# 深度学习理论基础

周洋
上海大学机自学院无人艇工程研究院 & 人工智能研究院

版本：3.14
更新：2025 年 2 月 13 日

# 目录

# 序章：写在前面

## 0.1 雅可比（Jacobi）矩阵的引入

### 0.1.1 泰勒（Taylor）展开

泰勒（Taylor）展开式：

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \tag{1}$$

**证明.** 首先对于任意连续函数，可以考虑 $n$ 次多项式拟合，如下：

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{2}$$

对上式进行连续求导得：

$$\begin{aligned}
&f'(x) = a_1 + 2 \cdot a_2 x + \ldots + n \cdot a_n x^{n-1} \\
&f''(x) = 1 \cdot 2 \cdot a_2 + \ldots + (n-1)(n) \cdot a_n x^{n-2} \\
&\ldots\ldots \\
&f^{(n)}(x) = 1 \cdot 2 \cdot 3 \cdot \ldots (n-2) \cdot (n-1) \cdot (n) a_n
\end{aligned} \tag{3}$$

令 $x = 0$，可得下面被称为 n 次多项式的麦克劳林（Maclaurin）公式：

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \ldots + \frac{f^{(n)}(0)}{n!}x^n \tag{4}$$

显然，多项式拟合也可以以 $(x - x_0)$ 的幂次展开，即写成，

$$f(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)^2 + \ldots + A_n(x - x_0)^n \tag{5}$$

再按上述方法进行求解，便可得到 Taylor 展开的表达式。

对上述一元函数的 Taylor 展开进行拓展：

多元函数（仅保留一阶项）：

$$\begin{aligned}
f(x_1, x_2, \cdots, x_n) &\approx f(0, 0, \cdots, 0) + \frac{\partial f}{\partial x_1}x_1 + \frac{\partial f}{\partial x_2}x_2 + \cdots + \frac{\partial f}{\partial x_n}x_n \\
&= f(0, 0, \cdots, 0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}x_i
\end{aligned} \tag{6}$$

上式中的求和部分也可以写成矩阵形式：

$$\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}x_i=\begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix} \tag{7}$$

多元向量值函数：

$$\begin{bmatrix}f_1(x_1,x_2,\cdots,x_n)\\ f_2(x_1,x_2,\cdots,x_n)\\ \cdots\\ f_m(x_1,x_2,\cdots,x_n)\end{bmatrix}\approx\begin{bmatrix}f_1(0,0,\cdots,0)\\ f_2(0,0,\cdots,0)\\ \cdots\\ f_m(0,0,\cdots,0)\end{bmatrix}+\underline{\text{JaccobiMatrix}}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix} \tag{8}$$

其中，JaccobiMatrix 为**雅可比（Jacobi）矩阵**，满足

$$\text{JaccobiMatrix}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix}_{m\times n} \tag{9}$$

### 0.1.2 多重积分的换元积分

对于重积分来说，Jacobi 行列式其实就是全微分的另一种表示形式：

$$\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\end{pmatrix}=\begin{pmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}\end{pmatrix}\begin{pmatrix}\mathrm{d}r\\ \mathrm{d}\theta\end{pmatrix} \tag{10}$$

$$\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\\ \mathrm{d}z\end{pmatrix}=\begin{pmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \phi}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \phi}\\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \phi}\end{pmatrix}\begin{pmatrix}\mathrm{d}r\\ \mathrm{d}\theta\\ \mathrm{d}\phi\end{pmatrix} \tag{11}$$

柱坐标系可以表示为如下向 Cartesian 坐标系的映射关系（微分同胚）：

$$\boldsymbol{X}(\boldsymbol{x}):\mathbb{R}^3\ni\boldsymbol{x}=\begin{bmatrix}r\\ \theta\end{bmatrix}\mapsto\boldsymbol{X}(\boldsymbol{x})=\begin{bmatrix}r\cos\theta\\ r\sin\theta\end{bmatrix}\in\mathbb{R}^3 \tag{12}$$

考察上述映射关系的雅可比（Jacobi）矩阵，有：

$$\boldsymbol{J}=\frac{\partial(x,y)}{\partial(r,\theta)}=\left[\frac{\partial X^i}{\partial x^j}\right]=\begin{bmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{bmatrix}=\begin{bmatrix}\boldsymbol{g}_r & \boldsymbol{g}_\theta\end{bmatrix} \tag{13}$$

求圆的面积：

$$S=\iint \mathrm{d}(x,y)=\iint \left|\boldsymbol{g}_r\times\boldsymbol{g}_\theta\right|\mathrm{d}\left(r,\theta\right)=\int_0^{2\pi}\mathrm{d}\theta\int_0^r r\,\mathrm{d}r=\pi r^2 \tag{14}$$

球坐标系可以表示为如下向 Cartesian 坐标系的映射关系（微分同胚）：

$$\boldsymbol{X}(\boldsymbol{x}):\mathbb{R}^3\ni\boldsymbol{x}=\begin{bmatrix} r\\ \theta\\ \phi\end{bmatrix}\mapsto\boldsymbol{X}(\boldsymbol{x})=\begin{bmatrix} r\sin\theta\cos\phi\\ r\sin\theta\sin\phi\\ r\cos\theta\end{bmatrix}\in\mathbb{R}^3 \tag{15}$$

考察上述映射关系的雅可比（Jacobi）矩阵，有：

$$\boldsymbol{J}=\frac{\partial(x,y,z)}{\partial(r,\theta,\phi)}=\left[\frac{\partial X^i}{\partial x^j}\right]=\begin{bmatrix}\sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi\\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi\\ \cos\theta & -r\sin\theta & 0\end{bmatrix}=\left[\boldsymbol{g}_r\quad\boldsymbol{g}_\theta\quad\boldsymbol{g}_\phi\right] \tag{16}$$

求球的体积：

$$S=\iiint \mathrm{d}(x,y,z)=\iiint \left|\boldsymbol{g}_r\times\boldsymbol{g}_\theta\cdot\boldsymbol{g}_\phi\right|\mathrm{d}\left(r,\theta,\phi\right)=\int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\int_0^r r^2\sin\theta\mathrm{d}r=\frac{4}{3}\pi r^3 \tag{17}$$

推广至 $N$ 维球体，进行一般化推导可得：

| 维度 | 1 | 2 | 3 | $\cdots$ | $N$ |
|---|---|---|---|---|---|
| 体积 | $2r$ | $\pi r^2$ | $\frac{4}{3}\pi r^3$ | $\cdots$ | $\frac{\pi^{n/2}}{(n/2)!}r^n$ |
| 表面积 | 2 | $2\pi r$ | $4\pi r^2$ | $\cdots$ | $n\frac{\pi^{n/2}}{(n/2)!}r^{n-1}$ |

这里阶乘的定义需要扩充至实数域的 $\Gamma$（Gamma）函数，见附录章节A.1。

# 1 矩阵求导（微分）

## 1.1 导数表示形式

严谨写法：

- Jacobi 矩阵转置形式：

$$\frac{\partial}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}^\top \tag{18}$$

- Jacobi 矩阵形式：

$$\frac{\partial}{\partial \boldsymbol{x}^\top} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{19}$$

人工智能与深度学习领域，大多并不区分两者整体的偏微分写法，需自行分辨。

## 1.2 基本关系式

注 以下关系式中，偏导数的矩阵形式，全部与 Jacobi 矩阵互为转置关系。

1. **标量函数对标量的导数：**
   全微分关系式：

$$y = f(x) \Rightarrow \mathrm{d}y = f'(x)\mathrm{d}x = \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x \tag{20}$$

2. **标量函数对向量的导数：**
   全微分关系式：

$$y = f(\boldsymbol{x}) \Rightarrow \mathrm{d}y = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \\ \vdots \\ \mathrm{d}x_n \end{bmatrix} = \left[\frac{\partial f}{\partial \boldsymbol{x}}\right]^\top \mathrm{d}\boldsymbol{x} \tag{21}$$

3. **向量函数对向量的导数：**
   全微分关系式：

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) \Rightarrow \mathrm{d}\boldsymbol{y} = \begin{bmatrix} \mathrm{d}y_1 \\ \mathrm{d}y_2 \\ \vdots \\ \mathrm{d}y_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \\ \vdots \\ \mathrm{d}x_n \end{bmatrix} = \left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right]^\top \mathrm{d}\boldsymbol{x} \tag{22}$$

4. **标量函数对矩阵的导数：**

全微分关系式：

$$
\begin{aligned}
y = f(\boldsymbol{X}) \Rightarrow \mathrm{d}y &= \operatorname{tr}\left(\begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{21}} & \cdots & \frac{\partial f}{\partial X_{n1}} \\ \frac{\partial f}{\partial X_{12}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{n2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{1m}} & \frac{\partial f}{\partial X_{2m}} & \cdots & \frac{\partial f}{\partial X_{nm}} \end{bmatrix} \begin{bmatrix} \mathrm{d}X_{11} & \mathrm{d}X_{12} & \cdots & \mathrm{d}X_{1n} \\ \mathrm{d}X_{21} & \mathrm{d}X_{22} & \cdots & \mathrm{d}X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}X_{m1} & \mathrm{d}X_{m2} & \cdots & \mathrm{d}X_{mn} \end{bmatrix}\right) \\
&= \operatorname{tr}\left(\left[\frac{\partial f}{\partial \boldsymbol{X}}\right]^\top \mathrm{d}\boldsymbol{X}\right)
\end{aligned} \tag{23}
$$

5. **矩阵函数对矩阵的导数：**

首先定义矩阵向量化（按列优先）：

$$
\operatorname{vec}(\boldsymbol{X}) = [X_{11}, \ldots, X_{m1}, X_{12}, \ldots, X_{m2}, \ldots, X_{1n}, \ldots, X_{mn}]^\top_{mn\times 1} \tag{24}
$$

则对照向量函数对向量的导数，有全微分关系式：

$$
\operatorname{vec}(\boldsymbol{Y}) = \operatorname{vec}(\boldsymbol{F})[\operatorname{vec}(\boldsymbol{X})] \Rightarrow \operatorname{vec}(\mathrm{d}\boldsymbol{Y}) = \left[\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}}\right]^\top \operatorname{vec}(\mathrm{d}\boldsymbol{X}) \tag{25}
$$

注意按照上述定义，标量函数对矩阵的导数为 $mn \times 1$ 的矩阵，与前述标量函数对矩阵的导数形式不兼容，不能混用。

## 1.3 常用运算法则

- 微分运算法则：

$$
\mathrm{d}(X+Y) = \mathrm{d}X + \mathrm{d}Y,\ \mathrm{d}(XY) = (\mathrm{d}X)Y + X\mathrm{d}Y,\ \mathrm{d}\left(X^\top\right) = (\mathrm{d}X)^\top \tag{26}
$$

- 迹运算法则：

$$
\begin{gathered}
\mathrm{d}\operatorname{tr}(X) = \operatorname{tr}(\mathrm{d}X),\ a = \operatorname{tr}(a),\ \operatorname{tr}\left(X^\top\right) = \operatorname{tr}(X) \\
\operatorname{tr}(X \pm Y) = \operatorname{tr}(X) \pm \operatorname{tr}(Y),\ \operatorname{tr}(XY) = \operatorname{tr}(YX)
\end{gathered} \tag{27}
$$

最后一式中要求 $X$ 与 $Y^\top$ 维度相同。

- 向量化运算法则：

$$
\operatorname{vec}(X+Y) = \operatorname{vec}(X) + \operatorname{vec}(Y),\ \operatorname{vec}(AXB) = \left(B^T \otimes A\right)\operatorname{vec}(X) \tag{28}
$$

其中，$\otimes$ 表示 Kronecker 积。$A_{m\times n}$ 与 $B_{p\times q}$ 的 Kronecker 积为 $A \otimes B = \left[A_{ij}B\right]_{mp\times nq}$。

- Kronecker 积运算法则：

$$
(X \otimes Y)^\top = X^\top \otimes Y^\top \tag{29}
$$

## 1.4 复合函数求导

- Jacobi 矩阵转置形式：

$$\boldsymbol{y}=\boldsymbol{f}[\boldsymbol{h}(\boldsymbol{x})]\Rightarrow \mathrm{d}\boldsymbol{y}=\left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\right]^{\top}\mathrm{d}\boldsymbol{h}=\left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\right]^{\top}\left[\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\right]^{\top}\mathrm{d}\boldsymbol{x}=\left[\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\right]^{\top}\mathrm{d}\boldsymbol{x} \tag{30}$$

  所以有，

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}=\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}} \tag{31}$$

- Jacobi 矩阵形式：

$$\boldsymbol{y}=\boldsymbol{f}[\boldsymbol{h}(\boldsymbol{x})]\Rightarrow \mathrm{d}\boldsymbol{y}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\mathrm{d}\boldsymbol{h}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\mathrm{d}\boldsymbol{x} \tag{32}$$

  所以有，

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} \tag{33}$$

# 2 线性回归

## 2.1 最小二乘估计

**例 2.1** 最小二乘估计的理解：

最小二乘估计最简单的实例即为：二维平面坐标系里有多于两个的数据点，现需找寻一条直线使得所有点到直线的 $y$ 方向距离最近。这时需要把所有点相对直线的距离误差加起来，再对参数直线的两个参数进行求导，并令导数为零，求解出两个参数。

如果考虑到 $n$ 维空间，则最小二乘的数学化表达如下：

考虑 $\boldsymbol{X} \in \mathbb{R}^{m \times n}, \boldsymbol{y} \in \mathbb{R}^m$，其中矩阵 $\boldsymbol{X}$ 为列满秩。求 $\boldsymbol{\omega}_* \in \mathbb{R}^n$，满足

$$\left|\boldsymbol{X}\boldsymbol{\omega}_* - \boldsymbol{y}\right|_{\mathbb{R}^m} = \inf_{\boldsymbol{\omega} \in \mathbb{R}^n} \left|\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right| \tag{34}$$

## 2.2 微分学方法

构建矩阵计算式：

$$\left|\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right|^2_{\mathbb{R}^m} = \left(\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}, \boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right)_{\mathbb{R}^m} = \left(\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right)^\top \left(\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right) \tag{35}$$

根据上式的计算结果，然后对每一个 $\omega_i (i = 1, \ldots n)$ 进行求导，最终令这 $n$ 个导数式为 0，求解出 $\omega_*$。

## 2.3 矩阵求导与复合函数求导

定义 Loss 函数 $L(\boldsymbol{\omega}) = \left(\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right)^\top \left(\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}\right)$。通过矩阵求导与复合函数求导得出 $\frac{\partial L}{\partial \boldsymbol{\omega}}$。该式为标量函数对向量的导数，可以用标量函数对矩阵的导数进行求解（加迹运算）。

## 2.4 投影定理

考虑到

$$\boldsymbol{X}\boldsymbol{\omega} = \begin{bmatrix} X_1, & X_2, & \ldots, & X_n \end{bmatrix} \begin{bmatrix} \omega^1 \\ \omega^2 \\ \vdots \\ \omega^n \end{bmatrix} = \sum_{i=1}^{n} \omega^i X_i \in \mathbb{R}^m \tag{36}$$

根据投影定理，

$$\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\perp\operatorname{span}\{X_i\}_{i=1}^{n}\Longleftrightarrow\left(X_i,\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)_{\mathbb{R}^m}=0,\quad i=1,\ldots,n \tag{37}$$

由于，

$$\left(X_i,\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)_{\mathbb{R}^m}=X_i^\top\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)\Longleftrightarrow\begin{bmatrix}X_1^\top\\X_2^\top\\\vdots\\X_n^\top\end{bmatrix}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)=0 \tag{38}$$

其中，$\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)\in\mathbb{R}^n$。

即有，

$$\boldsymbol{X}^\top\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}_*\right)=0\Longrightarrow\boldsymbol{X}^\top\boldsymbol{y}=\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{\omega}_* \tag{39}$$

其中，$\boldsymbol{X}^\top\boldsymbol{X}\in\mathbb{R}^{n\times n}$，且对称正定。

最后，可求得，

$$\boldsymbol{\omega}_*=\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{y}\right) \tag{40}$$

所以，最小二乘的投影解法可看作是高维向量向低维空间的投影。图示如下：

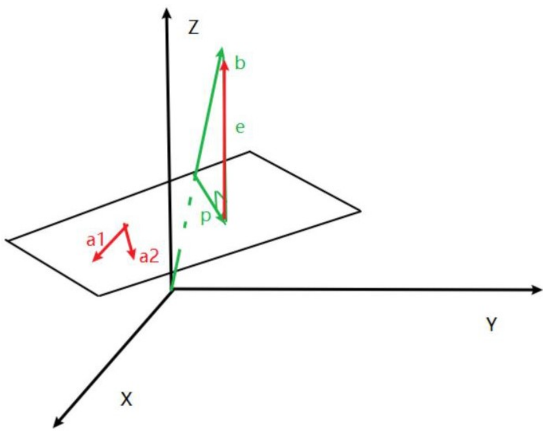


图 1: 最小二乘的投影理解

## 2.5 概率视角

当数据都在一条直线上时是最完美的情况，误差为 0。但现实中不可能出现这种情况，因为数据都带有一定的噪声。

假设噪声 $\epsilon \sim \mathcal{N}\left(0,\sigma^2\right)$，则有 $y_i = X_i\boldsymbol{\omega} + \epsilon$（$X_i$ 为 $\boldsymbol{X}$ 的行向量），得到 $y_i \mid X_i, \boldsymbol{\omega} \sim \mathcal{N}\left(X_i\boldsymbol{\omega}, \sigma^2\right)$，即满足

$$p(y_i \mid X_i, \boldsymbol{\omega}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - X_i\boldsymbol{\omega}\right)^2}{2\sigma^2}\right) \tag{41}$$

注意到，由于数据分布 $\boldsymbol{y}$ 独立同分布（IID），所以有

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}) = \prod_{i=1}^{n} p(y_i \mid X_i, \boldsymbol{\omega}) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - X_i\boldsymbol{\omega}\right)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^N} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})^\top \left(\sigma^{-2}\mathbf{I}\right)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})\right) \end{aligned} \tag{42}$$

对照附录章节B.6.2中的**多元 Gaussian 分布**公式(140)，可知 $\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\omega}, \sigma^2\mathbf{I}\right)$。

若用极大似然估计（参照附录章节B.5）来估计参数 $\boldsymbol{\omega}$，则可以令

$$\begin{aligned} L(\boldsymbol{\omega}) = \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega}) &= \log \prod_{i=1}^{n} p\left(y_i \mid X_i, \boldsymbol{\omega}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - X_i\boldsymbol{\omega}\right)^2}{2\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n}\left(\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{\left(y_i - X_i\boldsymbol{\omega}\right)^2}{2\sigma^2}\right) \end{aligned} \tag{43}$$

求得：

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} L(\boldsymbol{\omega}) = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - X_i\boldsymbol{\omega}\right)^2 = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\right\|^2 \tag{44}$$

以上说明，**最小二乘估计（LSE）$\Leftrightarrow$ Noise 为 Gaussian 的极大似然估计（MLE）**，即**最小二乘估计隐含了一个噪声服从正态分布的假设**。

进一步，若取先验分布 $\boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2\mathbf{I}\right)$，再通过 Bayes 公式（参照附录章节B.7）进行最大后验估计。首先可以计算得知：

$$\begin{aligned} p(\boldsymbol{\omega} \mid \text{data}) = p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) &= \frac{p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})}{p(\boldsymbol{X}, \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X})\, p(\boldsymbol{X} \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})}{p(\boldsymbol{y} \mid \boldsymbol{X})\, p(\boldsymbol{X})} \\ &= \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X})\, p(\boldsymbol{\omega} \mid \boldsymbol{X})}{p(\boldsymbol{y} \mid \boldsymbol{X})} \end{aligned} \tag{45}$$

或者

$$p(\boldsymbol{\omega} \mid \text{data}) = p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{\omega}, \boldsymbol{y} \mid \boldsymbol{X})}{p(\boldsymbol{y} \mid \boldsymbol{X})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X})\, p(\boldsymbol{\omega} \mid \boldsymbol{X})}{\int p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X})\, p(\boldsymbol{\omega} \mid \boldsymbol{X})\, \mathrm{d}\boldsymbol{\omega}} \tag{46}$$

注意以上相同的结果中，分母与参数 $\boldsymbol{\omega}$ 无关，且由于初始的 $\boldsymbol{\omega}$ 为先验产生，所以 $p(\boldsymbol{\omega} \mid \boldsymbol{X}) = p(\boldsymbol{\omega})$，接着可以根据类似极大似然估计的方法得到 $\boldsymbol{\omega}$ 的最大后验估计：

$$\begin{aligned}
\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} p(\boldsymbol{\omega} \mid \text{data}) = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X}) p(\boldsymbol{\omega})}{p(\boldsymbol{y} \mid \boldsymbol{X})} &= \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \log p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X}) p(\boldsymbol{\omega}) \\
&= \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \left(\log p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X}) + \log p(\boldsymbol{\omega})\right) \\
&= \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \left(\log\left(\frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\sigma_0}\right) - \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\right\|^2}{2\sigma^2} - \frac{\|\boldsymbol{\omega}\|^2}{2\sigma_0^2}\right) \\
&= \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \left(\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\right\|^2 + \frac{\sigma^2}{\sigma_0^2}\|\boldsymbol{\omega}\|^2\right)
\end{aligned} \tag{47}$$

观察上式结果，其与加了 L2 正则化（权重衰减）的 Loss Function 一致（防止过拟合，增加对参数的惩罚项。此种回归算法也被称为**岭（Ridge）回归**）：

$$L(\boldsymbol{\omega}) = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\right\|^2 + \lambda \|\boldsymbol{\omega}\|^2 \tag{48}$$

即**正则化（Generalized）的最小二乘估计（LSE）$\Leftrightarrow$ Noise 为 Gaussian 的 Bayes 最大后验估计（MAP）**。

注 MLE 为概率学派常用的参数估计方法，MAP 为贝叶斯学派常用的参数估计方法。MLE 的思想是通过数据得到参数，其完全依赖于数据，若数据过少而特征过多，则很容易出现过拟合。而 MAP 的思想是先给出一个预先估计，然后根据数据进行优化，这种情况下若先验较为靠谱则效果显著。若数据量大的情况下，MAP 与 MLE 将如出一辙。

## 2.6 Bayes 线性回归

1. 推断（Inference）：

   引入 Gaussian 先验：$p(\boldsymbol{\omega}) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_p\right)$。

   对参数的后验分布进行推断（与前述类似）：

$$p(\boldsymbol{\omega} \mid \text{data}) = p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}, \boldsymbol{X}) p(\boldsymbol{\omega} \mid \boldsymbol{X})}{p(\boldsymbol{y} \mid \boldsymbol{X})} \propto \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\omega}, \sigma^2 \mathbf{I}\right) \cdot \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_p\right) \tag{49}$$

   Gaussian 分布取 Gaussian 先验的共轭分布依然是 Gaussian 分布，于是可以得到后验分布也是一个 Gaussian 分布，有：

$$p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}) - \frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\omega}\right) \tag{50}$$

   假定最后得到的高斯分布为：$\mathcal{N}(\boldsymbol{\mu}_\omega, \boldsymbol{\Sigma}_\omega)$。对于上面的分布，采用配方法（对照**多元 Gaussian 分布**公式(140)）来得到最终的分布，指数上面 $\boldsymbol{\omega}$ 的二次项为：

$$-\frac{1}{2\sigma^2}\boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\omega} - \frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\omega} \tag{51}$$

于是有，

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-1} = \sigma^{-2}\boldsymbol{X}^{\top}\boldsymbol{X} + \boldsymbol{\Sigma}_p^{-1} = \boldsymbol{A} \tag{52}$$

再考虑 $\boldsymbol{\omega}$ 的一次项为：

$$\frac{1}{2\sigma^2}2\boldsymbol{y}^{\top}\boldsymbol{X}\boldsymbol{\omega} = \sigma^{-2}\boldsymbol{y}^{\top}\boldsymbol{X}\boldsymbol{\omega} \tag{53}$$

于是得到：

$$\boldsymbol{\mu}_{\omega}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-1} = \sigma^{-2}\boldsymbol{y}^{\top}\boldsymbol{X} \Rightarrow \boldsymbol{\mu}_{\omega} = \sigma^{-2}\boldsymbol{A}^{-1}\boldsymbol{X}^{\top}\boldsymbol{y} \tag{54}$$

2. 预测（Prediction）：
即给定一个 $\boldsymbol{X}^*$，求解 $\boldsymbol{y}^*$。由于 $f(\boldsymbol{X}^*) = \boldsymbol{X}^*\boldsymbol{\omega}$，代入参数后验，根据附录章节B.6.2中关于多元 Guassian 分布性质的定理B.13，有 $\boldsymbol{X}^*\boldsymbol{\omega} \sim \mathcal{N}\left(\boldsymbol{X}^*\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{X}^*\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\boldsymbol{X}^{*\top}\right)$，再添上噪声项 $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ 可最终得到：

$$\begin{aligned}
p\left(\boldsymbol{y}^* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}^*\right) &= \int_{\boldsymbol{\omega}} p\left(\boldsymbol{y}^* \mid \boldsymbol{\omega}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}^*\right) p\left(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}^*\right) \mathrm{d}\boldsymbol{\omega} \\
&= \int_{\boldsymbol{\omega}} p\left(\boldsymbol{y}^* \mid \boldsymbol{\omega}, \boldsymbol{X}^*\right) p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\omega} \\
&= \mathcal{N}\left(\boldsymbol{X}^*\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{X}^*\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\boldsymbol{X}^{*\top} + \sigma^2\mathbf{I}\right)
\end{aligned} \tag{55}$$

# 3 优化问题

## 3.1 梯度下降

梯度下降算法为下列格式：

$$x_{n+1} = x_n - \alpha \frac{\partial}{\partial x_n} L(x_n) \tag{56}$$

其中，$\alpha$ 为梯度下降的学习率。对上式进行多次迭代，$x_{n+1}$ 会趋于稳定值（从程序角度来说，是两次迭代的数值差距小于一个微小量 $\varepsilon$）。

接下来，将对上述整个过程进行数学上的证明。

梯度下降的本质是降低 $L(x)$ 函数值的大小，直到稳定值。接下来将从这两方面开始进行说明。第一，证明 $L(x)$ 函数值本身在下降；第二，函数值一定会下降到一个稳定的数值。

**证明.** 将上述第一方面用数学语言进行描述，即，$\exists$ 某一类条件，使得对于 $\forall$ 的 $x_{n+1}$ 和 $x_n$，

$$L(x_{n+1}) - L(x_n) < 0 \tag{57}$$

根据拉格朗日（Lagrange）中值定理，对于函数 $L(x)$，存在 $\xi \in [x, x+\Delta x]$，满足以下关系式，

$$L(x+\Delta x) - L(x) = L'(\xi) \cdot \Delta x, \quad \xi \in [x, x+\Delta x] \tag{58}$$

若 $\Delta x$ 为一微小量，则上式蜕化成泰勒（Taylor）一阶展开式，用 $x_{n+1}$ 和 $x_n$ 来表示，则有，

$$L(x_{n+1}) - L(x_n) = L'(x_n) \cdot (x_{n+1} - x_n), \quad |x_{n+1} - x_n| < \varepsilon,\ \varepsilon \to 0 \tag{59}$$

为保证 $L(x_{n+1}) - L(x_n) < 0$，即，

$$L'(x_n) \cdot (x_{n+1} - x_n) < 0 \tag{60}$$

由于 $L'(x_n)$ 为一固定表达式，无法更改；当且仅当，

$$x_{n+1} - x_n = -L'(x_n) \tag{61}$$

满足条件。（注意 = 条件，当且仅当 $L'(x_n)=0$）

考虑到 Taylor 一阶展开式的限制条件，这里需引入一微小量 $\alpha$，使得 $x_{n+1}$ 与 $x_n$ 的偏差不会太大。综合整理，得，

$$x_{n+1}=x_n-\alpha\frac{\partial}{\partial x_n}L(x_n) \tag{62}$$

由于以上步骤完全可逆，各部分上下均为充分必要条件，所以，第一部分证明完毕。即，证明了，当选取 $x_{n+1}=x_n-\alpha L'(x_n)$ 时，$L(x_{n+1})-L(x_n)<0$。

**证明.** 接下来证明第二部分，即 $x$ 会趋于稳定值。

第一部分的证明可以看出，梯度下降算法使得 $L(x_n)$ 一直下降。由于损失函数 $L(x)$ 具有单最小极值点，所以 $L(x_n)$ 会下降到函数的极值点处保持稳定，即当 $x_n\to\infty$ 时，$\forall\varepsilon\to 0$，使得 $|x_{n+1}-x_n|<\varepsilon$。

即，损失函数 $L(x_n)$ 会下降到一稳定数值，同时 $x_n$ 也会收敛于某一数值（从程序编写上来说，$x_n$ 不需要到 $\infty$，一般到 10 左右已经足够稳定）。证毕。

# 4 习题集

**例 4.1 (标量函数对向量求导问题)** $f=\left(\boldsymbol{x}^\top\boldsymbol{x}\right)^2$，其中 $\boldsymbol{x}=[2,1,3]^\top$ 为常数列向量，可知 $f$ 为标量函数。现根据矩阵求导术计算 $\frac{\partial f}{\partial \boldsymbol{x}}$。

**例 4.2 (标量函数对矩阵求导问题)** $f=\boldsymbol{a}\boldsymbol{X}\boldsymbol{b}^\top$，其中 $\boldsymbol{a}=[1,2,3]$ 为常数行向量，$\boldsymbol{b}=[2,1,3]$ 为常数行向量，$\boldsymbol{X}$ 是 $3\times3$ 的矩阵，形式如下：

$$\boldsymbol{X}=\begin{bmatrix}1&1&1\\1&2&1\\0&1&2\end{bmatrix}\tag{63}$$

可知 $f$ 为标量函数。现根据矩阵求导术计算 $\frac{\partial f}{\partial \boldsymbol{X}}$。

**例 4.3 (向量函数对向量求导问题)** $\boldsymbol{f}=\boldsymbol{x}^\top\boldsymbol{A}$，其中 $\boldsymbol{x}=[1,1,2]^\top$ 为常数列向量，$\boldsymbol{A}$ 是 $3\times3$ 的矩阵，形式如下：

$$\boldsymbol{A}=\begin{bmatrix}1&1&1\\0&1&2\\1&2&1\end{bmatrix}\tag{64}$$

可知 $\boldsymbol{f}$ 为行向量函数。现根据矩阵求导术计算 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$。

**例 4.4 (线性回归问题)** $l=\|\boldsymbol{X}\boldsymbol{\omega}-\boldsymbol{y}\|^2$，求 $\boldsymbol{\omega}$ 的最小二乘估计，即求 $\frac{\partial l}{\partial \boldsymbol{\omega}}$ 的零点。其中，

$$\boldsymbol{y}=\begin{bmatrix}1\\2\\3\end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix}1&2\\4&0\\8&1\end{bmatrix},\quad \boldsymbol{\omega}=\begin{bmatrix}\omega_1\\\omega_2\end{bmatrix}\in\mathbb{R}^{2\times1}\tag{65}$$

解法一般有以下四种：

1. 将矩阵展开成单独的表达式，把每一个单独表达式的最小二乘误差加起来，再令总误差对每一个参数进行求导，最终找出极值点，求解出每个参数数值；
2. 利用矩阵求导术求解出极值点的通项式，再令其为零向量，求解出参数数值；
3. 利用矩阵求导术的链式求导法则，建立计算流程求解；
4. 利用最小二乘法的几何意义，根据投影理论建立矩阵所满足的表达式进行求解。

请选用以上方法的起码 3 种，写明详细的求解过程，和最终结果。

**例 4.5 (*Bayes 线性回归问题)** 已知一元 Guassian 分布的概率分布密度函数为：

$$f(x)=\mathcal{N}(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\tag{66}$$

即随机变量 $X$ 服从均值为 $\mu$，方差为 $\sigma^2$ 的 Gaussian 分布。

多元 Guassian 分布的概率分布密度函数的向量值写法为（其中 $\boldsymbol{x}=[x_1,x_2,x_3,\cdots]^\top$）：

$$f(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{u})\right) \tag{67}$$

即随机变量 $\boldsymbol{X}$ 服从均值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$ 的多元 Gaussian 分布。

且满足：$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}$ 也服从多元 Gaussian 分布（$\boldsymbol{A}$ 为系数矩阵，$\boldsymbol{b}$ 为系数列向量），参数为 $(\boldsymbol{A}\boldsymbol{\mu},\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$。

**(1)** 问题一：对于前述线性回归问题，假设存在噪声 $\epsilon\sim\mathcal{N}\left(0,\sigma^2\right)$，则有 $y_i=X_i\boldsymbol{\omega}+\epsilon$（$X_i$ 为 $\boldsymbol{X}$ 的行向量），得到 $y_i\mid X_i,\boldsymbol{\omega}\sim\mathcal{N}\left(X_i\boldsymbol{\omega},\sigma^2\right)$，试证明：$\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\omega}\sim\mathcal{N}\left(\boldsymbol{X}\boldsymbol{\omega},\sigma^2\mathbf{I}\right)$。
提示：数据分布 $\boldsymbol{y}$ 独立同分布。

**(2)** 问题二：若用极大似然估计来估计参数 $\boldsymbol{\omega}$，即可以令

$$L(\boldsymbol{\omega})=\ln p(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\omega})=\log\prod_{i=1}^{n}p\left(y_i\mid X_i,\boldsymbol{\omega}\right) \tag{68}$$

试证明:此时 $\hat{\boldsymbol{\omega}}=\underset{\boldsymbol{\omega}}{\operatorname{argmax}}L(\boldsymbol{\omega})$ 的最终表达式,与线性回归的表达式 $\underset{\boldsymbol{\omega}}{\operatorname{argmin}}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\right\|^2$ 完全一致。

**(3)** 问题三：进一步，若取先验分布 $\boldsymbol{\omega}\sim\mathcal{N}\left(\mathbf{0},\sigma_0^2\mathbf{I}\right)$，再通过 Bayes 公式进行最大后验估计，即通过 Bayes 公式计算：$p(\boldsymbol{\omega}\mid\mathrm{data})=p(\boldsymbol{\omega}\mid\boldsymbol{X},\boldsymbol{y})$。
试证明：此时极大似然估计 $\hat{\boldsymbol{\omega}}=\underset{\boldsymbol{\omega}}{\operatorname{argmax}}\left(\ln p(\boldsymbol{\omega}\mid\mathrm{data})\right)$ 的最终表达式，与如下加了 L2 正则化（权重衰减）的线性回归表达式一致（防止过拟合，增加对参数的惩罚项。此种回归算法也被称为岭（Ridge）回归）：

$$\hat{\boldsymbol{\omega}}=\underset{\boldsymbol{\omega}}{\operatorname{argmin}}\left(\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\right\|^2+\lambda\|\boldsymbol{\omega}\|^2\right) \tag{69}$$

**(4)** 问题四：将如上 Gaussian 先验一般化，令 $p(\boldsymbol{\omega})\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}_p\right)$。试通过配方法证明：此时后验分布 $p(\boldsymbol{\omega}\mid\mathrm{data})=p(\boldsymbol{\omega}\mid\boldsymbol{X},\boldsymbol{y})$ 也服从 Gaussian 分布 $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}},\boldsymbol{\Sigma}_{\boldsymbol{\omega}})$，且满足：

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-1}&=\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{\Sigma}_p^{-1}=\boldsymbol{A}\\ \boldsymbol{\mu}_\omega&=\sigma^{-2}\boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{y}\end{aligned} \tag{70}$$

**(5)** 问题五：在反复后验得到最准确的模型参数 $\boldsymbol{\omega}$ 后，给定一个 $\boldsymbol{X}^*$，预测 $\boldsymbol{y}^*$。注意需添加噪声项 $\epsilon\sim\mathcal{N}\left(0,\sigma^2\right)$。试证明：此时 $p\left(\boldsymbol{y}^*\mid\boldsymbol{X},\boldsymbol{y},\boldsymbol{X}^*\right)\sim\mathcal{N}\left(\boldsymbol{X}^*\boldsymbol{\mu}_{\boldsymbol{\omega}},\boldsymbol{X}^*\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\boldsymbol{X}^{*\top}+\sigma^2\mathbf{I}\right)$。

# A 微积分相关

## A.1 特殊函数

**定义 A.1 (Γ 函数的定义及其性质)** 在区间 $(0,\infty)$ 上，Γ 函数被定义为

$$\Gamma(x)=\int_0^{\infty}\mathrm{e}^{-t}t^{x-1}\mathrm{d}t \tag{71}$$

其满足如下基本性质

$$\Gamma(x+1)=x\Gamma(x) \tag{72}$$

**证明.** 根据 Γ 函数的定义，其性质可通过分部积分法直接证明：

$$\begin{aligned}\Gamma(x+1)&=\int_0^{\infty}\mathrm{e}^{-t}t^{x}\mathrm{d}t=-\int_0^{\infty}t^{x}\mathrm{d}\left(\mathrm{e}^{-t}\right)\\&=-\,t^{x}\mathrm{e}^{-t}\big|_{t=0}^{\infty}+x\int_0^{\infty}\mathrm{e}^{-t}t^{x-1}\mathrm{d}t=x\Gamma(x)\end{aligned} \tag{73}$$

进一步，根据

$$\Gamma(1)=\int_0^{\infty}\mathrm{e}^{-t}\mathrm{d}t=1 \tag{74}$$

可得到重要的递推公式：

$$\Gamma(n+1)=n!,\quad n=0,1,2,\cdots \tag{75}$$

所以，Γ 函数可看作**阶乘运算**在实数或复数域的拓展。当定义域取正整数时，$\Gamma(n+1)$ 即为阶乘 $n!$。

**注** 需要指出，Γ 函数在区间 $(-\infty,0]$ 上没有定义（积分不收敛）。

**例 A.1** 求解 $(1/2)!$，即 $\Gamma(3/2)$。

$$(1/2)!=\Gamma(3/2)=\int_0^{\infty}\mathrm{e}^{-t}\sqrt{t}\mathrm{d}t=\int_0^{\infty}t\mathrm{e}^{-t^2}\mathrm{d}t^2=2\int_0^{\infty}t^2\mathrm{e}^{-t^2}\mathrm{d}t \tag{76}$$

通过分部积分（$u=t,\mathrm{d}v=t\mathrm{e}^{-t^2}\mathrm{d}t,v=-\frac{1}{2}\mathrm{e}^{-t^2}$）继续计算：

$$(1/2)!=\Gamma(3/2)=-\,t\mathrm{e}^{-t^2}\Big|_0^{\infty}+\int_0^{\infty}\mathrm{e}^{-t^2}\mathrm{d}t=\int_0^{\infty}\mathrm{e}^{-t^2}\mathrm{d}t \tag{77}$$

构造

$$I=\int_{-\infty}^{\infty}\mathrm{e}^{-x^2}\mathrm{d}x \tag{78}$$

则有

$$I^2=\int_{-\infty}^{\infty}\mathrm{e}^{-y^2}\mathrm{d}y\int_{-\infty}^{\infty}\mathrm{e}^{-x^2}\mathrm{d}x=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{-(y^2+x^2)}\mathrm{d}y\mathrm{d}x \tag{79}$$

利用极坐标变换对上述二重积分进行求解（令 $x=r\cos\theta, y=r\sin\theta, \mathrm{d}y\mathrm{d}x=r\mathrm{d}\theta\mathrm{d}r$）：

$$I^2=\int_0^{\infty}\int_0^{2\pi}\mathrm{e}^{-r^2}r\mathrm{d}\theta\mathrm{d}r=2\pi\int_0^{\infty}r\mathrm{e}^{-r^2}\mathrm{d}r=-\pi\mathrm{e}^{-u}\big|_0^{\infty}=\pi \tag{80}$$

考虑到公式(78)的积分必然大于零，所以有 $I=\sqrt{\pi}$。

从公式(77)继续计算，可得：

$$(1/2)!=\Gamma(3/2)=\int_0^{\infty}\mathrm{e}^{-t^2}\mathrm{d}t=\frac{1}{2}\int_{-\infty}^{\infty}\mathrm{e}^{-t^2}\mathrm{d}t=\frac{\sqrt{\pi}}{2} \tag{81}$$

# B 概率论相关

符号体系：

$X$ 为事件，当其存在不同结果时，被定义为随机变量。例如抛三个硬币，$X$ 表示正面朝上的个数。$x$ 为 $X$ 的取值，例如 $X$ 的可能值为 $x_1, x_2, \cdots$。

$P(X)$ 表示事件 $X$ 的概率，$p$ 则为 $P\{X=x\}$ 中某个具体的概率数值。

## B.1 离散型随机变量

**定义 B.1 (离散型随机变量)** 若一个随机变量具有可数多个可能取值，则称这个随机变量为离散型的。对于一个离散型随机变量 $X$，定义 $X$ 的概率分布列 $p(x)$ 为

$$p(x) = P\{X = x\} \tag{82}$$

**定义 B.2 (期望)** 期望刻画随机变量所有可能取值的加权平均：

$$\mu = E[X] = \sum_{x:p(x)>0} x p(x) = \sum_{i=1}^{n} x_i p(x_i) \tag{83}$$

**命题 B.1 (随机变量函数的期望)** 如果 $X$ 是一个离散型随机变量，其可能取值为 $x_i$, $i \geq 1$，相应的取值概率为 $p(x_i)$，那么，对任一实值函数 $g$，都有

$$E\left[g(X)\right] = \sum_i g(x_i) p(x_i) \tag{84}$$

**证明.** 将和号 $\sum\limits_i g(x_i)p(x_i)$ 中具有相同 $g(x_i)$ 数值的项合并。假设 $y_j (j \geq 1)$ 表示 $g(x_i) (i \geq 1)$ 的不同取值，则有

$$\begin{aligned}
\sum_i g(x_i)p(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p(x_i) = \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) = \sum_j y_j \sum_{i:g(x_i)=y_j} p(x_i) \\
&= \sum_j y_j P\left\{g(X) = y_j\right\} = \sum_j y_j p(y_j) = E\left[g(X)\right]
\end{aligned} \tag{85}$$

随机变量 $X$ 的期望 $E[X]$，也称为 $X$ 的均值或**一阶矩**。$E[X^n] (n \geq 1)$ 称为 $X$ 的 $n$ 阶矩。根据命题B.1可知：

$$E[X^n] = \sum_{x:p(x)>0} x^n p(x) \tag{86}$$

**定义 B.3 (方差)** 方差刻画随机变量的取值相对于均值的偏离程度：

$$\sigma^2 = \mathrm{Var}[X] = E\left[(X-\mu)^2\right] = E\left[X^2\right] - (E[X])^2 \tag{87}$$

**证明.** 结合命题B.1，$\mathrm{Var}[X]=E\left[X^2\right]-(E[X])^2$ 的具体证明过程如下：

$$\begin{aligned}\mathrm{Var}[X]&=E\left[(X-\mu)^2\right]=\sum_x(x-\mu)^2p(x)=\sum_x\left(x^2-2\mu x+\mu^2\right)p(x)\\&=\sum_x x^2p(x)-2\mu\sum_x xp(x)+\mu^2\sum_x p(x)=E\left[X^2\right]-\mu^2\end{aligned}\tag{88}$$

### B.1.1 伯努利（Bernoulli）随机变量

**定义 B.4 (伯努利（Bernoulli）随机变量)** 一项试验，要么成功要么失败。每次成功的概率为 $p$，则有

$$\begin{aligned}p(0)&=p\{X=0\}=1-p\\p(1)&=p\{X=1\}=p\end{aligned}\tag{89}$$

则称 $X$ 为伯努利（Bernoulli）随机变量。

Bernoulli 随机变量的期望：$E[X]=p$；方差：$\mathrm{Var}[X]=p(1-p)$。

### B.1.2 二项随机变量

**定义 B.5 (二项随机变量)** 假设将 Bernoulli 随机变量重复进行 $n$ 次，则称为参数是 $(n,p)$ 的二项随机变量，可记为 $X\sim\mathcal{B}(n,p)$，满足

$$p\{X=k\}=\mathrm{C}_n^k p^k(1-p)^{n-k}=\binom{n}{k}p^k(1-p)^{n-k}\tag{90}$$

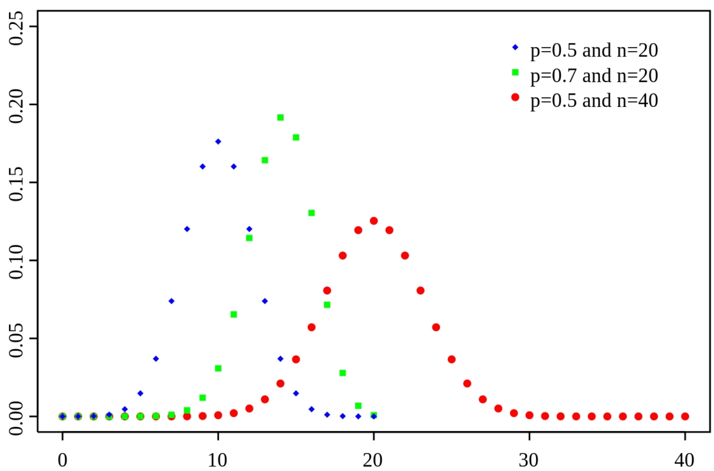


图 2: 不同参数下的二项随机变量概率分布

二项随机变量的期望：$E[X]=np$；方差：$\mathrm{Var}[X]=np(1-p)$。

### B.1.3 泊松（Poisson）随机变量

**定义 B.6 (泊松（Poisson）随机变量)** 当二项随机变量的 $n$ 很大而 $p$ 很小时，泊松（Poisson）随机变量可作为二项随机变量的近似，其中 $\lambda$ 为 $np$：

$$p\{X=k\}=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\quad k=0,1,2,\cdots \tag{91}$$

**证明.** 根据二项随机变量的概率公式(90)，有

$$p\{X=k\}=\mathrm{C}_n^k\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{92}$$

注意到当 $n\to\infty$ 取极限时，有

$$\frac{\mathrm{C}_n^k}{n^k}\to\frac{1}{k!},\quad \left(1-\frac{\lambda}{n}\right)^n\to\mathrm{e}^{-\lambda} \tag{93}$$

所以公式(91)得证。

Poisson 随机变量的**期望**和**方差**均为 $\lambda$。

Poisson 随机变量多出现在当 $X$ 表示在一定的时间或空间内出现的事件个数这种场合。在一定时间内某交通路口所发生的事故个数，是一个典型的例子。

设所观察的时间段为 $[0,1)$，取一个很大的自然数 $n$，把 $[0,1)$ 分为等长的 $n$ 段：

$$l_1=\left[0,\frac{1}{n}\right],l_2=\left[\frac{1}{n},\frac{2}{n}\right],\cdots,l_i=\left[\frac{i-1}{n},\frac{i}{n}\right],\cdots,l_n=\left[\frac{n-1}{n},1\right] \tag{94}$$

在每段 $l_i$ 内，恰发生一个事故的概率，近似的与这段时间的长 $\frac{1}{n}$ 成正比，可设为 $\frac{\lambda}{n}$，且各段是否发生事故是独立的，则 $X$ 应服从二项随机变量 $\mathcal{B}\left(n,\frac{\lambda}{n}\right)$。此时的情形可用 Poisson 分布进行计算。

## B.2 连续型随机变量

**定义 B.7** 对于随机变量 $X$，若存在一个非负的可积函数 $f(x)$，使得对任意实数 $x$，有：

$$P\{X<x\}=P\{X\le x\}=F(x)=\int_{-\infty}^{x}f(x)\mathrm{d}x \tag{95}$$

则称 $X$ 为连续型随机变量。其中 $f(x)$ 为 $X$ 的概率分布密度函数。

从 $-\infty$ 到 $\infty$，上述积分为总概率 1。无法计算单个 $x$ 的概率，只能计算区间概率。

**注** 当提到一个随机变量 $X$ 的概率分布，指的是它的分布函数。当 $X$ 是连续型时，指的是他的概率密度；当 $X$ 是离散型时，指的是它的分布列规律。

**定义 B.8 (连续型随机变量的期望)** 可用与离散型随机变量类似的方法进行定义：

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \tag{96}$$

**命题 B.2 (连续型随机变量函数的期望)** 与离散型随机变量函数的期望类似：

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x \tag{97}$$

**定义 B.9 (连续型随机变量的方差)** 连续型随机变量的方差与离散型随机变量完全一致：

$$\sigma^2 = \mathrm{Var}[X] = E\left[(X-\mu)^2\right] = E\left[X^2\right] - (E[X])^2 \tag{98}$$

### B.2.1 高斯（Gaussian）分布

**定义 B.10 (高斯（Gaussian）分布/正态分布)** 若存在如下概率分布密度函数：

$$f(x) = \mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{99}$$

则称随机变量 $X$ 服从均值为 $\mu$，方差为 $\sigma^2$ 的 Gaussian 分布，记为 $X \sim \mathcal{N}(\mu,\sigma^2)$。

公式(99)的证明详见附录章节B.6.1。

下面进一步证明：如果 $X$ 是一个服从参数为 $(\mu,\sigma^2)$ 的 Gaussian 分布的随机变量，那么 $Y = aX + b$ 也服从 Gaussian 分布，且参数为 $(a\mu + b, a^2\sigma^2)$。

**证明.** 假设 $a > 0$（$a < 0$ 时的证明类似），设 $F_Y$ 为 $Y$ 的分布函数，则有

$$F_Y(x) = P\{Y \leqslant x\} = P\{aX + b \leqslant x\} = P\left\{X \leqslant \frac{x-b}{a}\right\} = F_X\left(\frac{x-b}{a}\right) \tag{100}$$

其中 $F_X$ 为 $X$ 的分布函数。求导可得 $Y$ 的密度函数为

$$\begin{aligned} f_Y(x) &= \frac{1}{a} f_X\left(\frac{x-b}{a}\right) = \frac{1}{\sqrt{2\pi}a\sigma} \exp\left\{-\left(\frac{x-b}{a} - \mu\right)^2 / 2\sigma^2\right\} \\ &= \frac{1}{\sqrt{2\pi}a\sigma} \exp\left\{-(x-b-a\mu)^2 / 2(a\sigma)^2\right\} \end{aligned} \tag{101}$$

以上证明说明：如果 $X$ 是一个参数为 $(\mu,\sigma^2)$ 的 Gaussian 随机变量，则 $Z = (X-\mu)/\sigma$ 是一个参数为 $(0,1)$ 的 Gaussian 随机变量，即**标准正态随机变量**；反之亦然，即可通过 $X = \sigma Z + \mu$ 将 $Z \sim \mathcal{N}(0,1)$ 转变为 $X \sim \mathcal{N}(\mu,\sigma^2)$。

一般将标准正态随机变量的分布函数记为$\Phi(x)$。

下面再证明：一个参数为$(\mu,\sigma^2)$的 Gaussian 随机变量，$E(X)=\mu$，$\mathrm{Var}(X)=\sigma^2$。

**证明.** 先计算标准正态随机变量 $Z=(X-\mu)/\sigma$ 的期望和方差。由于

$$E[Z]=\int_{-\infty}^{\infty} x f_Z(x)\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\mathrm{e}^{-x^2/2}\mathrm{d}x=-\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\Big|_{-\infty}^{\infty}=0 \tag{102}$$

因此，

$$\mathrm{Var}(Z)=E\left[Z^2\right]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2\mathrm{e}^{-x^2/2}\mathrm{d}x \tag{103}$$

通过分部积分（$u=x,\mathrm{d}v=x\mathrm{e}^{-x^2/2}\mathrm{d}x, v=\mathrm{e}^{-x^2/2}$）得到：

$$\mathrm{Var}(Z)=\frac{1}{\sqrt{2\pi}}\left(-x\mathrm{e}^{-x^2/2}\Big|_{-\infty}^{\infty}+\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2}\mathrm{d}x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2}\mathrm{d}x=1 \tag{104}$$

上式最后一步结果的得出可参考附录章节A.1中例A.1的计算过程。

由 $X=\mu+\sigma Z$ 得到

$$E[X]=\mu+\sigma E[Z]=\mu \tag{105}$$

从而

$$\mathrm{Var}(X)=\sigma^2\mathrm{Var}(Z)=\sigma^2 \tag{106}$$

## B.3 随机变量的联合分布

**定义 B.11** 若是离散型随机变量，则可以定义 $X$ 和 $Y$ 的联合概率分布列：

$$p(x,y)=P\{X=x,Y=y\} \tag{107}$$

若是连续型随机变量，则可以定义 $X$ 和 $Y$ 的联合概率分布函数：

$$F(a,b)=P\{X\leqslant a,Y\leqslant b\}=\int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\mathrm{d}x\mathrm{d}y,\quad -\infty<a,b<\infty \tag{108}$$

其中，函数 $f(x,y)$ 为 $X$ 和 $Y$ 的联合密度函数。

### B.3.1 联合分布的期望与协方差

**命题 B.3 (多元随机变量函数的期望)** 如果 $X_1,\cdots,X_n$ 服从多元分布列 $p(x_1,\cdots,x_n)$，则有：

$$E\left[g(X_1,\cdots,X_n)\right]=\sum_{x_1}\cdots\sum_{x_n}g(x_1,\cdots,x_n)p(x_1,\cdots,x_n) \tag{109}$$

如果 $X_1,\cdots,X_n$ 具有联合分布密度 $f(x_1,\cdots,x_n)$，则有：

$$E\left[g(X_1,\cdots,X_n)\right]=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(x_1,\cdots,x_n)f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n \tag{110}$$

以下命题B.4和B.5即为在联合分布情况下关于期望性质的扩展。

**命题 B.4 (随机变量和的期望)** 对于随机变量 $X_1,X_2,\cdots,X_n$，有

$$E\left[\sum_{i=1}^{n}X_i\right]=\sum_{i=1}^{n}E[X_i] \tag{111}$$

**证明.** 以离散型随机变量为例，令 $X_i(x)$ 表示不同随机变量 $X_i$ 的取值，可直接计算：

$$\begin{aligned}E\left[\sum_{i=1}^{n}X_i\right]&=\sum_{x}[X_1(x)+X_2(x)+\cdots+X_n(x)]p(x)=\sum_{i=1}^{n}\left[\sum_{x}X_i(x)p(X_i(x))\right]\\&=E[X_1]+E[X_2]+\cdots+E[X_n]\end{aligned} \tag{112}$$

**注** 二项随机变量的期望和方差即可根据命题B.4对 Bernoulli 随机变量进行求和得到。

**命题 B.5 (随机变量乘积的期望)** 对于相互独立的随机变量 $X_1,X_2,\cdots,X_n$，有

$$E\left[\prod_{i=1}^{n}X_i\right]=\prod_{i=1}^{n}E[X_i] \tag{113}$$

**证明.** 以离散型随机变量为例，令 $X_i(x)$ 表示不同随机变量 $X_i$ 的取值，可直接计算：

$$\begin{aligned}E\left[\prod_{i=1}^{n}X_i\right]&=\sum_{x}[X_1(x)\cdot X_2(x)\cdots X_n(x)]p(X_1(x)X_2(x)\cdots X_n(x))\\&=\prod_{i=1}^{n}\sum_{x}X_i(x)p(X_i(x))=E[X_1]\cdot E[X_2]\cdots E[X_n]\end{aligned} \tag{114}$$

**定义 B.12 (协方差)** 协方差刻画两个变量变化的相关性：

$$\sigma_X\sigma_Y=\mathrm{Cov}(X,Y)=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y] \tag{115}$$

根据命题B.5，若 $X$ 和 $Y$ 相互独立，则协方差 $\sigma_X\sigma_Y=0$。

**定义 B.13 (协方差矩阵)** 协方差矩阵刻画多个变量相关性的统一表达。假设随机变量为 $X_1, X_2, X_3, \cdots$，期望分别为 $\mu_1, \mu_2, \mu_3, \cdots$，方差分别为 $\sigma_1^2, \sigma_2^2, \sigma_3^2, \cdots$，则有：

$$
\begin{aligned}
\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^\top &= \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_n \end{bmatrix} \begin{bmatrix} \sigma_1 & \sigma_2 & \cdots & \sigma_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_n \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n\sigma_1 & \sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix} \\
&= \begin{bmatrix} E\left[(X_1-\mu_1)^2\right] & E\left[(X_1-\mu_1)(X_2-\mu_2)\right] & \cdots & E\left[(X_1-\mu_1)(X_n-\mu_n)\right] \\ E\left[(X_2-\mu_2)(X_1-\mu_1)\right] & E\left[(X_2-\mu_2)^2\right] & \cdots & E\left[(X_2-\mu_2)(X_n-\mu_n)\right] \\ \vdots & \vdots & \ddots & \vdots \\ E\left[(X_n-\mu_n)(X_1-\mu_1)\right] & E\left[(X_n-\mu_n)(X_2-\mu_2)\right] & \cdots & E\left[(X_n-\mu_n)^2\right] \end{bmatrix} \\
&= E\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top\right]
\end{aligned} \tag{116}
$$

### B.3.2 独立随机变量的联合分布

**命题 B.6 (独立随机变量的和)** 对于离散型独立随机变量 $X$ 和 $Y$，$X+Y$ 的分布列为：

$$
P\{X+Y=n\} = \sum_{k=0}^{n} P\{X=k, Y=n-k\} = \sum_{k=0}^{n} P\{X=k\}P\{Y=n-k\} \tag{117}
$$

对于连续型独立随机变量 $X$ 和 $Y$，$X+Y$ 的累积分布函数为：

$$
\begin{aligned}
F_{X+Y}(a) = P\{X+Y \leqslant a\} &= \iint_{x+y\leqslant a} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)\mathrm{d}x f_Y(y)\mathrm{d}y = \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\mathrm{d}y
\end{aligned} \tag{118}
$$

分布函数 $F_{X+Y}$ 称为分布函数 $F_X$ 和 $F_Y$（分别表示 $X$ 和 $Y$ 的分布函数）的卷积。

对上式(118)求导，可得 $X+Y$ 的密度函数：

$$
\begin{aligned}
f_{X+Y}(a) &= \frac{\mathrm{d}}{\mathrm{d}a}\int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\mathrm{d}y = \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}a}F_X(a-y)f_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)\mathrm{d}y
\end{aligned} \tag{119}
$$

**命题 B.7 (独立二项随机变量的和（卷积）)** 如果 $X \sim \mathcal{B}(n,p)$ 和 $Y \sim \mathcal{B}(m,p)$，且 $X$ 和 $Y$ 相互独立，那么 $X+Y$ 也服从二项分布，且满足 $X+Y \sim \mathcal{B}(n+m,p)$。

**命题 B.8 (独立正态随机变量的和（卷积）)** 如果 $X_i (i=1,2,\cdots,n)$ 是 $n$ 个相互独立的随机变量，且分别服从参数为 $(\mu_i, \sigma_i^2)$ 的正态 Gaussian 分布，则 $\sum\limits_{i=1}^{n} X_i$ 也服从正态 Gaussian 分布，且参数为 $(\sum\limits_{i=1}^{n}\mu_i, \sum\limits_{i=1}^{n}\sigma_i^2)$。

以上命题B.7和B.8均可通过命题B.6进行计算证明。

## B.4 极限定理

### B.4.1 大数定律

**定理 B.9 (弱大数定律)** 设 $X_1, X_2, \cdots$ 为独立同分布的随机变量序列，其公共期望 $E[X_i] = \mu$ 有限，则对任何 $\epsilon > 0$，有

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty \tag{120}$$

**定理 B.10 (强大数定律)** 设 $X_1, X_2, \cdots$ 为独立同分布的随机变量序列，其公共期望 $E[X_i] = \mu$ 有限，则下式以概率 1 成立：

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu, \quad n \to \infty \tag{121}$$

即也可表示为：

$$P\left\{\lim_{n\to\infty} (X_1 + \cdots + X_n)/n = \mu\right\} = 1 \tag{122}$$

大数定律表明：**独立同分布随机变量序列的均值以概率 1 收敛到分布的均值**。简单来说，即**在试验不变的条件下，重复试验多次，随机事件的频率近似于它的概率**。

### B.4.2 中心极限定理

**定理 B.11 (中心极限定理)** 设 $X_1, X_2, \cdots$ 为独立同分布的随机变量序列，其公共分布的均值为 $\mu$，方差为 $\sigma^2$，则随机变量

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \tag{123}$$

的分布当 $n \to \infty$ 时趋向于标准正态分布。即对任何 $-\infty < a < \infty$，

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \mathrm{e}^{-x^2/2} \mathrm{d}x, \quad n \to \infty \tag{124}$$

中心极限定理表明：**大量独立随机变量的和近似服从正态分布**。

**注** 对于二项分布的随机变量，如果 $n$ 足够大，那么分布的偏度就比较小。这种情况下，如果使用适当的连续性校正，那么 $\mathcal{B}(n, p)$ 的一个很好的近似是 Gaussian 分布 $\mathcal{N}(np, np(1-p))$。该结论即为棣莫弗-拉普拉斯（De Moivre - Laplace）极限定理（上述中心极限定理的一个特殊情形）。

**大数定律**和**中心极限定理**的证明需要借助**马尔可夫（Markov）不等式**（只知道分布的均值，导出概率上界）和**切比雪夫（Chebyshev）不等式**（只知道分布的均值和方差，导出概率上界）。

## B.5 极大似然估计（MLE）

用于根据**观测数据**和假设的**概率分布**推断最可能得**模型参数值**。

1. 根据假设的数据分布（如 Gaussian 分布）写出似然函数（数据已知，评估参数）：

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} p(x_i|\theta) \tag{125}$$

其中，$\theta$ 为模型参数，如 Gaussian 分布中的 $(\mu, \sigma)$。

2. 对似然函数取对数，并整理：

$$H(\theta) = \ln L(\theta) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} p(x_i|\theta) \tag{126}$$

3. 求导数，令其为 0，得到似然方程（需确保其二阶导数 $< 0$）：

$$\hat{\theta} = \operatorname{argmax} H(\theta) \tag{127}$$

4. 解方程，得到概率模型的参数估计。

## B.6 Gaussian 分布的证明

### B.6.1 一元 Gaussian 分布

一元 Gaussian 分布公式的证明方法较多，比较直接的还是 Gauss 自己的推导：从 $n$ 个观测值 $(x_1, x_2, \cdots, x_n)$ 中根据极大似然估计来估算真实的数值。

令第 $i (1 \le i \le n)$ 次测量误差为 $e_i = x_i - \theta$。假设随机观测误差的概率密度函数为 $f(e)$，则似然函数为误差的联合概率密度函数，如下：

$$L(\theta) = \prod_{i=1}^{n} f(e_i) = \prod_{i=1}^{n} f(x_i - \theta) \tag{128}$$

对上式取对数，求导后令其为 0，则可得到似然方程：

$$\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)} = 0 \tag{129}$$

记 $g(x)=\frac{f'(x)}{f(x)}$，可简化为：

$$\sum_{i=1}^{n} g(x_i-\theta)=0 \tag{130}$$

$n$ 次独立试验后，估计的真值应当趋近于观测值的算术平均数：

$$\hat{\theta}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \tag{131}$$

然而，想要结合公式(130)和(131)给出 $g(x)$ 的具体表达式并不现实。由于这里的观测值是任意的，可以先随意构造一种简化的样本，如：

$$x_n=nx,\quad x_1=x_2=\cdots=x_{n-1}=0,\quad -\infty<x<\infty \tag{132}$$

此时有 $\hat{\theta}=x$，代回公式(130)得到 $(n-1)g(-x)+g((n-1)x)=0$。由于当 $n=2$ 时，$g(-x)=-g(x)$，说明 $g(x)$ 为奇函数。以上整理可知：

$$mg(x)=g(mx),\quad -\infty<x<\infty,\quad m=0,1,2,\cdots \tag{133}$$

由于上式恒成立，唯一满足此式的连续函数即为 $g(x)=ax$（$a$ 为常数），则有

$$f(x)=axf'(x)\rightarrow f(x)=C\mathrm{e}^{ax^2},\quad C,a\text{为常数} \tag{134}$$

对上式进行正则化（概率密度函数 $f(x)$ 的积分为 1），令

$$I=\int_{-\infty}^{\infty} C\mathrm{e}^{ay^2}\mathrm{d}y=1 \tag{135}$$

则有

$$I^2=C^2\int_{-\infty}^{\infty}\mathrm{e}^{ay^2}\mathrm{d}y\int_{-\infty}^{\infty}\mathrm{e}^{ax^2}\mathrm{d}x=C^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{a^2(y^2+x^2)}\mathrm{d}y\mathrm{d}x \tag{136}$$

利用极坐标变换对上述二重积分进行求解（令 $x=r\cos\theta, y=r\sin\theta, \mathrm{d}y\mathrm{d}x=r\mathrm{d}\theta\mathrm{d}r$）：

$$I^2=C^2\int_0^{\infty}\int_0^{2\pi}\mathrm{e}^{ar^2}r\mathrm{d}\theta\mathrm{d}r=2\pi C^2\int_0^{\infty}r\mathrm{e}^{ar^2}\mathrm{d}r=\frac{\pi C^2}{a}\mathrm{e}^{ar^2}\bigg|_0^{\infty}=-\frac{\pi C^2}{a} \tag{137}$$

上式积分可积，必须要求 $a<0$。进一步计算可得：

$$a=-C^2\pi\rightarrow f(x)=C\mathrm{e}^{-C^2\pi x^2} \tag{138}$$

注意上式是观测误差的概率密度函数，也可以视为真值 $\theta=0$（即平均值 $\mu=0$）的情形。为了求出剩余的唯一常数 $C$，则需要进一步求解方差 $\sigma^2$，并令其为 1（具体计算见附录章节B.2.1中的计算证明），可最终得到（概率密度函数为正，去掉 $C$ 的负数解）

$$C=\frac{1}{\sqrt{2\pi}}\rightarrow f(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}=\Phi(x) \tag{139}$$

以上即为**标准正态随机变量** $\mathcal{N}(0,1)$。

**注** 由于上述分析是在假设特殊样本情况下给出的，所以仅证明了 $f(x)$ 的形式是极大似然估计和样本均值相等的必要条件。由于可以通过所求解出的 $f(x)$ 形式，代回似然方程(129)证明仅存在唯一解 $\hat{\theta}=\bar{x}$，即公式(131)，所以充分必要性均可得到证明。

之后对于一般 Gaussian 分布 $\mathcal{N}(\mu,\sigma^2)$ 的证明可回到附录章节B.2.1。

### B.6.2 多元 Gaussian 分布

**定理 B.12 (高维（多元）Gaussian 分布)** Gaussian 分布可推广至多变量的高维（多元）形式，其中 $\boldsymbol{x}=[x_1,x_2,x_3,\cdots]^\top$ 为多元函数的向量值写法：

$$f(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{u})\right) \tag{140}$$

**证明.** 假设随机变量 $\boldsymbol{Z}=[Z_1,Z_2,Z_3,\cdots]^\top$，其中 $Z_i\sim\mathcal{N}(0,1)(i=1,2,\cdots,n)$，自变量为 $\boldsymbol{z}=[z_1,z_2,z_3,\cdots]^\top$，即从标准正态随机变量开始，由一元函数向多元函数扩充。首先计算其期望向量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$：

$$\boldsymbol{u}=[u_1,u_2,\cdots,u_n]^\top=\boldsymbol{0},\quad \boldsymbol{\Sigma}=\boldsymbol{\sigma}\boldsymbol{\sigma}^\top=\begin{bmatrix}\sigma_1\\\sigma_2\\\vdots\\\sigma_n\end{bmatrix}\begin{bmatrix}\sigma_1 & \sigma_2 & \cdots & \sigma_n\end{bmatrix}=\mathbf{I} \tag{141}$$

以上表明 $\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I})$，且随机变量 $\boldsymbol{Z}$ 中两两互为独立事件，进一步可知：

$$p(z_1,\cdots,z_n)=p(z_1)\cdots p(z_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\cdot\mathrm{e}^{-\frac{1}{2}\cdot(z_i)^2}=\frac{1}{(2\pi)^{\frac{n}{2}}}\cdot\mathrm{e}^{-\frac{1}{2}\cdot(\boldsymbol{z}^\top\boldsymbol{z})} \tag{142}$$

且满足

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}p(z_1,\cdots,z_n)\,\mathrm{d}z_1\cdots\mathrm{d}z_n=1 \tag{143}$$

为了向一般多元 Gaussian 分布进行推广，需要向满足任意 Gaussian 分布的随机变量 $\boldsymbol{X}$ 进行线性变换的构造（$\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$），如下：

$$Z_i=\frac{X_i-\mu_i}{\sigma_i}(i=1,2,\cdots,n)\rightarrow\boldsymbol{Z}=\boldsymbol{A}^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}\right) \tag{144}$$

其中，矩阵 $\boldsymbol{A}$ 为一个 $n\times n$ 的方阵，具体元素与 $\sigma_i(i=1,2,\cdots,n)$ 相关。虽然不太容易给出其完整表达，但可以建立矩阵 $\boldsymbol{A}$ 与随机变量 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 之间的关系：

$$\boldsymbol{\Sigma}=E\left[\left(\boldsymbol{X}-\boldsymbol{\mu}\right)\left(\boldsymbol{X}-\boldsymbol{\mu}\right)^\top\right]=E\left[(\boldsymbol{A}\boldsymbol{Z})(\boldsymbol{A}\boldsymbol{Z})^\top\right]=E\left[\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}^\top\right]=\boldsymbol{A}E\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right]\boldsymbol{A}^\top=\boldsymbol{A}\boldsymbol{A}^\top \tag{145}$$

对上式两边取行列式，还可得到：

$$|\boldsymbol{\Sigma}| = \left|\boldsymbol{A}\boldsymbol{A}^{\top}\right| = |\boldsymbol{A}|^2 \tag{146}$$

将线性变换关系式(144)代入(142)，并结合关系式(145)，可知：

$$\begin{aligned} p\left(z_1\left(x_1, \cdots, x_n\right), \cdots\right) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}\cdot\left[\left(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{\top}\left(\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right]} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}\cdot\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\left(\boldsymbol{A}\boldsymbol{A}^{\top}\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}\cdot\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]} \end{aligned} \tag{147}$$

考虑到

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p\left(z_1\left(x_1, \cdots, x_n\right), \cdots\right) \mathrm{d}z_1 \cdots \mathrm{d}z_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}\cdot\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]}\right) |\boldsymbol{J}| \mathrm{d}x_1 \cdots \mathrm{d}x_n \end{aligned} \tag{148}$$

其中，$|\boldsymbol{J}|$ 为线性变换(144)的 Jacobi 行列式（换元积分），满足

$$\boldsymbol{J} = \left(\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}}\right)^{\top} = \boldsymbol{A}^{-1} \rightarrow |\boldsymbol{J}| = \left|\boldsymbol{A}^{-1}\right| = |\boldsymbol{A}|^{-1} \tag{149}$$

注意上式需使用章节1.2中所介绍的矩阵微分知识进行求解。

将公式(149)代回公式(148)并结合公式(146)，可得：

$$p\left(x_1, \cdots, x_n\right) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot \mathrm{e}^{-\frac{1}{2}\cdot\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]} \tag{150}$$

上式考虑到 $p\left(x_1, \cdots, x_n\right)$ 为正，所以公式(146)只取了正数解。

公式(150)所求得的 $p\left(x_1, \cdots, x_n\right)$ 即为高维（多元）情形下的 Gaussian 分布函数 $f(\boldsymbol{x})$，满足 $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。

**定理 B.13 (多元 Gaussian 分布性质)** 如果 $\boldsymbol{X}$ 是一个服从参数为 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的多元 Gaussian 分布的随机变量，那么 $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ 也服从多元 Gaussian 分布（$\boldsymbol{A}$ 为系数矩阵，$\boldsymbol{b}$ 为系数列向量），且参数为 $(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top})$。

该定理的证明可以结合附录章节B.2.1中的证明，以及矩阵的基本运算得出。

## B.7 贝叶斯（Bayes）公式

回顾**大数定律**：大样本统计下，发生的频率接近于真实概率——频率派。但是很多事情并不一定有足够的样本，特别是新兴事物。

**贝叶斯派**：概率是主观值，取决于判断。

**贝叶斯定理**：

Hypothesis-$H$（想要知道概率的事件）

Evidence-$E$（掌握的新信息）

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} \times P(H) \tag{151}$$

该公式可以用条件概率证明。

可以简单理解为（底层逻辑）：后验概率 = 修正因子 × 先验概率。

实际使用中，更多需要将 $P(E)$ 分解为 $H$ 情形下和非 $H$ 情形下 $E$ 发生的条件概率之和（也可以以积分形式表示为全概率），如下图所示：

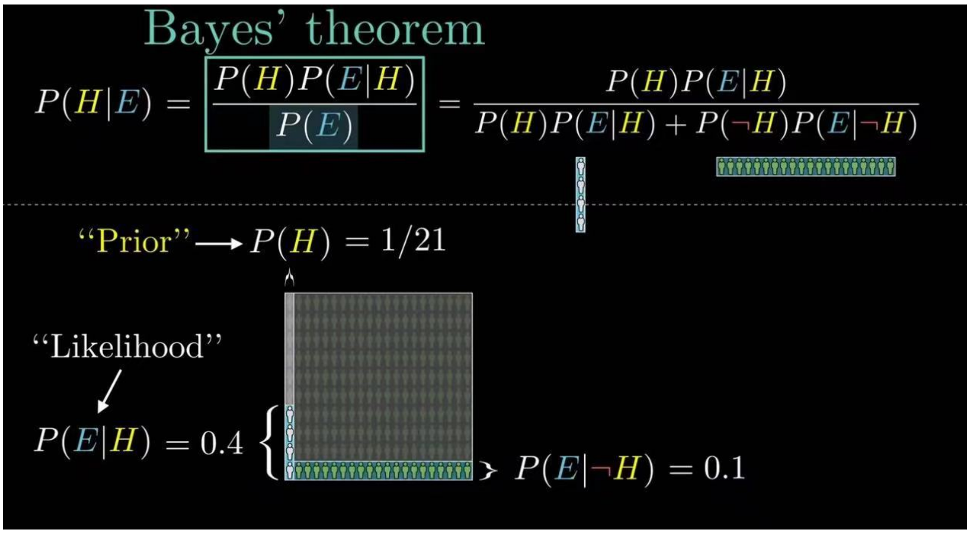


图 3: Bayes 公式的几何理解

上图还直观展示了各个概率的动态状态，即每部分的面积都可以是变动的。

该公式还可以理解为 $\frac{\text{TP}}{\text{TP+FP}}$，即统计概念中的**准确率**。

注 贝叶斯公式的本质（上层思维）是通过对想要知道概率的事件的先验判断，加上该事件基础下发生新事件的条件概率，预测出现新事件后对该事件的更新概率判断。

新事件和想要知道概率的事件可能具有某种**因果关系**，也可能没有。以下列举几种：

- 通过给定人物特征（知性、稳重等）推理属于农民还是图书管理员-弱因果；
- 通过今天升起蓝月亮判断明天太阳是否东升西落-无因果；
- 通过衣柜中发现女性内衣判断老公是否出轨-强因果。

**贝叶斯推理**三步走：

1. 先验（假设）；
2. 新数据/信息（证据），也称似然估计；
3. 后验。

之后进入重复迭代。

上述也可以理解为：大胆假设，小心求证，不断调整，快速迭代。

**深入理解：**

- 先验概率在后验概率中占了比较大的比重，就算出现了新的小概率事件，依然不会“一棍子打死”，概率会慢慢下降；
- 但一次次迭代，不断出现反常识的新事件，后验概率也会逐渐降低；
- 先验不能走入极端，无论 0%（不依赖先验，非理性）还是 100%（过度依赖先验，偏执），经过贝叶斯公式都不会更新后验概率。即要学会接受事件认知的不确定性。