n = 1, \dots, N$$
$$\xi_n \geq 0$$



支持向量机

● 处理噪声和离群点

利用拉格朗日乘子法求解：

$$L(w, b, \xi, a, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{n=1}^N \xi_n - \sum_{n=1}^N a_n \{t_n y(x_n) - 1 + \xi_n\} - \sum_{n=1}^N \mu_n \xi_n$$

$$a_n \geq 0; \mu_n \geq 0$$

KKT条件：

$$a_n \geq 0$$

$$t_n y(x_n) - 1 + \xi_n \geq 0$$

$$a_n (t_n y(x_n) - 1 + \xi_n) = 0$$

$$\mu_n \geq 0$$

$$\xi_n \geq 0$$

$$\mu_n \xi_n = 0$$

优化 $w, b, \{\xi_n\}$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^N a_n t_n x_n$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^N a_n t_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n = C - \mu_n$$

代入 L 化简：

$$\tilde{L}(a) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m x_n^T x_m$$

支持向量机

● 处理噪声和离群点

得到其对偶形式:

$$\max_a \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m x_n^T x_m$$

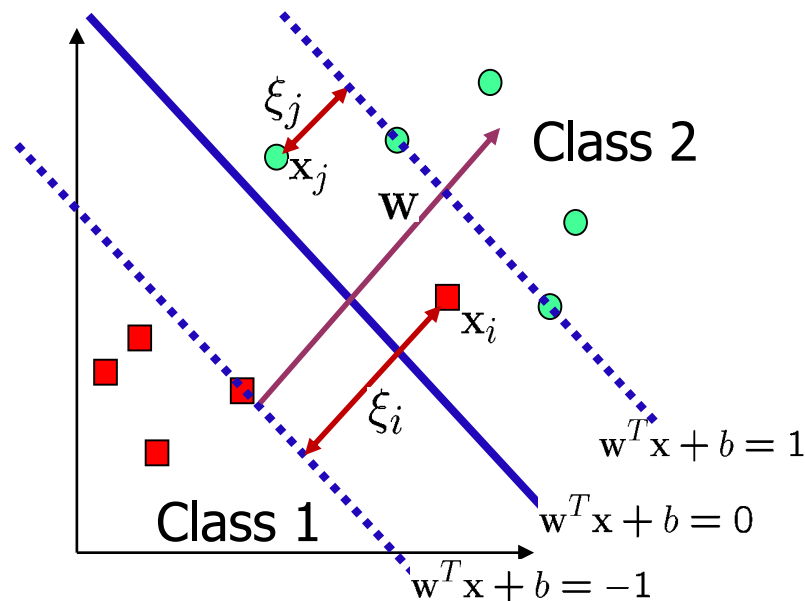
$$s.t. \ 0 \leq a_n \leq C, n = 1, 2, \dots, N$$

$$\sum_{n=1}^N a_n t_n = 0 \quad \text{二次规划问题}$$

对于新样本预测的分类器:

$$y(x) = \sum_{n=1}^N a_n t_n x^T x_n + b$$

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} (t_n - \sum_{m \in \mathcal{S}} a_m t_m x_n^T x_m)$$



非支持向量: $a_n = 0$

支持向量: $t_n y(x_n) = 1 - \xi_n$

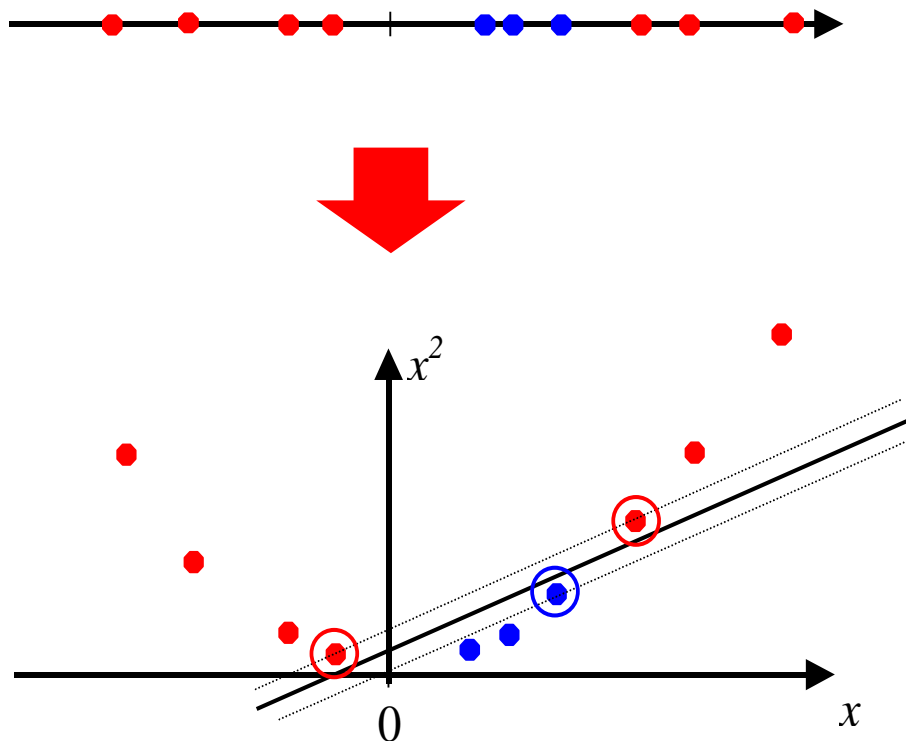
$$a_n < C \Rightarrow \mu > 0 \Rightarrow \xi_n = 0$$

$$a_n = C \Rightarrow \xi_n \leq 1 \text{ or } \xi_n > 1$$

支持向量机

● 非线性支持向量机

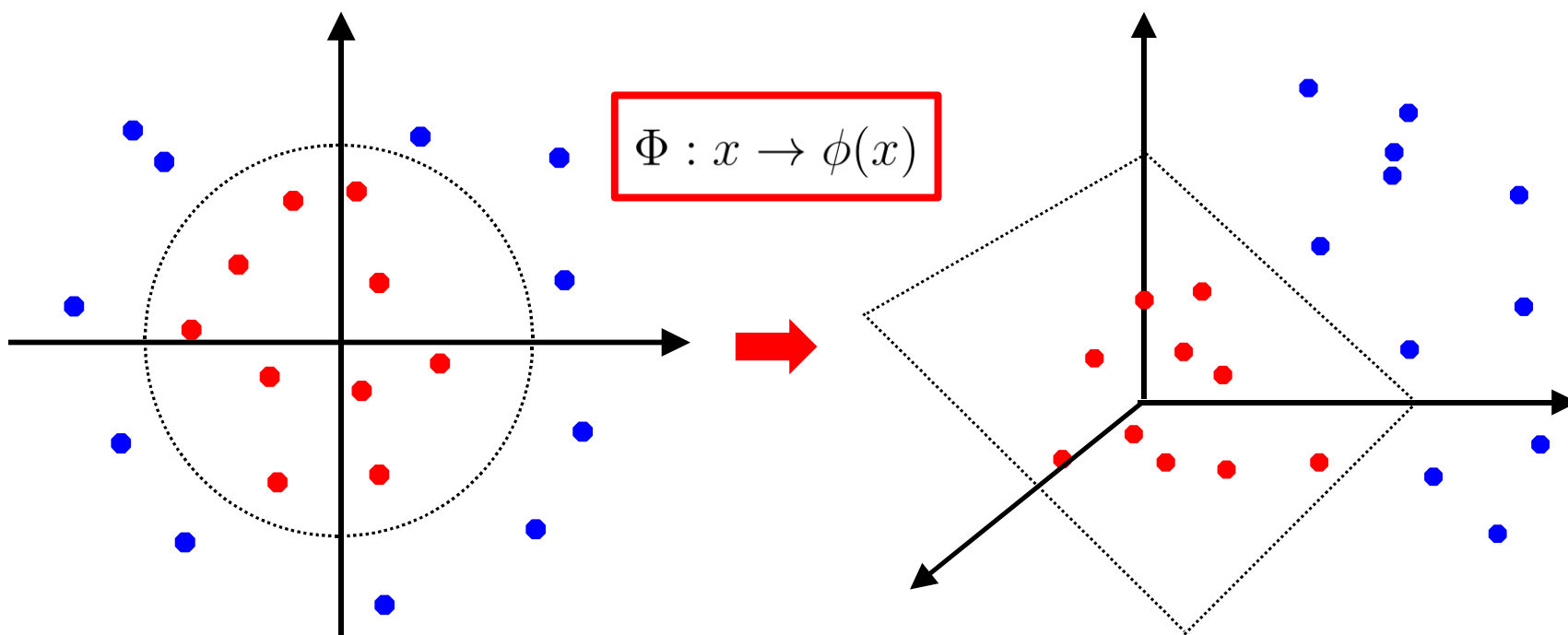
线性模型在解决复杂分类问题时适应性较差。而对于非线性可分的数据样本，可能通过适当的函数变换，将其在**高维空间**中转化为线性可分。



支持向量机

- 非线性支持向量机

可以把样本 x 映射到某个高维特征空间 $\phi(x)$ ，并在其中使用线性分类器。



支持向量机

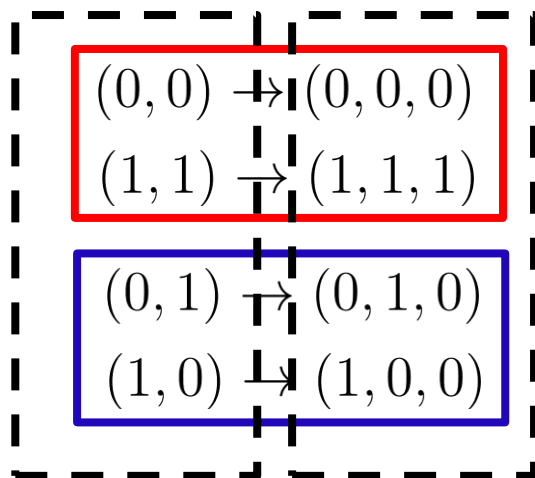
● XOR问题

二维样本集 $x = (x_1, x_2)$

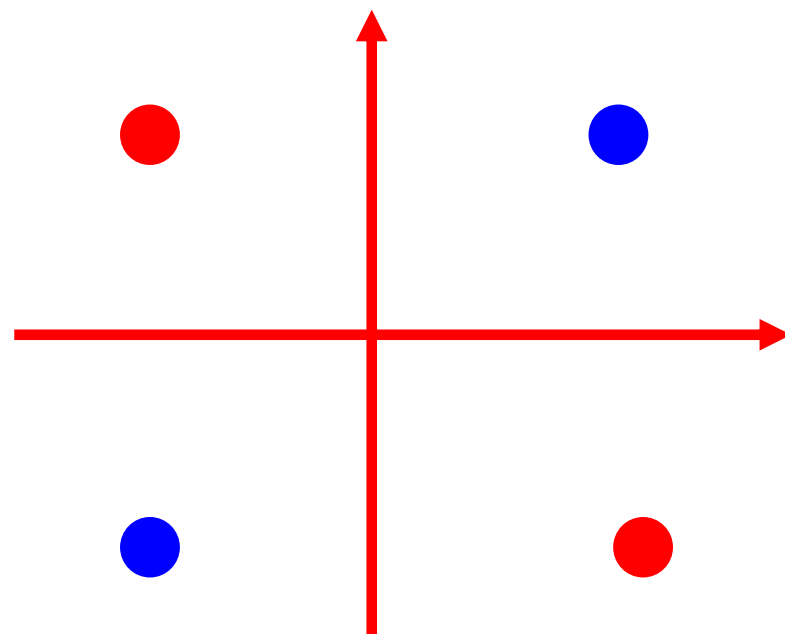
第一类(0, 0)和 (1, 1)， 第二类(1, 0)和 (0, 1)

将二维数据映射到三维

映射函数 $\phi(x) = (x_1, x_2, x_1x_2)$



线性不可分 线性可分



支持向量机

● 非线性支持向量机

利用一个固定的非线性映射将数据映射到特征空间学习的线性分类器等价于基于原始数据学习的非线性分类器。

$$y(x) = w^T x + b \quad \rightarrow \quad y(x) = w^T \phi(x) + b$$

决策时

$$y(x) = \sum_{n=1}^N a_n t_n x^T x_n + b \quad \rightarrow \quad y(x) = \sum_{n=1}^N a_n t_n \boxed{k(x, x_n)} + b$$

$$k(x, x_n) = \phi(x)^T \phi(x_n)$$

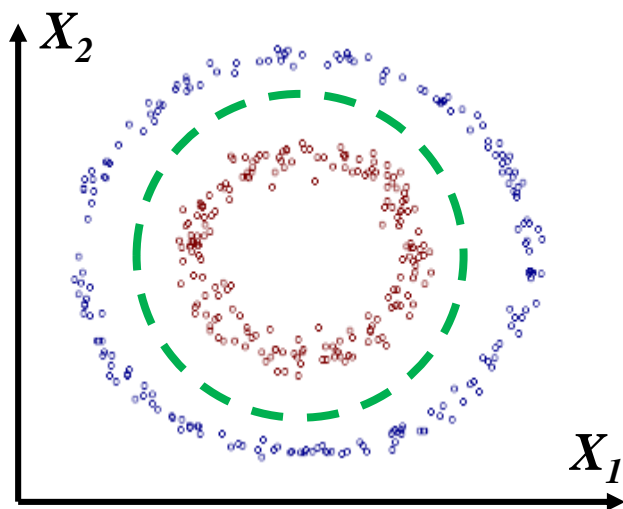
↓
核函数

核函数在特征空间中直接计算数据映射后的内积就像在原始输入数据的函数中计算一样，大大简化了计算过程。

支持向量机

● 非线性支持向量机

利用一个固定的非线性映射将数据映射到特征空间学习的线性分类器等价于基于原始数据学习的非线性分类器。



$$a_1 X_1^2 + a_2 (X_2 - c)^2 + a_3 = 0$$

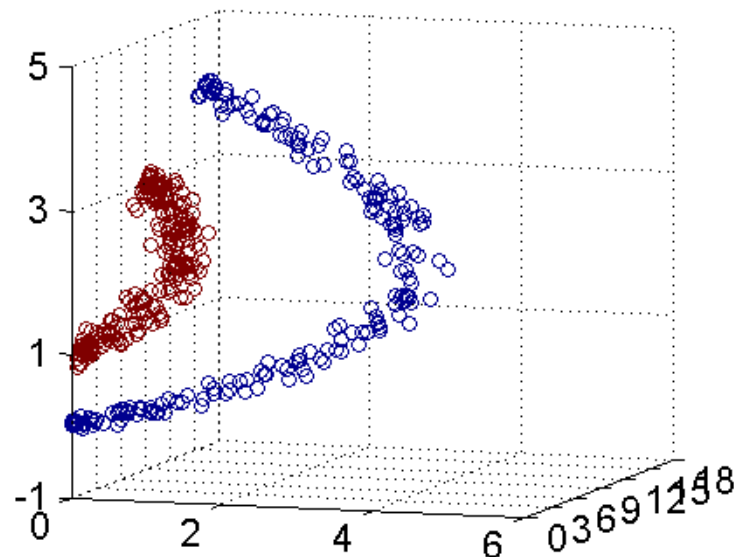
$$Z_1 = X_1^2; Z_2 = X_2^2; Z_3 = X_2$$

$$\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$a_1 X_1 + a_2 X_1^2 + a_3 X_2 + a_4 X_2^2 + a_5 X_1 X_2 + a_6 = 0$$

$$Z_1 = X_1; Z_2 = X_1^2; Z_3 = X_2; Z_4 = X_2^2; Z_5 = X_1 X_2$$

$$\rightarrow \sum_{i=1}^5 a_i Z_i + a_6 = 0 \quad \phi : \mathbb{R}^2 \rightarrow \mathbb{R}^5$$



支持向量机

● 非线性支持向量机

利用一个固定的非线性映射将数据映射到特征空间学习的线性分类器等价于基于原始数据学习的非线性分类器。

$$a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_2^2 + a_5X_1X_2 + a_6 = 0 \quad \phi: \mathbb{R}^2 \rightarrow \mathbb{R}^5$$

原始样本增加到三维

$$\begin{aligned} &a_1X_1^3 + a_2X_2^3 + a_3X_3^3 + a_4X_1^2X_2 + a_5X_1^2X_3 + a_6X_2^2X_1 + \\ &a_7X_2^2X_3 + a_8X_3^2X_1 + a_9X_3^2X_2 + a_{10}X_1X_2X_3 + a_{11}X_1^2 + \\ &a_{12}X_2^2 + a_{13}X_3^2 + a_{14}X_1X_2 + a_{15}X_2X_3 + a_{16}X_1X_3 + \\ &a_{17}X_1 + a_{18}X_2 + a_{19}X_3 + a_{20} = 0 \end{aligned} \quad \phi: \mathbb{R}^3 \rightarrow \mathbb{R}^{19}$$

**维数大大增加
计算变得非常困难**

$$k(x_1, x_2) = (< x_1, x_2 > + 1)^2$$

利用核函数直接在原来的低维空间中进行计算不需要显式地写出映射后的结果，避免了先映射到高维空间中然后再根据内积的公式进行计算

支持向量机

● 非线性支持向量机

根据问题和数据的不同, 选择带有不同的核函数。

一些常用的核函数:

线性核: $k(x_1, x_2) = x_1^T x_2$

多项式核: $k(x_1, x_2) = (< x_1, x_2 > + R)^d$

高斯核: $k(x_1, x_2) = \exp\{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\}$

Sigmoid核: $k(x_1, x_2) = \tanh(\beta_0 x_1^T x_2 + \beta_1)$

如何判断一个函数是
否可以作为核函数?

Mercer定理:

$\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ 上的映射 k 是一个有效核函数(也称Mercer核函数)当且仅当对于训练样本其相应的核函数矩阵是对称半正定的, 即对于任何平方可积函数 $g(x)$ 有 $\int \int k(x, y) g(x) g(y) dx dy \geq 0$ 。

序列最小优化算法

- J. C. Platt(1999年提出)

支持向量机的学习问题可以形式化为求解**具有全局最优解的凸二次规划问题**。许多方法可以用于求解这一问题，但当训练样本容量很大时，这些算法往往效率较低，以致无法使用。

序列最小优化算法(Sequential Minimal Optimization, SMO)是一种启发式算法。基本思想是：**如果所有变量都满足此优化问题的KKT条件，那么这个问题的解就得到了**。

SMO算法的特点是不不断地将原二次规划问题分解为只有两个变量的二次规划问题，并对子问题进行解析求解，直到所有变量都满足KKT条件为止。因为子问题解析解存在，所以每次计算子问题都很快，虽然子问题次数很多，但是总体上还是高效的。

序列最小优化算法

● SMO算法解凸二次规划问题

$$\min_a \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(x_n, x_m) - \sum_{n=1}^N a_n$$

$$s.t. \ 0 \leq a_n \leq C, n = 1, 2, \dots, N$$

$$\sum_{n=1}^N a_n t_n = 0$$

子问题有两个变量，一个是违反KKT条件最严重的，另一个有约束条件自动确定。两个变量中只有一个是自由变量。假设 a_1, a_2 为两个变量， a_3, a_4, \dots, a_N 固定，那么：

$$a_1 = -t_1 \sum_{n=2}^N a_n t_n = 0$$

即 a_2 确定， a_1 也随之确定。

SMO算法包括：**求解两个变量二次规划的解析方法和选择变量的启发式方法。**

序列最小优化算法

● SMO算法

输入：训练数据集 $T = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ ，其中 $x_i \in \mathcal{X} = R^n$ ， $y_i \in \mathcal{Y} = \{-1, +1\}, i = 1, 2, \dots, N$ ，精度 ϵ ；

输出：近似解 \hat{a} 。

(1) 取初值 $a^{(0)} = 0$ ，令 $k = 0$ ；

(2) 选取优化变量 a_1^k, a_2^k ，解析求解两个变量的最优化问题，求得最优解 a_1^{k+1}, a_2^{k+1} ，更新 a 为 a^{k+1} ；

(3) 若在精度 ϵ 范围内满足停止条件
其中，

$$g(x_i) = \sum_{j=1}^N a_j y_j K(x_j, x_i) + b$$

$$\begin{aligned} \sum_{i=1}^N a_i y_i &= 0 \\ 0 \leq a_i &\leq C, i = 1, 2, \dots, N \\ y_i \cdot g(x_i) &= \begin{cases} \geq 1, & \{x_i | a_i = 0\} \\ = 1, & \{x_i | 0 < a_i < C\} \\ \leq 1, & \{x_i | a_i = C\} \end{cases} \end{aligned}$$

则转(4)；否则令 $k = k + 1$ ，转(2)；

(4) 取 $\hat{a} = a^{(k+1)}$ 。

第6讲：采样方法

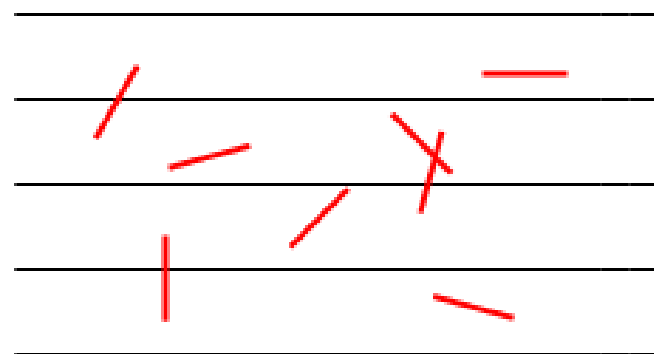
Chapter 6: Sampling Methods

布丰投针问题

- 1777年法国科学家布丰（Buffon）提出一种计算圆周率 π 的方法—随机投针法。

1、在平面上画上间距为 a 的平行线

2、取一根长度为 l ($l < a$) 的针，
随机投掷于平面上，针与任一线相交的概率 p 为



$$p = \frac{2l}{\pi a}$$

π 的值可计算为：

$$\pi = \frac{2l}{ap} \approx \frac{2l}{a} \left(\frac{N}{n} \right)$$

N 为总的投针次数， n 为与平行线相交次数

试验者	时间	投掷次数	相交次数	圆周率估计值
Wolf	1850年	5000	2532	3.1596
Smith	1855年	3204	1218.5	3.1554
C.De Morgan	1860年	600	382.5	3.137
Fox	1884年	1030	489	3.1595
Lazzerini	1901年	3408	1808	3.1415929

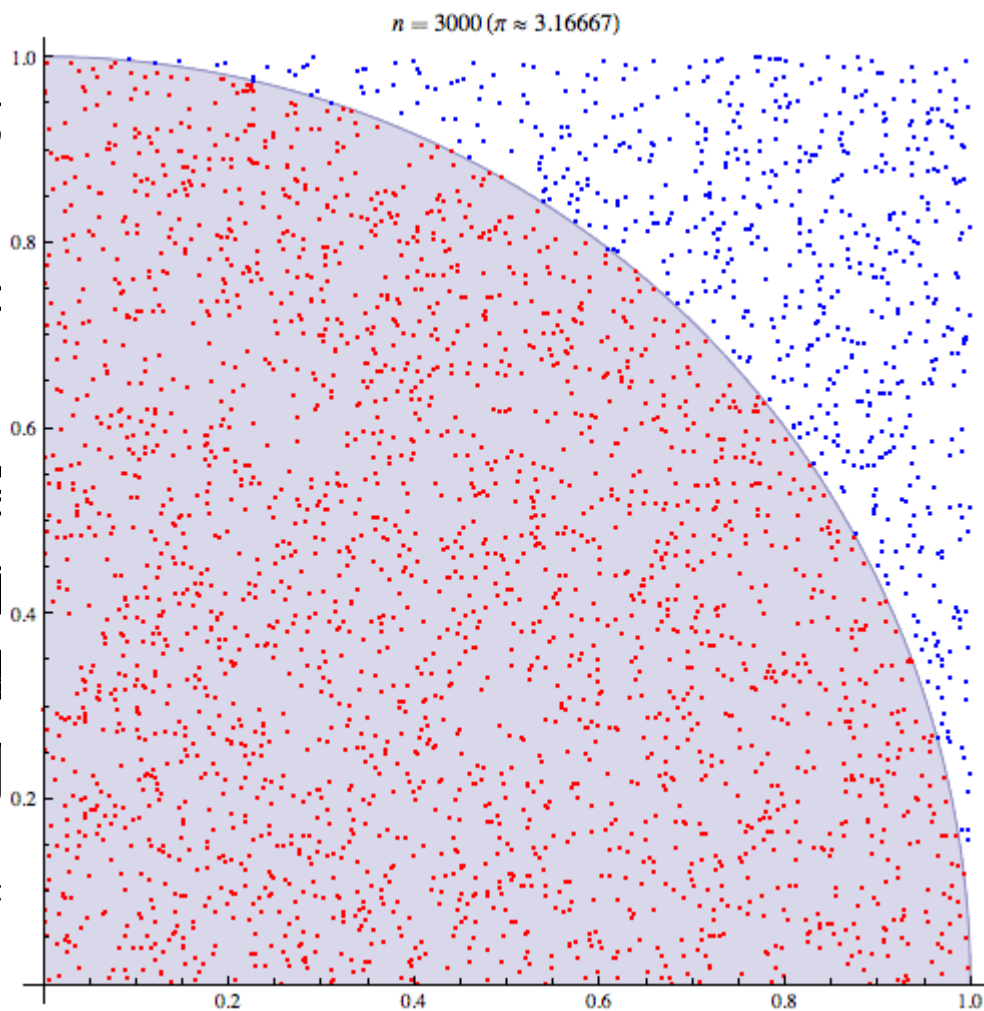
计算 π - 另一个例子

- 正方形内部

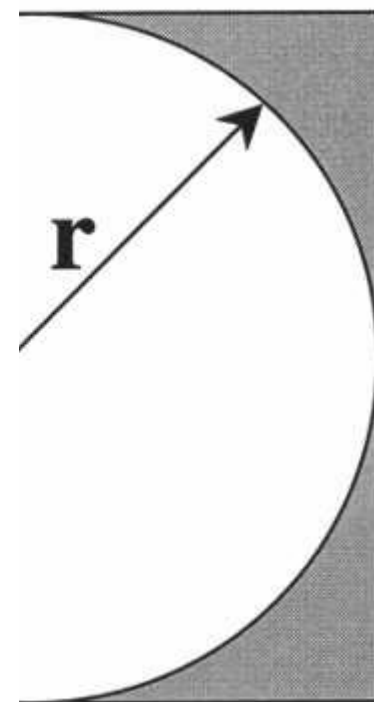
$$\frac{A_{\text{circle}}}{A_{\text{square}}} =$$

在正方形内随机
个点，计算个
距离，判断
计为 N_c ，则

$$\pi \approx 4$$



$\pi/4$

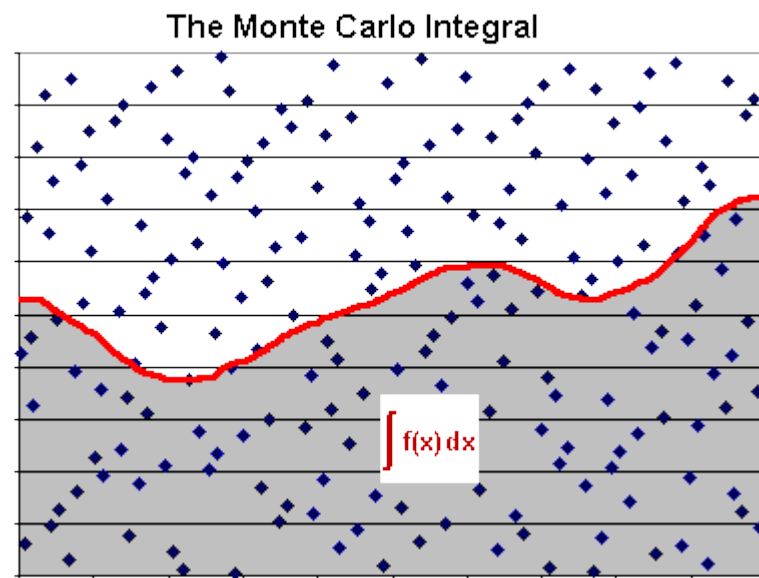


复杂函数定积分计算

- 上面的方法称为 **蒙特卡罗方法 (Monte Carlo method, Monte Carlo Simulation)** ,可以用于计算任一个函数的定积分。
- 如要求 $\int_a^b f(x) dx$ 积分, 而 $f(x)$ 积分的解析形式又很难求, 可通过数值方法求其近似值:

$$\begin{aligned}\int_a^b f(x) dx &= \int_a^b \frac{f(x)}{q(x)} q(x) dx \\ &\approx \frac{1}{N} \sum_{i=1}^N \frac{f(x_i)}{q(x_i)}\end{aligned}$$

其中 $q(x)$ 是某种容易采样的分布, x_i 是服从 $q(x)$ 分布的随机样本





姚期智

(Andrew Chi-Chih Yao, 1946年12月24日 -)

计算机科学家
2000年图灵奖得主

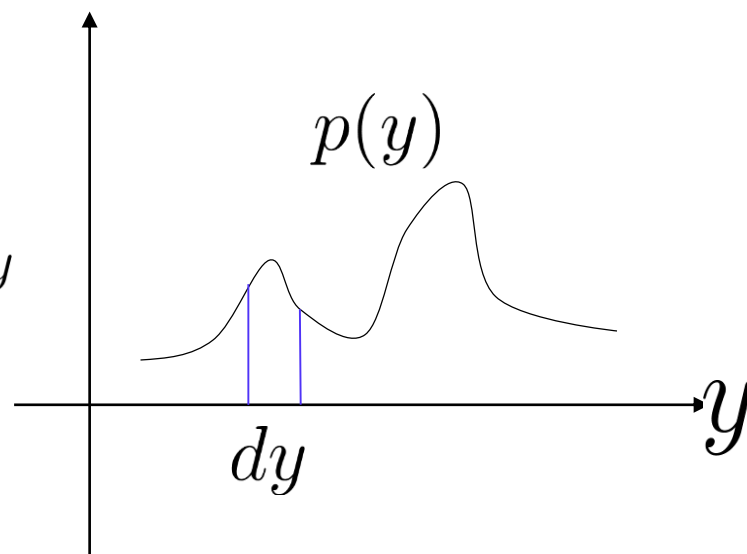
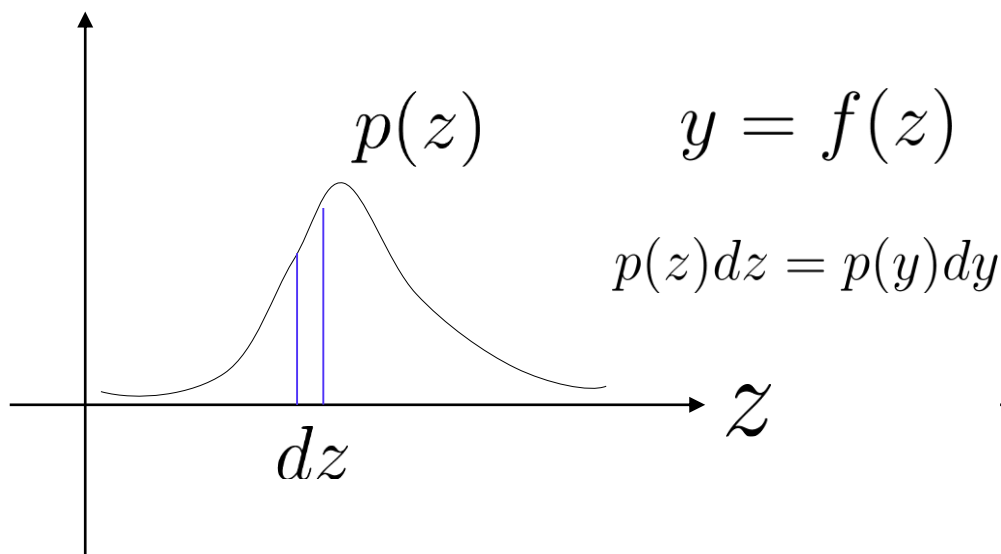
2000年，因为姚期智对计算理论，包括**伪随机数生成，密码学与通信复杂性**的诸多贡献，计算机协会（ACM）授予他该年度的图灵奖。

基本采样法 (Basic Sampling)

- 思想：从基本概率分布中产生新变量的分布

➤ 均匀分布 (Uniform distribution) : $p(z) = 1 \quad z \in (0, 1)$

➤ 产生非均匀分布: $p(y), \quad y = f(z)$



基本采样法 (Basic Sampling)

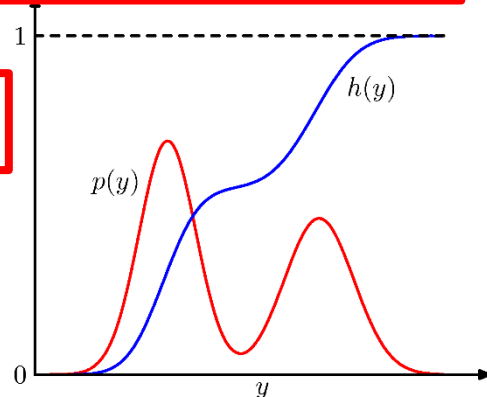
➤ 均匀分布: $p(z) = 1 \quad z \in (0, 1)$

➤ 非均匀分布: $p(y), \quad y = f(z)$

$$\left. \begin{array}{l} p(y) = p(z) \left| \frac{dz}{dy} \right| \\ p(z) = 1 \end{array} \right\} \rightarrow z = h(y) \equiv \int_{-\infty}^y p(\hat{y}) d\hat{y}$$

累积分布函数 (CDF)

$$\rightarrow y = h^{-1}(z)$$



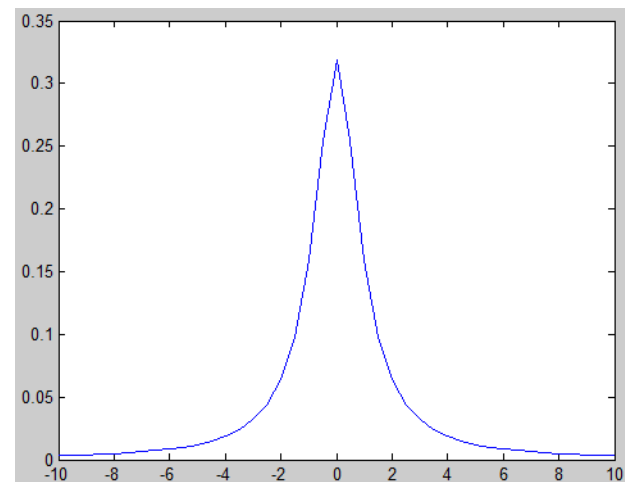
基本采样法 (Basic Sampling)

- 例子：标准柯西分布

已知 $z \sim U(0, 1)$

求 $y = f(z)$

使 $p(y) = \frac{1}{\pi} \frac{1}{1+y^2}$



➡
$$h(y) = \int_{-\infty}^y p(\hat{y}) d\hat{y} = \frac{1}{\pi} \arctan y + \frac{1}{2}$$

➡
$$y = h^{-1}(z) = \pi \tan(z - \frac{1}{2})$$

基本采样法 (Basic Sampling)

- **练习：指数分布** $p(y) = \lambda \exp(-\lambda y) \quad y \in [0, \infty)$

求 $y = f(z)$

➡ $h(y) = \int_{-\infty}^y p(\hat{y}) d\hat{y} = 1 - \exp(-\lambda y)$

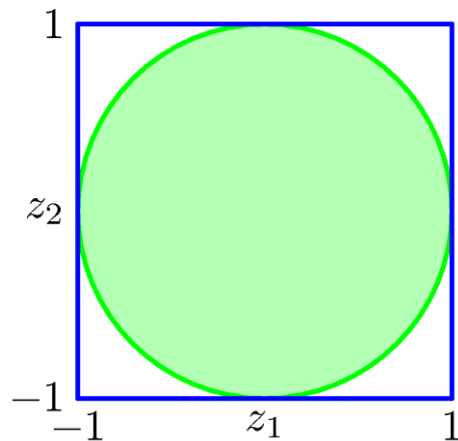
➡ $y = h^{-1}(z) = -\lambda^{-1} \ln(1 - z)$

- **多变量分布形式：**

$$p(y_1, \dots, y_M) = p(z_1, \dots, z_M) \left| \frac{\partial(z_1, \dots, z_M)}{\partial(y_1, \dots, y_M)} \right|$$

基本采样法 (Basic Sampling)

● 高斯分布:



$$z_1, z_2 \sim \text{unif}$$

$$y_1 = \sqrt{-2 \ln z_1} \cos(2\pi z_2)$$

$$y_2 = \sqrt{-2 \ln z_1} \sin(2\pi z_2)$$

$$p(z_1, z_2) = \frac{1}{\pi}, \quad (z_1 + z_2)^2 \leq 1$$

$$y_1 = z_1 \left(\frac{-2 \ln z_1}{r^2} \right)^{1/2}$$

$$y_2 = z_2 \left(\frac{-2 \ln z_2}{r^2} \right)^{1/2}$$

$$r^2 = (z_1 + z_2)^2$$

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

Box-Muller变换

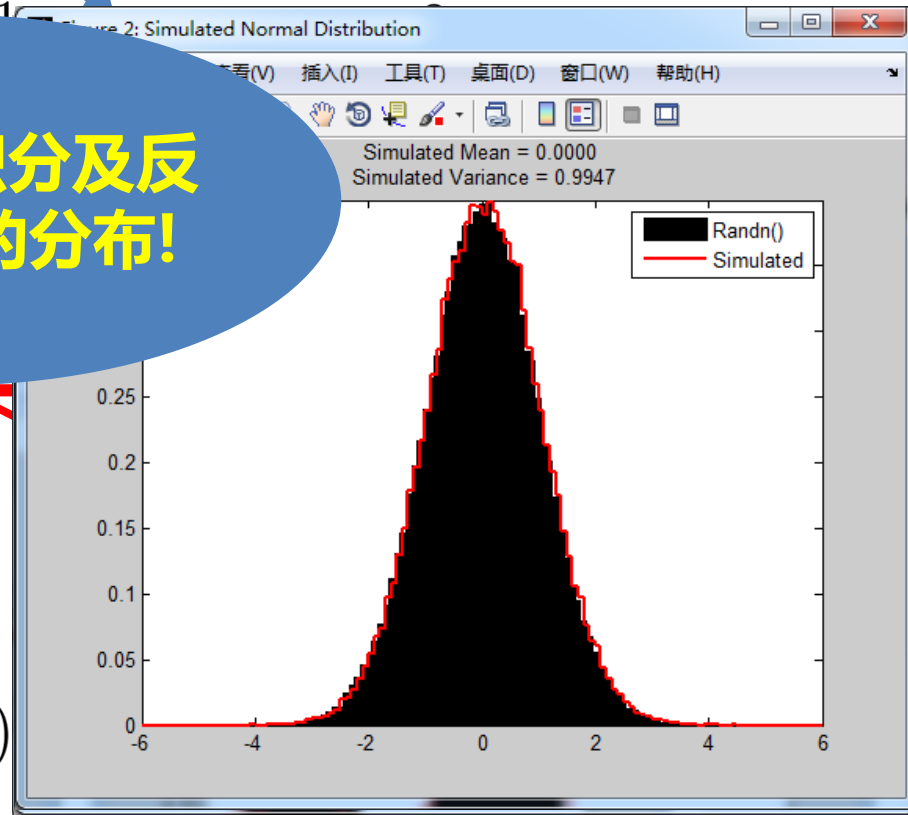
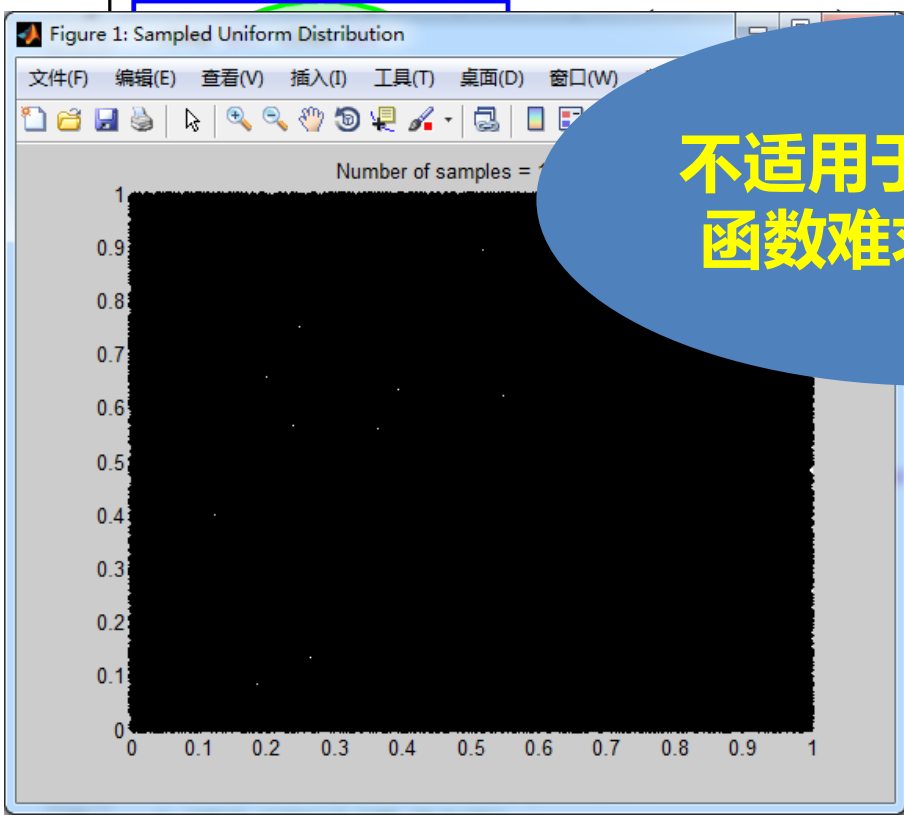
$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]$$

基本采样法 (Basic Sampling)

- 高斯分布:

$$z_1, z_2 \sim \text{Uniform}(0, 1)$$
$$y_1 = \sqrt{-2 \ln z_1} \cos(2\pi z_2)$$
$$y_2 = \sqrt{-2 \ln z_1} \sin(2\pi z_2)$$

不适用于积分及反函数难求的分布!



拒绝采样 (Rejection Sampling)

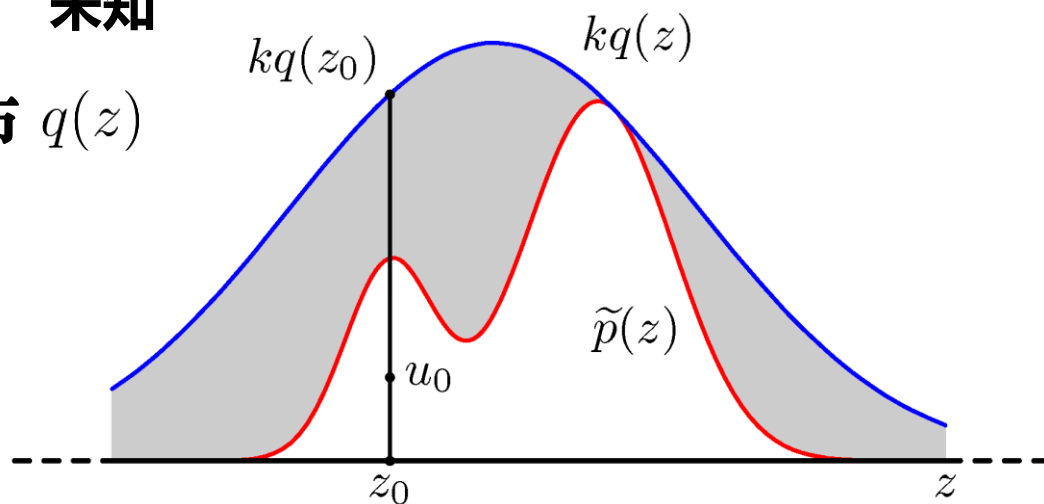
- 对一个很复杂的分布 $p(z)$ 进行采样，但又无法给出其具体的解析形式，但是每个 z 可以计算其概率

$$p(z) = \frac{1}{Z_p} \tilde{p}(z)$$

已知
未知

- 借助一个容易采样的分布 $q(z)$ 去逼近 $\tilde{p}(z)$

$q(z)$ 称为 **建议分布**
proposal distribution

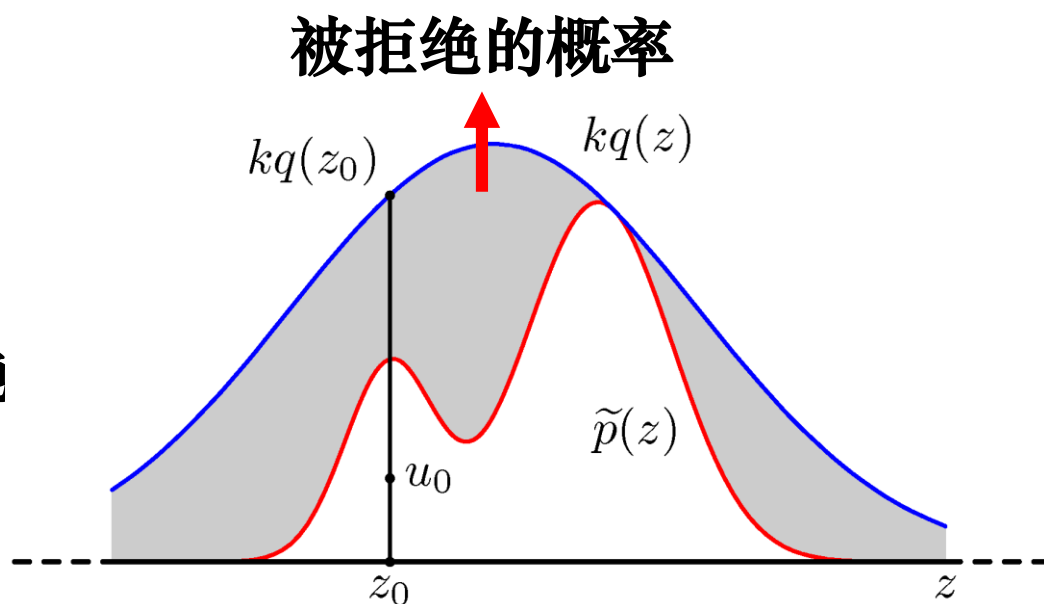


拒绝采样 (Rejection Sampling)

- 找到一个常数 k $\tilde{p}(z) \leq kq(z)$ \rightarrow 对比函数 comparison function

- 采样步骤

- 1、从 $q(z)$ 中产生 z_0
- 2、从 $[0, kq(z_0)]$ 均匀分布中产生 u_0
- 3、如果 $u_0 > \tilde{p}(z_0)$ 拒绝采样，否则接受采样



$$p(\text{accept}) = \int \{\tilde{p}(z)/kq(z)\}q(z)dz = \frac{1}{k} \int \tilde{p}(z)dz$$

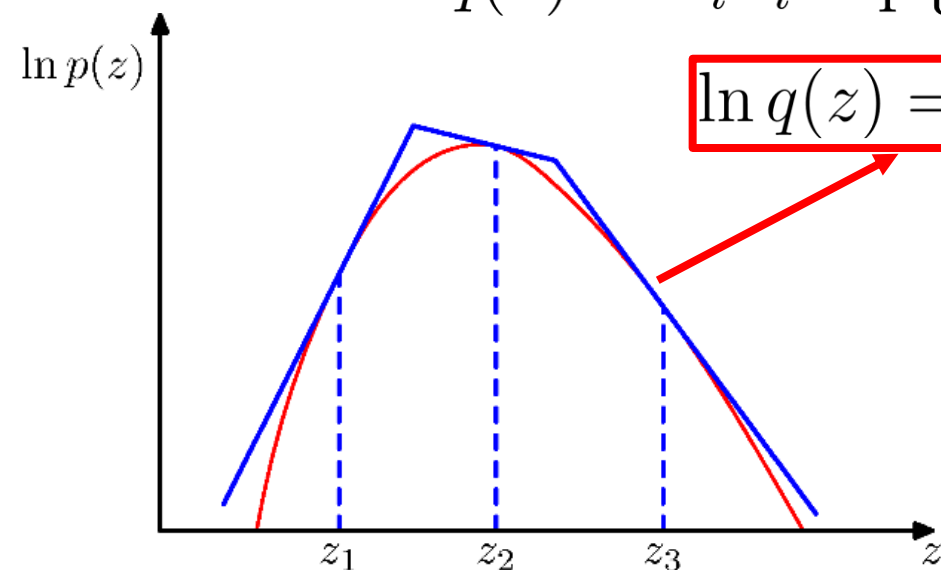
拒绝采样又称为接受-拒绝采样 (Acceptance-Rejection Sampling)

自适应拒绝采样 (Adaptive Rejection Sampling)

- 实际应用中，往往很难找到合适的 $q(z)$
- 特别地，当 $p(z)$ 为log凸函数时，可采用ARS

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_{i-1})\} \quad z_{i-1} < z \leq z_i$$

$$\ln q(z) = C - \lambda(z - z_{i-1}), \quad z_{i-1} < z \leq z_i$$



在log域执行拒绝采样

- 如果满足，接受
- 如果拒绝，重新逼近

重要性采样 (Importance Sampling)

- 对于某个变量 $z \sim p(z)$
- 一个关于 z 函数 $f(z)$ ，预测 $f(z)$ 的关于 z 的期望：

$$\mathbb{E}(f) = \int f(z)p(z)dz$$

$p(z)$ 很复杂，但可以估算每个出现 z 的概率 $p(z)$

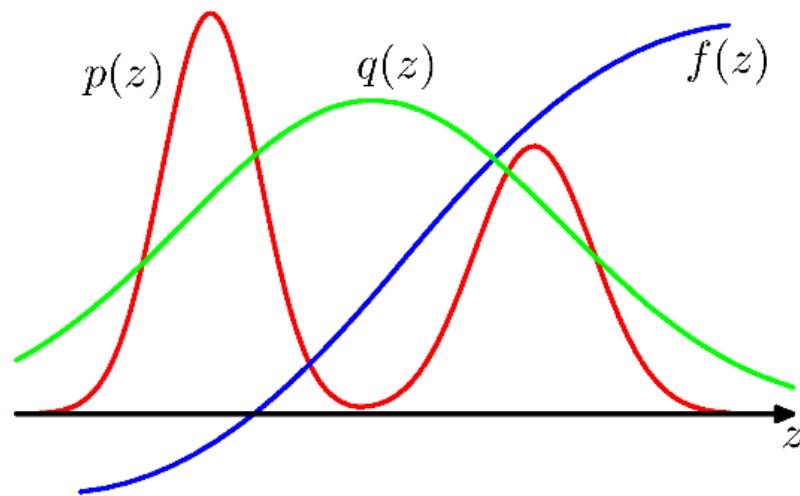
- 要估计 $\mathbb{E}(f) = \int f(z)p(z)dz$
- 可以按如下方式计算-将空间网络化

$$\mathbb{E}(f) \approx \sum_{l=1}^L p(z^{(l)})f(z^{(l)}) \longrightarrow \text{维数灾难}$$

重要性采样 (Importance Sampling)

- 与拒绝采样类似，借助一个容易采样的建议分布 $q(z)$

$$\begin{aligned}\mathbb{E}(f) &= \int f(z)p(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{L} \sum_{l=1}^L \boxed{\frac{p(z^{(l)})}{q(z^{(l)})}} f(z^{(l)})\end{aligned}$$



重要性权重

注: $p(z) = \frac{1}{Z_p} \tilde{p}(z)$

重要性采样 (Importance Sampling)

- 同样地，我们也希望 $q(z)$ 具有类似的性质，即

$$q(z) = \frac{1}{Z_q} \tilde{q}(z)$$


- 再看 $\mathbb{E}(f)$

$$\begin{aligned}\mathbb{E}(f) &= \int f(z)p(z)dz \\ &= \frac{Z_q}{Z_p} \int f(z) \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z) dz \\ &\approx \boxed{\frac{Z_q}{Z_p}} \frac{1}{L} \sum_{l=1}^L \tilde{r}_l f(z^{(l)})\end{aligned}$$

$$\text{其中 } \tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$$

重要性采样 (Importance Sampling)

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q} \int \tilde{p}(z) dz \\ &= \int \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z) dz \\ &\approx \frac{1}{L} \sum_{l=1}^L \tilde{r}_l\end{aligned}$$


$$\mathbb{E}(f) \approx \sum_{l=1}^L w_l f(z^{(l)})$$

$$w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(l)})/q(z^{(l)})}{\sum_m \tilde{p}(z^{(m)})/q(z^{(m)})}$$

马尔可夫蒙特卡罗方法 (Markov Chain Monte Carlo)

- 蒙特卡罗 (Monte Carlo) 蒙特卡洛坐落于欧洲地中海之滨、法国的东南方，有一个版图很小的国家**摩纳哥公国**。蒙特卡罗是世界著名的赌城，是摩纳哥的标志，**世界三大赌城之一**。富丽堂皇的蒙地卡罗赌场，建成于一八六三年，是一幢古色古香以及巍峨的宫殿式建筑物。1856年，摩纳哥亲王Charles三世为了解决财政危机，才在市区北边开设了第一家赌场。
- 蒙特卡罗方法于20世纪40年代美国在第二次世界大战中研制原子弹的“曼哈顿计划”计划的成员**S.M.乌拉姆**和**J.冯·诺伊曼**首先提出。数学家冯·诺伊曼用驰名世界的赌城—摩纳哥的Monte Carlo—来命名这种方法，为它蒙上了一层神秘色彩。在这之前，蒙特卡罗方法就已经存在。1777年，法国Buffon提出用投针实验的方法求圆周率 π 。这被认为是蒙特卡罗方法的起源。



蒙特卡罗方法 (Monte Carlo Method)

注:

$$p(z) = \frac{1}{Z_p} \tilde{p}(z)$$

- 首先产生一个采样点 $z^{(\tau)}$
- 根据建议概率 $q(z|z^{(\tau)})$ 产生新的采样点
- 依次类推, 产生马尔可夫链 $z^{(1)}, z^{(2)}, \dots$
- 要求 $q(z|z^{(\tau)})$ 尽可能简单, 便于产生采样点;
- 有一个准则去决定是接受还是拒绝产生的采样点

基本Metropolis采样法

- 建议概率:

$$q(\mathbf{z}_t)$$

- 接受概率:

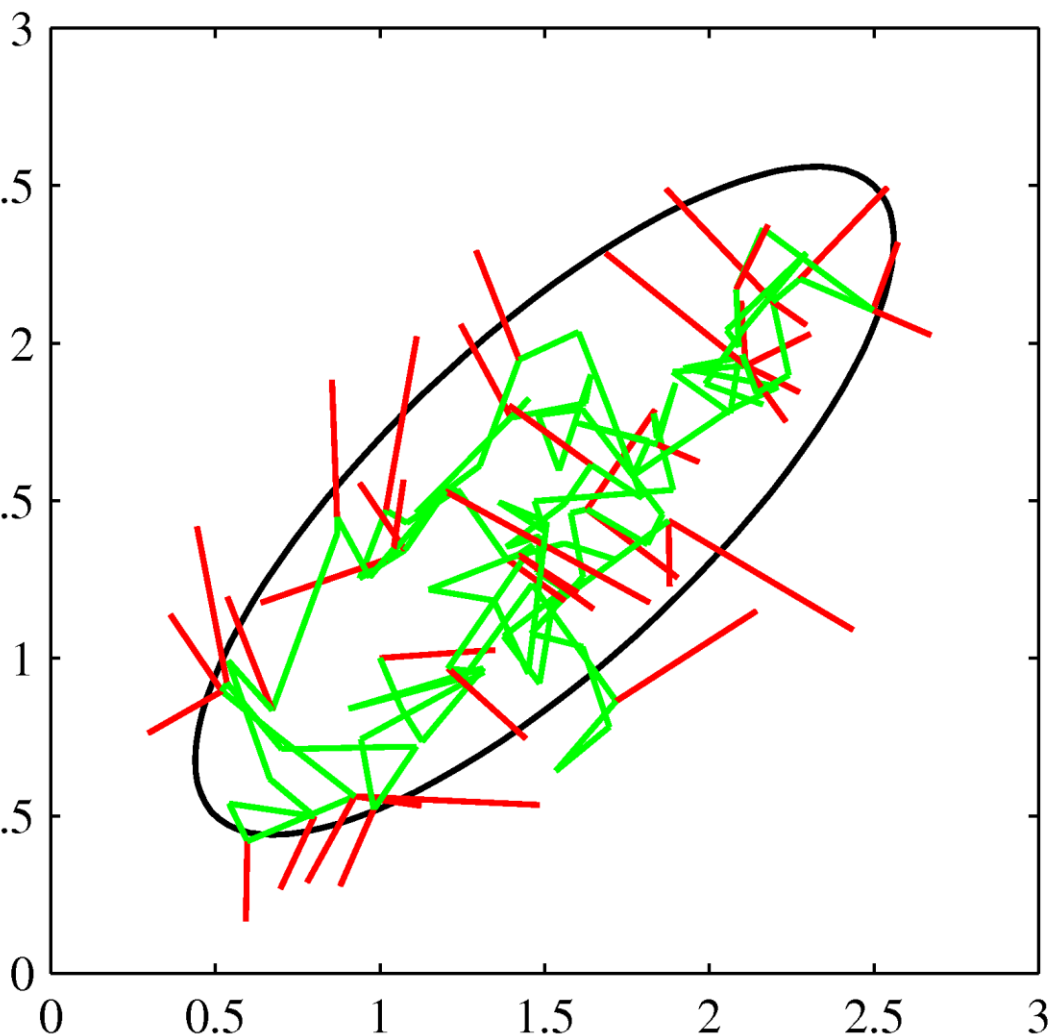
$$A(\mathbf{z}_t)$$

- 在(0,1)的均匀分布

- 如果 $u < A(\mathbf{z}_t)$

- 否则拒绝;

$\tau \rightarrow \infty$



马尔可夫链 (Markov Chain)

- 一阶马尔可夫链 (First Order Markov Chain)

$$p(z^{(m+1)} | z^{(1)}, \dots, z^{(m)}) = p(z^{(m+1)} | z^{(m)})$$

- 高阶马尔可夫 (High Order Markov Chain)

$$p(z^{(m+1)} | z^{(1)}, \dots, z^{(m)}) = p(z^{(m+1)} | z^{(m)}, \dots, z^{(m-n)})$$

- 转移概率 (Transition Probabilities)

$$T(z^{(m)}, z^{(m+1)}) \equiv p(z^{(m+1)} | z^{(m)})$$

- 转移概率矩阵

$$T = \begin{bmatrix} T(1,1), T(1,2), \dots, T(1,m) \\ T(2,1), T(2,2), \dots, T(2,m) \\ \vdots & \ddots & \vdots \\ T(m,1), T(m,2), \dots, T(m,m) \end{bmatrix}$$

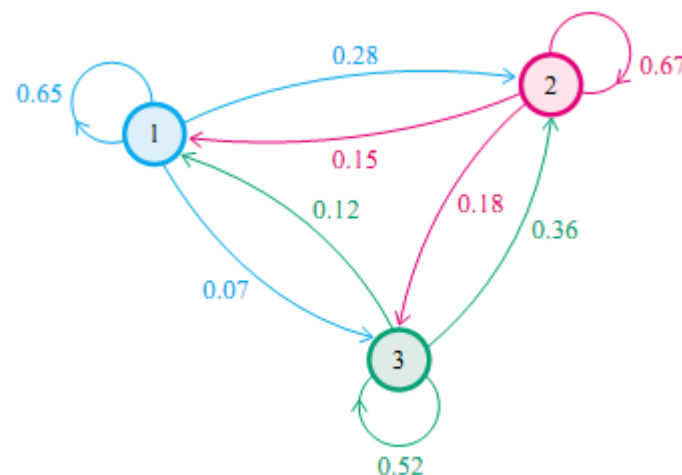
- 如果对所有的 $z^{(m)}$ 都有相同的转移概率 T_m ，则称为齐次马尔可夫 (Homogeneous Markov)

马尔可夫链 (Markov Chain)

● 例子：

社会学家经常把人按其经济状况分成3类：下层(lower-class)、中层(middle-class)、上层(upper-class)，我们用1,2,3 分别代表这三个阶层。社会学家们发现决定一个人的收入阶层的最重要的因素就是其父母的收入阶层。如果一个人的收入属于下层类别，那么他的孩子属于下层收入的概率是 0.65, 属于中层收入的概率是 0.28, 属于上层收入的概率是 0.07。事实上，从父代到子代，收入阶层的变化的转移概率如下

		子代		
	State	1	2	3
父代	1	0.65	0.28	0.07
	2	0.15	0.67	0.18
	3	0.12	0.36	0.52



马尔可夫链 (Markov Chain)

● 例子:

写成矩阵的形式: $T = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$

如果把当前这一段人处在下、中、上层的比例

那么他们子女所处阶层的分布比例将是: p_1

他们孙子代各阶层的分布比例将是: $p_2 = p_1 T$

第n代子孙各阶层的分布将是: $p_n = p_{n-1} T$

假设初始概率分布为: $p_0 = [0.21, 0.68, 0.11]$

计算后代各阶层的分布为:

第n代人	下层	中层	上层
0	0.210	0.680	0.110
1	0.252	0.554	0.194
2	0.270	0.512	0.218
3	0.278	0.497	0.225
4	0.282	0.490	0.226
5	0.285	0.489	0.225
6	0.286	0.489	0.225
7	0.286	0.489	0.225
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...

马尔可夫链 (Markov Chain)

- 一个状态的边缘分布可以表示为

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)} | z^{(m)}) p(z^{(m)})$$

- 平稳性 (Stationary, 或不变性 Invariant)

$$p^*(z) = \sum_{z'} T(z', z) p^*(z')$$

- 细致平稳 (Detailed balance) - 充分条件

$$p^*(z) T(z, z') = p^*(z') T(z', z)$$

$$\sum_{z'} p^*(z') T(z', z) = \sum_{z'} p^*(z) T(z, z') = p^*(z) \sum_{z'} p(z' | z) = p^*(z)$$

马尔可夫链 (Markov Chain)

- 当 $m \rightarrow \infty$ 马尔可夫链各状态趋于平稳, 即

$$p_m = p_{m-1}T \xrightarrow{m \rightarrow \infty} p = pT \quad p = [p(1), \dots, p(j), \dots]$$

平稳分布

- 同时

$$\lim_{m \rightarrow \infty} T^m = \begin{bmatrix} p(1), \dots, p(j), \dots \\ p(1), \dots, p(j), \dots \\ \dots, \dots \\ p(1), \dots, p(j), \dots \\ \dots, \dots \end{bmatrix} \quad \sum_j p(j) = 1$$

Metropolis-Hastings 方法

- **思想**：对于需要采样的一分布 $p(z)$ ，构造一个转移矩阵为 T 的马尔可夫链，使它的平稳分布恰好为 $p(z)$
- 假设有一个转移矩阵 $Q(z, z') = q(z'|z)$ ， $q(z)$ 为容易采样的分布
- 通常情况下，该转移矩阵难以满足细致平稳条件

$$p(z)q(z'|z) \neq p(z')q(z|z')$$

- 引入 $a(z, z')$ 使

$$p(z)q(z'|z)a(z, z') = p(z')q(z|z')a(z', z)$$

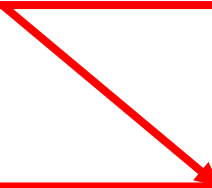
其中：

$$\left. \begin{aligned} a(z, z') &= p(z')q(z|z') \\ a(z', z) &= p(z)q(z'|z) \end{aligned} \right\} \text{接受率}$$

Metropolis-Hastings 方法

- 步骤:

- 1、初始化马尔可夫链状态 $z = z_0$
- 2、对 $\tau = 1, 2, \dots$ 循环以下过程采样
 - 1) 第 τ 个时刻马尔可夫链状态为 $z = z^{(\tau)}$ 采样 $z^* = q(z|z^{(\tau)})$
 - 2) 从均匀分布中采样 $u \sim \text{uniform}[0, 1]$
 - 3) 如果 $u < a(z^{(\tau)}, z^*) = p(z^*)q(z^{(\tau)}|z^*)$ 则接受 $z^{(\tau+1)} = z^*$
 - 4) 否则 $z^{(\tau+1)} = z^{(\tau)}$



如果 $a(z^{(\tau)}, z)$ 过小，则采样效率较低！

Metropolis-Hastings 方法

- 在细致平稳条件两边乘以因子 C

$$p(z)q(z'|z)a(z, z') \cdot C = p(z')q(z|z')a(z', z) \cdot C$$

细致平稳条件并没有打破!!!

- 同比例放大 $a(z, z')$, $a(z', z)$ 使最大的为1, 令

$$\begin{aligned} A(z, z') &= \min \left\{ 1, \frac{p(z)q(z'|z)}{p(z')q(z|z')} \right\} \\ &= \min \left\{ 1, \frac{\tilde{p}(z)q(z'|z)}{\tilde{p}(z')q(z|z')} \right\} \quad p(z) = \frac{1}{Z_p} \tilde{p}(z) \end{aligned}$$

Metropolis-Hastings 方法

- 步骤:

- 1、初始化马尔可夫链状态 $z = z_0$
- 2、对 $\tau = 1, 2, \dots$ 循环以下过程采样
 - 1) 第 τ 个时刻马尔可夫链状态为 $z = z^{(\tau)}$ 采样 $z^* = q(z|z^{(\tau)})$
 - 2) 从均匀分布中采样 $u \sim \text{uniform}[0, 1]$
 - 3) 如果 $u < A(z^{(\tau)}, z^*) = \min \left\{ 1, \frac{\tilde{p}(z^*)q(z'|z^*)}{\tilde{p}(z')q(z^*|z')} \right\}$ 则接受 $z^{(\tau+1)} = z^*$
 - 4) 否则 $z^{(\tau+1)} = z^{(\tau)}$

吉布斯采样 (Gibbs Sampling)

- 一种特殊的M-H采样算法
- 针对多元分布进行采样 $p(\mathbf{z}) = p(z_1, \dots, z_M)$

每次只改变一个维度上的值，保持其他维度不变

$$p(z_1, z_2, z_3)$$

首先：初始化 $(z_1^{(0)}, z_2^{(0)}, z_3^{(0)})$

在第 τ 步，假设已经产生了 $(z_1^{(\tau)}, z_2^{(\tau)}, z_3^{(\tau)})$

➡ 根据 $p(z_1 | z_2^{(\tau)}, z_3^{(\tau)})$ 产生 $z_1^{(\tau+1)}$

➡ 根据 $p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)})$ 产生 $z_2^{(\tau+1)}$

➡ 根据 $p(z_3 | z_1^{(\tau+1)}, z_2^{(\tau+1)})$ 产生 $z_3^{(\tau+1)}$

与M-H的关系

- 建议概率：

$$q_k(\mathbf{z}^*|\mathbf{z}) = p(z_k^*|\mathbf{z}_{\setminus k})$$

- 接受概率：

$$A_k(z^*, z^{(\tau)}) = \min \left\{ 1, \frac{\tilde{p}(z^*)q_k(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q_k(z^*|z^{(\tau)})} \right\}$$

$$p(\mathbf{z}) = p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k}) \quad \mathbf{z}_{\setminus k}^* = \mathbf{z}_{\setminus k}$$

$$A(\mathbf{z}^*, \mathbf{z}) = \frac{p(\mathbf{z}^*)q_k(\mathbf{z}|\mathbf{z}^*)}{p(\mathbf{z})q_k(\mathbf{z}^*|\mathbf{z})} = \frac{p(z_k^*|\mathbf{z}_{\setminus k}^*)p(\mathbf{z}_{\setminus k}^*)p(z_k|\mathbf{z}_{\setminus k}^*)}{p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k})p(z_k^*|\mathbf{z}_{\setminus k})} = 1$$

Slice Sampling

- Metropolis 方法的缺点：
 - 步长太短：走得太慢（可能随机散步）
 - 步长太长：拒绝率很好，效率较差；
- SLICE采样可以自适应调整步长
 - 将 \mathcal{Z} 空间扩展成 (z, u) 空间

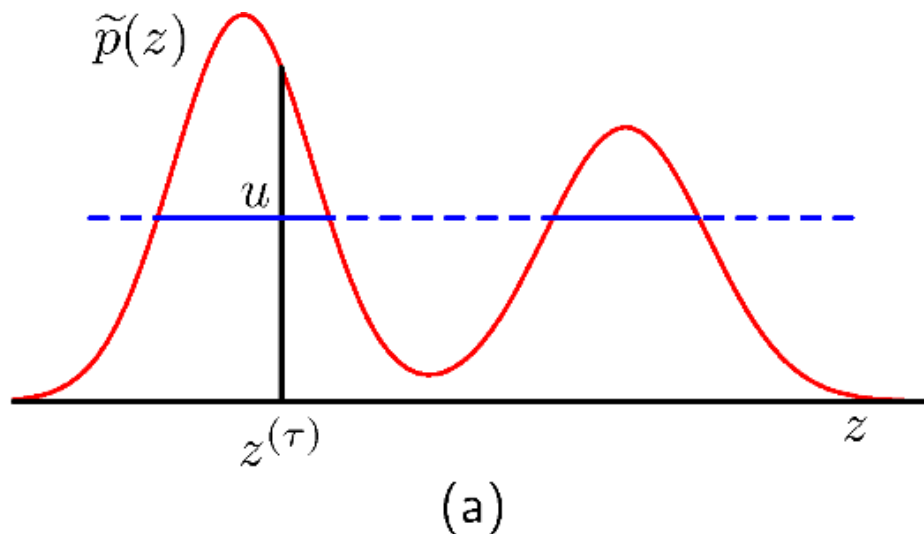
$$\hat{p}(z, u) = \begin{cases} 1/Z_p & \text{if } 0 \leq u \leq \tilde{p}(z) \\ 0 & \text{otherwise} \end{cases}$$

$$\int \hat{p}(z, u) du = \int_0^{\tilde{p}(z)} \frac{1}{Z_p} du = \frac{\tilde{p}(z)}{Z_p} = p(z)$$

Slice Sampling

$$\hat{p}(z, u) = \begin{cases} 1/Z_p & \text{if } 0 \leq u \leq \tilde{p}(z) \\ 0 & \text{otherwise} \end{cases} \quad \int \hat{p}(z, u) du \int_0^{\tilde{p}(z)} \frac{1}{Z_p} du = \frac{\tilde{p}(z)}{Z_p} = p(z)$$

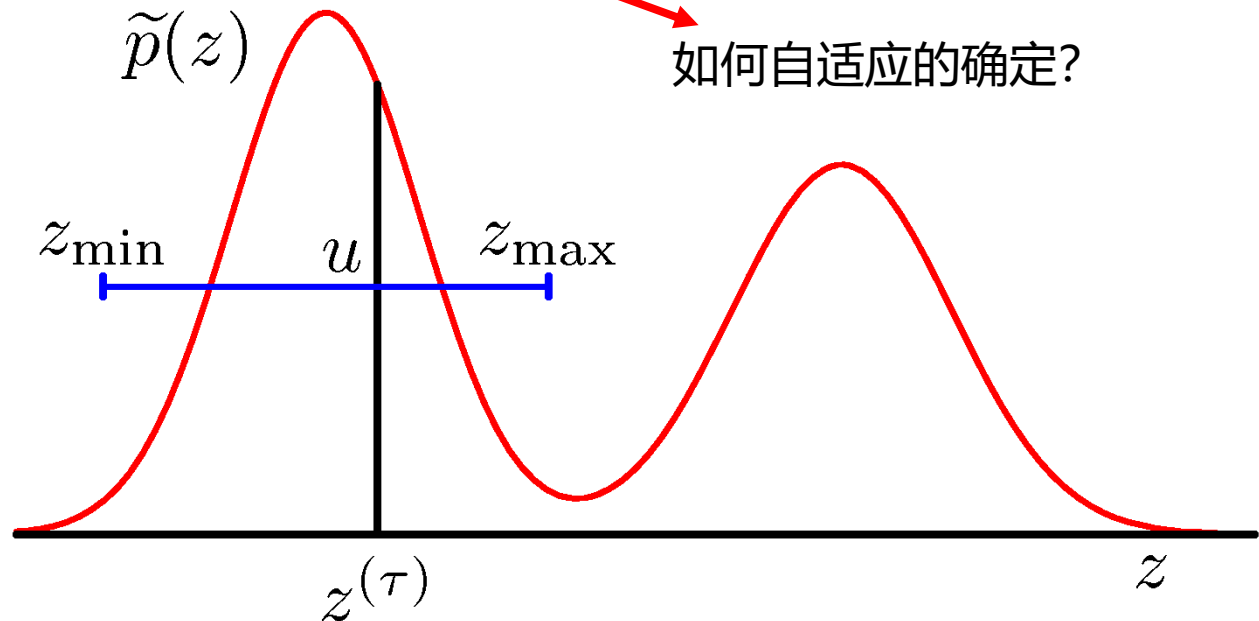
- 第一步：给定 z ，在 $0 \leq u \leq \tilde{p}(z)$ 范围内均匀分布产生 u
- 第二步：给定 u ，在 $\{z : \tilde{p}(z) > u\}$ 范围内均匀分布产生 z



Slice Sampling

- 在实际应用中，很难确定范围

- 第一步：给定 z ，在 $0 \leq u \leq \tilde{p}(z)$ 范围内均匀分布产生 u
- 第二步：给定 u ，在 $z_{\min} \leq z \leq z_{\max}$ 范围内均匀分布产生 z



(b)