第 3 章　数学回顾

3.1　微积分

3.1.1 导数

对于一元函数$y = f(x)$，记其**一阶导数**(first derivative)为$\dfrac{dy}{dx}$或$f'(x)$，其定义为

$$\frac{dy}{dx} \equiv f'(x) \equiv \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \equiv \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{3.1}$$

几何上，(一阶)导数就是函数$y = f(x)$在x处的切线斜率，参见图 3.1。

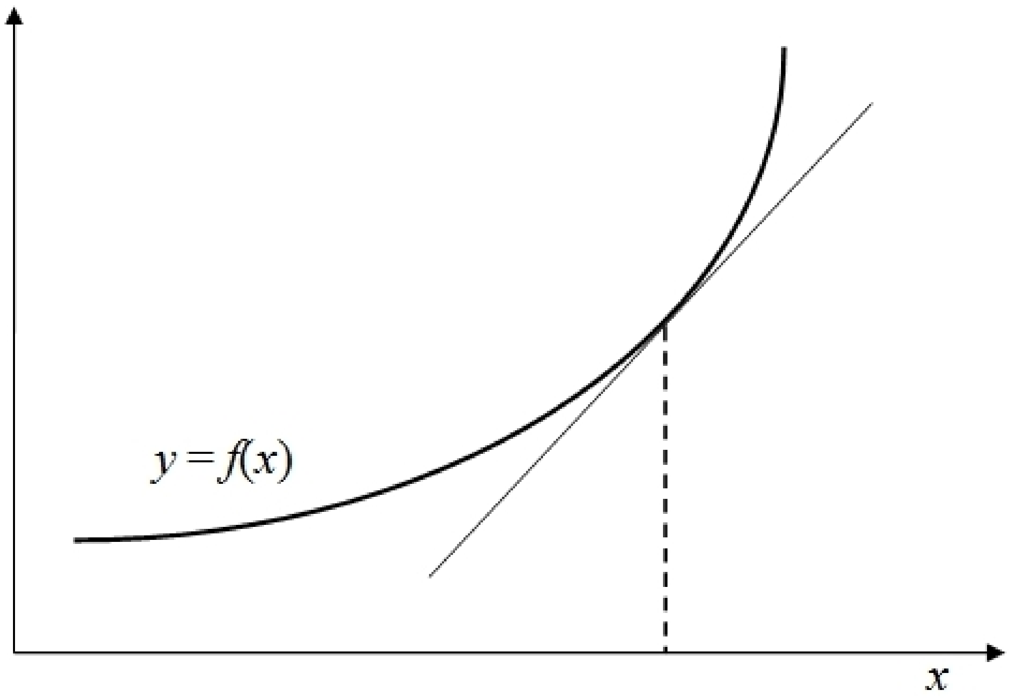

图 3.1 导数的示意图

一阶导数 $f'(x)$ 仍然是 x 的函数，故可定义 $f'(x)$ 的导数，即**二阶导数**(second derivative)：

$$\frac{d^2 y}{dx^2} \equiv \frac{d\left(\dfrac{dy}{dx}\right)}{dx} \equiv f''(x) \equiv \left[f'(x)\right]' \tag{3.2}$$

直观上，二阶导数表示切线斜率(一阶导数)的变化速度，即曲线 $f(x)$ 的弯曲程度，也称“曲率”(curvature)。

3.1.2 偏导数

对于多元函数$y = f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n)$，其中列向量$\mathbf{x} \equiv (x_1\ x_2 \cdots x_n)'$，可定义$y$对于$x_1$的**偏导数**(partial derivative)为

$$\frac{\partial y}{\partial x_1} \equiv \frac{\partial f(x_1, x_2, \cdots, x_n)}{\partial x_1}$$

$$\equiv \lim_{\Delta x_1 \to 0} \frac{f(x_1 + \Delta x_1, x_2, \cdots, x_n) - f(x_1, x_2, \cdots, x_n)}{\Delta x_1} \tag{3.3}$$

将多元函数 $f(\mathbf{x})$ 的所有偏导数写成一个列向量，即为**梯度向量**(gradient vector)：

$$\nabla f(\mathbf{x}) \equiv \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \equiv \begin{pmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{3.4}$$

由于偏导数$\dfrac{\partial y}{\partial x_1}(\mathbf{x})$依然是$(x_1, x_2, \cdots, x_n)$的函数，故可进一步求其自身的二阶偏导数，

$$\frac{\partial^2 y}{\partial x_1^2} \equiv \frac{\partial\left(\dfrac{\partial y}{\partial x_1}\right)}{\partial x_1} \tag{3.5}$$

以及混合偏导数，比如：

$$\frac{\partial^2 y}{\partial x_1 \partial x_2} \equiv \frac{\partial\left(\dfrac{\partial y}{\partial x_1}\right)}{\partial x_2} \qquad (3.6)$$

在一般情况下（要求二阶偏导数为连续函数），混合偏导数与求导顺序无关，即

$$\frac{\partial^2 y}{\partial x_1 \partial x_2} = \frac{\partial^2 y}{\partial x_2 \partial x_1} \qquad (3.7)$$

将多元函数 $f(\mathbf{x})$ 的所有二阶偏导数排成一个矩阵，即为**黑塞矩阵**(Hessian Matrix)：

$$\nabla^2 f(\mathbf{x}) \equiv \mathbf{H}(\mathbf{x}) \equiv \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \equiv \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\,\partial \mathbf{x}'} \equiv \frac{\partial\left(\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)}{\partial \mathbf{x}'}$$

$$\equiv \begin{pmatrix} \dfrac{\partial^2 y}{\partial x_1^2} & \cdots & \dfrac{\partial^2 y}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 y}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 y}{\partial x_n^2} \end{pmatrix} \tag{3.8}$$

其中，$\dfrac{\partial\left(\dfrac{\partial f(\mathbf{x})}{\partial\mathbf{x}}\right)}{\partial\mathbf{x}'}$表示对梯度(列)向量的每个分量分别求偏导，并排成一行(故分母为$\partial\mathbf{x}'$)。

由于混合偏导数与求导顺序无关，故黑塞矩阵为对称矩阵。

梯度向量与黑塞矩阵在最优化中起着重要作用。

有时我们需要同时考虑多个响应变量，比如 $y_1 = f_1(\mathbf{x}),\ \cdots,\ y_m = f_m(\mathbf{x})$，故存在$m$个函数关系。

可将其写为从向量$\mathbf{x}$到向量$\mathbf{y}$的“向量值函数”(vector-valued function)：

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \tag{3.9}$$

其中，$\mathbf{y} \equiv (y_1 \cdots y_m)'$。

将所有 $y_i = f_i(\mathbf{x})\ \ (i=1,\cdots,m)$ 的梯度向量写为行向量，然后叠放在一起，即可得到**雅各比矩阵**(Jacobian Matrix)：

$$\mathbf{J}(\mathbf{x}) \equiv \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \equiv \begin{pmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_m}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{pmatrix} \tag{3.10}$$

3.1.3 方向导数

偏导数给出了当$\mathbf{x}$沿着某坐标轴(比如x_1)变动时，函数$y = f(\mathbf{x})$变化的速率。

有时我们想知道当$\mathbf{x}$沿着任意方向变化时，函数$f(\mathbf{x})$的变化率。

任意给定$\mathbf{x}^*$，考虑$f(\mathbf{x})$在$\mathbf{x}^*$处，沿着任意方向$\mathbf{v} = (v_1 \cdots v_n)$的**方向导数**(directional derivative)，其中将$\mathbf{v}$的长度标准化为 1，即$\|\mathbf{v}\| \equiv \sqrt{v_1^2 + \cdots + v_n^2} = 1$。

沿着方向$\mathbf{v}$，经过$\mathbf{x}^*$的直线方程为

$$\mathbf{x} = \mathbf{x}^* + t\mathbf{v} \tag{3.11}$$

其中，随着参数$t \in \mathbf{R}$变化，可得到整条直线。函数$f(\mathbf{x})$沿着此直线方向的取值可写为

$$g(t) \equiv f(\mathbf{x}^* + t\mathbf{v}) = f(x_1^* + tv_1, \cdots, x_n^* + tv_n) \tag{3.12}$$

使用链式法则(chain rule)，可得函数$g(t)$在$t = 0$处的导数，即方向导数：

$$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{v}} \equiv g'(t)\Big|_{t=0} = \frac{\partial f(\mathbf{x}^*)}{\partial x_1} v_1 + \cdots + \frac{\partial f(\mathbf{x}^*)}{\partial x_n} v_n$$

$$= \begin{pmatrix} \frac{\partial f(\mathbf{x}^*)}{\partial x_1} & \cdots & \frac{\partial f(\mathbf{x}^*)}{\partial x_n} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \nabla f(\mathbf{x}^*)' \mathbf{v}$$

(3.13)

方向导数$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{v}}$是各偏导数$\frac{\partial f(\mathbf{x}^*)}{\partial x_j}$ $(j=1,\cdots,n)$的线性组合，而组合的权重为方向$\mathbf{v}$沿着各坐标轴的分量$(v_1 \cdots v_n)$。

命题 3.1 梯度向量$\nabla f(\mathbf{x})$为函数 $f(\mathbf{x})$增长最快的方向，而负梯度方向$-\nabla f(\mathbf{x})$为该函数下降最快的方向。

证明： 在n维空间$\mathbb{R}^n$中，记梯度向量$\nabla f(\mathbf{x}^*)$与方向$\mathbf{v}$的夹角为θ，则根据线性代数知识，此夹角的余弦为

$$\cos\theta = \frac{\nabla f(\mathbf{x}^*)'\mathbf{v}}{\left\|\nabla f(\mathbf{x}^*)\right\|\left\|\mathbf{v}\right\|} \tag{3.14}$$

由于$\left\|\mathbf{v}\right\| = 1$，故沿方向$\mathbf{v}$的方向导数可写为

$$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{v}} = \nabla f(\mathbf{x}^*)'\mathbf{v} = \left\|\nabla f(\mathbf{x}^*)\right\|\cos\theta \qquad (3.15)$$

由于$-1 \le \cos\theta \le 1$，故当$\cos\theta = 1$(即$\mathbf{v}$的方向与$\nabla f(\mathbf{x}^*)$相同)，方向导数$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{v}} = \left\|\nabla f(\mathbf{x}^*)\right\|$(梯度向量的长度)达到最大值。

因此，梯度向量$\nabla f(\mathbf{x}^*)$为在$\mathbf{x}^*$处，函数$f(\mathbf{x})$增加最快的方向，称为“梯度上升”(gradient ascent)。

反之，当$\cos\theta = -1$(即$\mathbf{v}$的方向与$\nabla f(\mathbf{x}^*)$相反)，方向导数$\dfrac{\partial f(\mathbf{x}^*)}{\partial \mathbf{v}} = -\left\|\nabla f(\mathbf{x}^*)\right\|$(梯度向量长度的负数)达到最小值。

因此，负梯度向量$-\nabla f(\mathbf{x}^*)$为在$\mathbf{x}^*$处，函数$f(\mathbf{x})$下降最快的方向，称为“梯度下降”(gradient descent)。

函数$f(\mathbf{x})$上升最快与下降最快的方向正好相反。

在几何上，应如何想象梯度向量？这可通过函数 $f(\mathbf{x})$ 的“水平集”(level set)来考察。

任意给定 $\mathbf{x}^*$，函数 $f(\mathbf{x})$ 相应的水平集可定义为

$$C \equiv \left\{ \mathbf{x} : f(\mathbf{x}) = f(\mathbf{x}^*) \right\} \quad (3.16)$$

水平集也称为“等值集”(contour set)；比如，地形图的“等高线”或气压图的“等压线”，参见图 3.2。

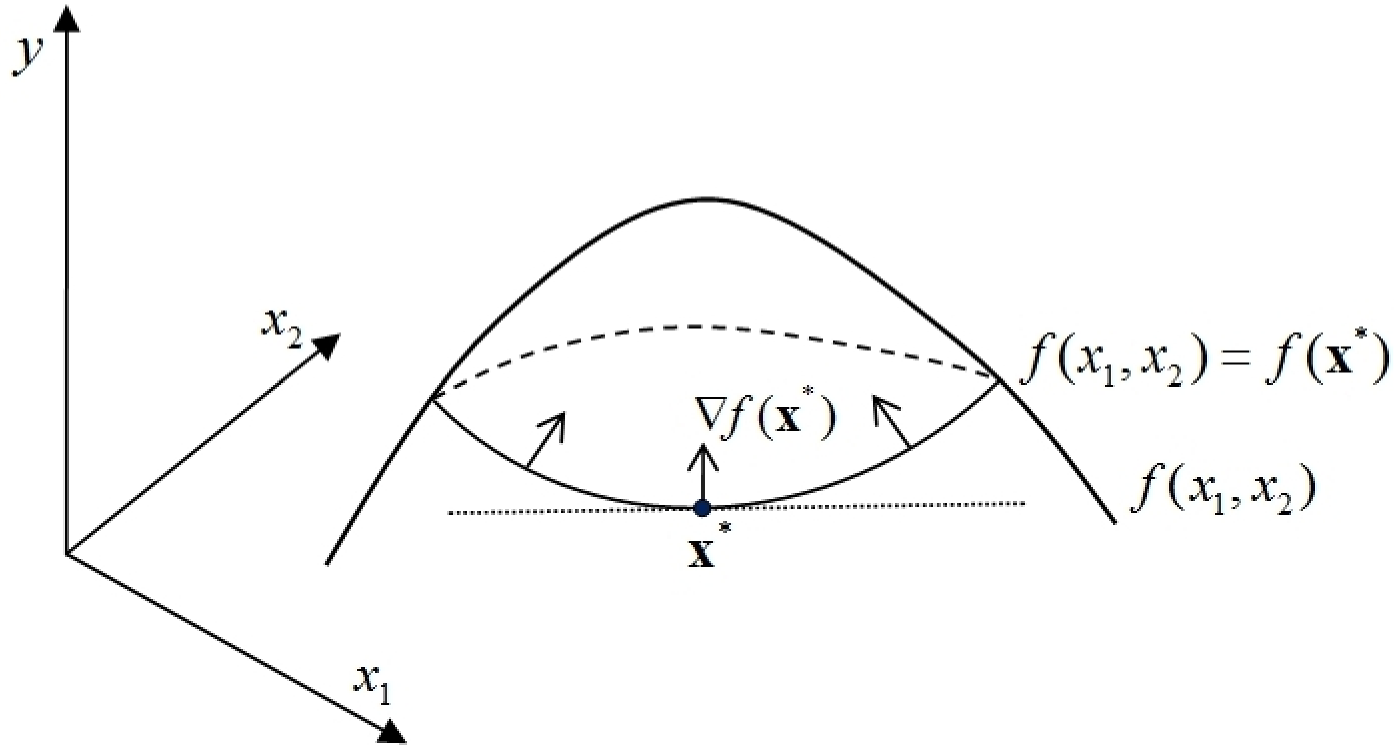

图 3.2 梯度向量与水平集(等值集)正交

命题 3.2 梯度向量 $\nabla f(\mathbf{x}^*)$ 与水平集 $C \equiv \{\mathbf{x} : f(\mathbf{x}) = f(\mathbf{x}^*)\}$ 正交。

证明： 给定水平集 $C \equiv \{\mathbf{x} : f(\mathbf{x}) = f(\mathbf{x}^*)\}$。假定

$$\mathbf{x}(t) = (x_1^* + x_1(t), \cdots, x_n^* + x_n(t)) \quad (3.17)$$

为在水平集 C 上，经过 $\mathbf{x}^*$ 的一条任意曲线。当 $t = 0$ 时，$\mathbf{x}(t) = \mathbf{x}^*$；而当 t 变化时，即得到 $\mathbb{R}^n$ 空间的一条曲线。

根据微积分知识，在$\mathbf{x}^*$处，曲线$\mathbf{x}(t)$的切线方向为

$$\frac{d\mathbf{x}(0)}{dt} = \left(\frac{dx_1(0)}{dt} \cdots \frac{dx_n(0)}{dt} \right)' \qquad (3.18)$$

将此曲线方程的表达式(3.17)代入水平集的方程(3.16)，则有

$$f(\mathbf{x}(t)) = f(x_1^* + x_1(t), \cdots, x_n^* + x_n(t)) = f(\mathbf{x}^*) \quad (3.19)$$

在$t = 0$处，将此方程两边对t求导可得：

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_1}\frac{dx_1(0)}{dt}+\cdots+\frac{\partial f(\mathbf{x}^*)}{\partial x_n}\frac{dx_n(0)}{dt}=0 \qquad (3.20)$$

上式可写为

$$\nabla f(\mathbf{x}^*)'\frac{d\mathbf{x}(0)}{dt}=0 \qquad (3.21)$$

其中，$\frac{d\mathbf{x}(0)}{dt}=\left(\frac{dx_1(0)}{dt}\cdots\frac{dx_n(0)}{dt}\right)'$为曲线$\mathbf{x}(t)$在$\mathbf{x}^*$处的切线方向，而梯度向量$\nabla f(\mathbf{x}^*)$与此切线方向正交(垂直)。

由于$\mathbf{x}(t)$为水平集C上的任意曲线，而它们在$\mathbf{x}^*$处的切线方向均与梯度向量$\nabla f(\mathbf{x}^*)$垂直，故在水平集C这一曲面上，通过点$\mathbf{x}^*$的一切曲线在点$\mathbf{x}^*$处的切线都在同一个平面上，即水平集C在点$\mathbf{x}^*$的切平面。

因此，故梯度向量$\nabla f(\mathbf{x}^*)$与水平集C正交。

3.1.4 向量微分

在进行最优化时，常需对向量求微分(vector differentiation)。

(1) **线性函数的向量微分规则** 对于线性函数 $y = \mathbf{x}'\boldsymbol{\beta}$，其向量微分为 $\dfrac{\partial(\mathbf{x}'\boldsymbol{\beta})}{\partial\mathbf{x}} = \boldsymbol{\beta}$。

将此线性函数展开写：

$$y = \mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \qquad (3.22)$$

其中，参数向量 $\boldsymbol{\beta} \equiv (\beta_1\ \beta_2\ \cdots\ \beta_n)'$。

由于$\dfrac{\partial y}{\partial x_i} = \beta_i$，故其梯度向量为

$$\frac{\partial(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}} \equiv \begin{pmatrix} \dfrac{\partial(\mathbf{x}'\boldsymbol{\beta})}{\partial x_1} \\ \vdots \\ \dfrac{\partial(\mathbf{x}'\boldsymbol{\beta})}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \boldsymbol{\beta} \tag{3.23}$$

此向量微分的规则类似于对一次函数求导。

(2) **二次型的向量微分规则** 对于二次(型)函数 $y = \mathbf{x}'\mathbf{A}\mathbf{x}$，其向量微分为 $\dfrac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$。特别地，如果 $\mathbf{A}$ 为对称矩阵，则 $\dfrac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$。

将此二次(型)函数展开写：

$$y = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j \quad (3.24)$$

其中，二次型的矩阵$\mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$。

考虑对第k个变量x_k求偏导。由于在双重求和式$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$中，只有$\sum_{j=1}^{n} a_{kj}x_k x_j$(当$x_i = x_k$)与$\sum_{i=1}^{n} a_{ik}x_i x_k$(当$x_j = x_k$)才包含$x_k$，故

$$\frac{\partial y}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$$

$$= \begin{pmatrix} a_{k1} \cdots a_{kn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} a_{1k} \cdots a_{nk} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{3.25}$$

上式适用于所有$k = 1, \cdots, n$，故共有n个方程。

将这n个方程叠放可得

$$
\begin{aligned}
\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &\equiv \begin{pmatrix} \dfrac{\partial y}{\partial x_1} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{pmatrix} \\
&= \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\
&= \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} = (\mathbf{A} + \mathbf{A}')\mathbf{x}
\end{aligned} \tag{3.26}
$$

其中，$\mathbf{A}'$为矩阵$\mathbf{A}$的转置。

如果$\mathbf{A}$为对称矩阵，则$\mathbf{A}' = \mathbf{A}$，上式可简化为

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} \tag{3.27}$$

对二次型的向量微分类似于对二次函数求导。

(3) **复合函数的向量微分规则** 对于复合函数$y = f\left(\mathbf{x}(\mathbf{z})\right)$，其向量微分为$\dfrac{\partial y}{\partial \mathbf{z}} = \left(\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}\right)' \dfrac{\partial f}{\partial \mathbf{x}}$。

将此复合函数(composite function)展开写：

$$y = f\left(\mathbf{x}(\mathbf{z})\right) = f\left(x_1(z_1,\cdots,z_K),\cdots,x_n(z_1,\cdots,z_K)\right) \tag{3.28}$$

其中，$\mathbf{x} = (x_1,\cdots,x_n)$，而$\mathbf{z} = (z_1,\cdots,z_K)$。

考虑对 z_k 求偏导数。根据微积分的链式法则(chain rule)，此复合函数的偏导数可写为

$$\frac{\partial y}{\partial z_k}=\frac{\partial f}{\partial x_1}\cdot\frac{\partial x_1}{\partial z_k}+\cdots+\frac{\partial f}{\partial x_n}\cdot\frac{\partial x_n}{\partial z_k}=\begin{pmatrix}\frac{\partial x_1}{\partial z_k} & \cdots & \frac{\partial x_n}{\partial z_k}\end{pmatrix}\begin{pmatrix}\frac{\partial f}{\partial x_1}\\ \vdots \\ \frac{\partial f}{\partial x_n}\end{pmatrix} \tag{3.29}$$

上式对$k=1,\cdots,K$均成立，将这些方程叠放，可得梯度向量：

$$\frac{\partial y}{\partial \mathbf{z}}=\begin{pmatrix}\dfrac{\partial y}{\partial z_1}\\ \vdots \\ \dfrac{\partial y}{\partial z_K}\end{pmatrix}=\begin{pmatrix}\dfrac{\partial x_1}{\partial z_1} & \cdots & \dfrac{\partial x_n}{\partial z_1}\\ \cdots & \cdots & \cdots \\ \dfrac{\partial x_1}{\partial z_K} & \cdots & \dfrac{\partial x_n}{\partial z_K}\end{pmatrix}\begin{pmatrix}\dfrac{\partial f}{\partial x_1}\\ \vdots \\ \dfrac{\partial f}{\partial x_n}\end{pmatrix}=\left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}}\right)'\frac{\partial f}{\partial \mathbf{x}} \tag{3.30}$$

其中，$\left(\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}\right)'$为$\mathbf{x}(\mathbf{z})$的雅各比矩阵$\dfrac{\partial \mathbf{x}}{\partial \mathbf{z}}$之转置。

上式似乎与一元复合函数的微分规则不同，但如果将上式两边同时转置则有

$$\left(\frac{\partial y}{\partial \mathbf{z}}\right)' = \left(\frac{\partial f}{\partial \mathbf{x}}\right)' \left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}}\right) \tag{3.31}$$

这意味着，如果将梯度向量定义为行向量，则复合函数的向量微分规则在形式上与一元复合函数相同。

在较早的文献中，为了保持这种形式上的一致，一般将梯度向量定义为偏导数的行向量。

3.2 最优化

3.2.1 一元最优化

机器学习的主要方法为最优化(optimization)，尤其是最小化问题(minimization)。有时也涉及最大化问题(maximization)，比如最大似然估计(Maximum Likelihood Estimation，简记 MLE)。

最大化问题可等价地写为最小化问题，只要将目标函数加上负号即可：

$$\max_{x} f(x) = \min_{x} \left[-f(x)\right] \tag{3.32}$$

因此，我们主要讨论最小化问题。

考虑以下无约束(unconstrained)的一元最小化问题(参见图3.3)，

$$\min_{x} \ f(x) \tag{3.33}$$

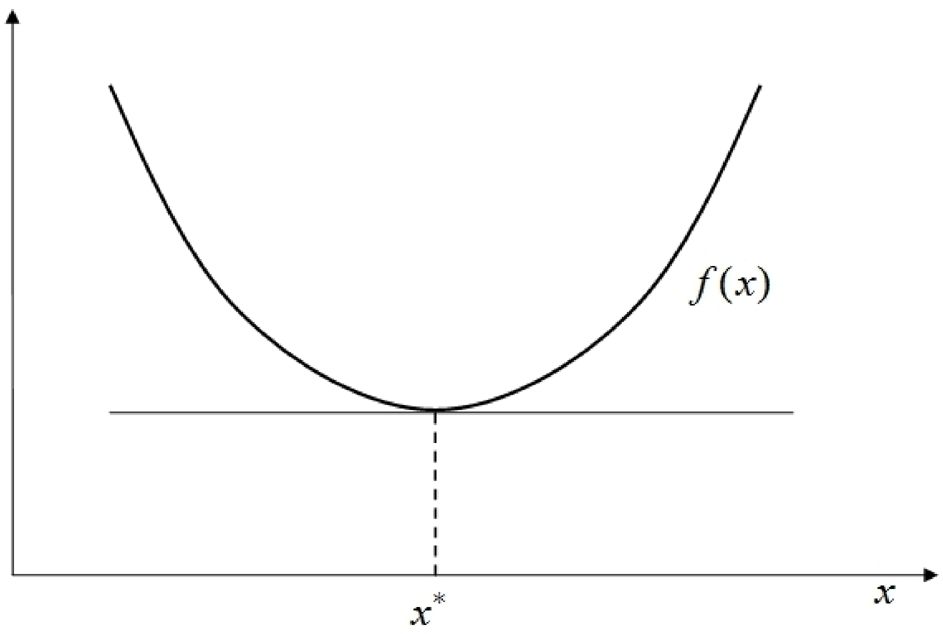

图 3.3 最小化的示意图

从图 3.3 可知，函数 $f(x)$ 在其“山谷”底部 x^* 处达到局部最小值。

在山底 x^* 处，$f(x)$ 的切线恰好为水平线，故切线斜率为 0。故一元最小化问题的必要条件为

$$f'(x^*) = 0 \tag{3.34}$$

由于此最小化的必要条件涉及一阶导数，故称为**一阶条件** (first order condition)。

直观上，如果 $f'(x^*) < 0$，则在 x^* 处增加 x 可使函数值 $f(x)$ 进一步下降，故 $f(x^*)$ 不是最小值；反之，如果 $f'(x^*) > 0$，则在 x^* 处减少 x 可使函数值 $f(x)$ 进一步下降，故 $f(x^*)$ 也不是最小值。因此，在最小值处，必然有 $f'(x^*) = 0$。

根据同样的逻辑，最大化问题的一阶条件与最小化问题相同，都要求在最优值 x^* 处的切线斜率为 0，即 $f'(x^*) = 0$，参见图 3.4。

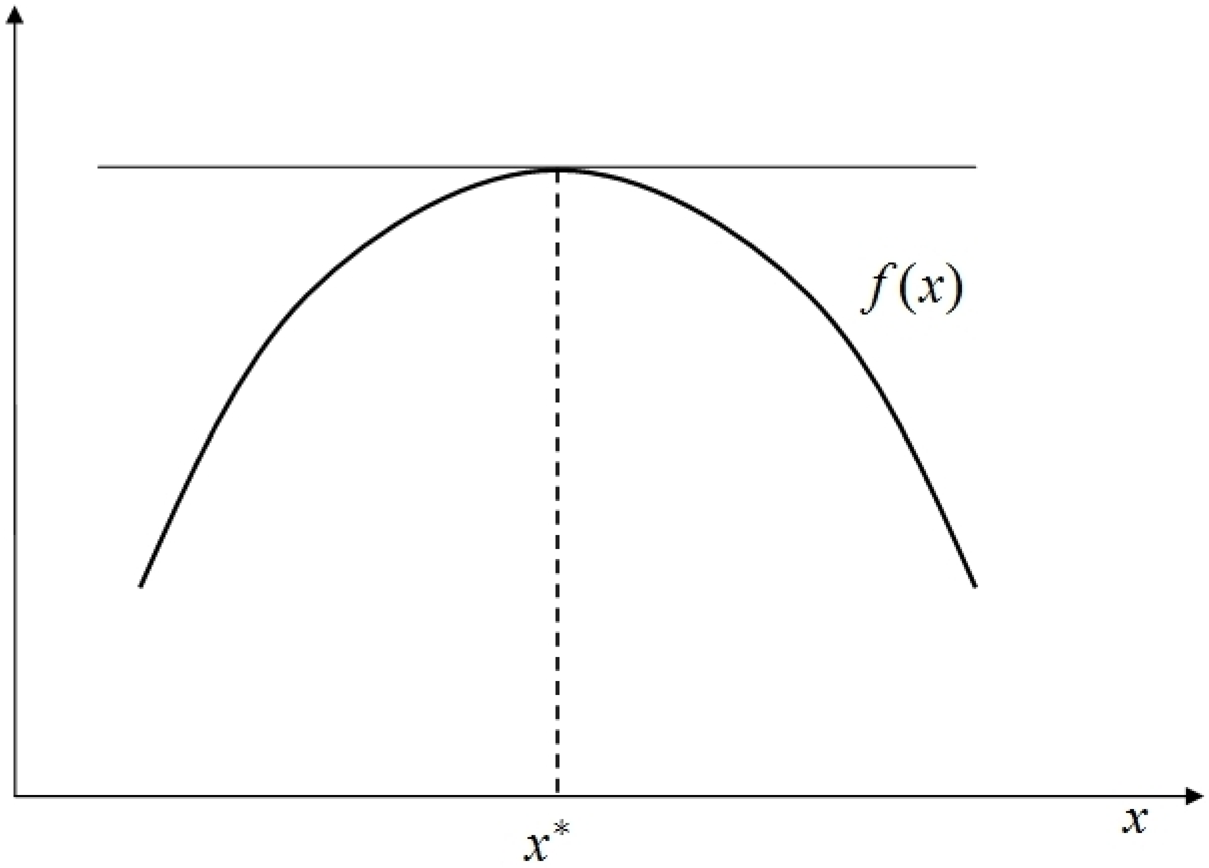

图 3.4 最大化的示意图

最小化问题与最大化问题的区别在于最优化的**二阶条件**(second order condition)，即最小化要求 $f''(x^*) \geq 0$(函数 $f(x)$在x^*处为凸函数)，而最大化要求二阶导数 $f''(x^*) \leq 0$(函数 $f(x)$在 x^*处为凹函数)。

对于最优化的一阶条件与二阶条件，下面给出较为严格的证明。假设 $f(x)$在 x^*处达到局部最小值，则对于在 x^*附近任意小的扰动Δx，都有

$$f(x^*) \leq f(x^* + \Delta x) \qquad (3.35)$$

假设函数 $f(x)$ 二阶连续可导(twice continuously differentiable)，可将上式右边在 x^* 处进行二阶“泰勒展开”(Taylor expansion)：

$$f(x^* + \Delta x) = f(x^*) + f'(x^*)\Delta x + \frac{1}{2} f''(x^* + \theta\Delta x)(\Delta x)^2 \tag{3.36}$$

其中，$0 < \theta < 1$。将此式代入(3.35)可得“基本不等式”(fundamental inequality)：

$$f'(x^*)\Delta x + \frac{1}{2} f''(x^* + \theta\Delta x)(\Delta x)^2 \geq 0 \tag{3.37}$$

上式对任意小的Δx都成立。

如果$\Delta x > 0$，在上式两边同除以Δx，并求右极限($\Delta x \to 0^+$)可得

$$\lim_{\Delta x \to 0^+} \left[f'(x^*) + \frac{1}{2} f''(x^* + \theta \Delta x)(\Delta x) \right] = f'(x^*) \geq 0 \tag{3.38}$$

反之，如果$\Delta x < 0$，在上式两边同除以Δx，并求左极限$(\Delta x \to 0^-)$可得

$$\lim_{\Delta x \to 0^+}\left[f'(x^*) + \frac{1}{2} f''(x^* + \theta \Delta x)(\Delta x) \right] = f'(x^*) \le 0 \tag{3.39}$$

综合不等式(3.38)与(3.39)可知，最小化问题的必要条件为$f'(x^*) = 0$。将此一阶条件代入基本不等式(3.37)可得：

$$f''(x^* + \theta \Delta x)(\Delta x)^2 \ge 0 \tag{3.40}$$

在上式中，由于$(\Delta x)^2 \ge 0$，故

$$f''(x^* + \theta \Delta x) \geq 0 \qquad (3.41)$$

由于此式对于任意小的Δx都成立，而二阶导数$f''(x)$假设为连续函数，故最小化的二阶条件要求：

$$f''(x^*) \geq 0 \qquad (3.42)$$

根据同样的逻辑，“严格局部最小值”(strict local minimum)的充分条件包括：一阶条件 $f'(x^*)=0$，二阶条件 $f''(x^*)>0$。如果此一阶条件与二阶条件均成立，则基本不等式(3.37)取严格大于号(>)，故 x^* 为严格局部最小值。

同理可证，最大化的二阶条件要求 $f''(x^*)\leq 0$。“严格局部最大值”(strict local maximum)的充分条件包括：一阶条件 $f'(x^*)=0$，二阶条件 $f''(x^*)<0$。

3.2.2 多元最优化

更一般地，考虑以下无约束的多元最小化问题，

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \equiv f(x_1, x_2, \cdots, x_n) \qquad (3.43)$$

其中，$\mathbf{x} \equiv (x_1 \ x_2 \cdots x_n)'$。

此最小化问题的一阶条件要求，在最优值$\mathbf{x}^*$处，梯度向量等于$\mathbf{0}$：

$$\nabla f(\mathbf{x}^*) = \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}} = \begin{pmatrix} \dfrac{\partial f(\mathbf{x}^*)}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x}^*)}{\partial x_n} \end{pmatrix} = \mathbf{0} \qquad (3.44)$$

假设函数 $f(\mathbf{x})$ 在 $\mathbf{x}^* \equiv (x_1^* \ x_2^* \cdots x_n^*)'$ 达到最小值，则一元函数 $f(x_1, x_2^*, \cdots, x_n^*)$ 在 $x_1 = x_1^*$ 达到最小值。故根据一元最小化的一阶条件可得，$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_1} = \dfrac{\partial f(x_1^*, x_2^*, \cdots, x_n^*)}{\partial x_1} = 0$。由此可知，在最优值 $\mathbf{x}^*$ 处，所有的偏导数均等于 0：

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_1} = \frac{\partial f(\mathbf{x}^*)}{\partial x_2} = \cdots = \frac{\partial f(\mathbf{x}^*)}{\partial x_n} = 0 \qquad (3.45)$$

多元最大化的一阶条件与此相同。

假设 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 达到局部最小值，则对于在 $\mathbf{x}^*$ 附近任意小的扰动 $\Delta\mathbf{x}$，都有

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + h\Delta\mathbf{x}) \qquad (3.46)$$

其中，$h > 0$ 为任意小的正数，$\Delta\mathbf{x} \equiv (\Delta x_1\ \Delta x_2 \cdots \Delta x_n)'$ 为 n 维空间 $\mathbb{R}^n$ 的一个方向，而 Δx_j 为对 x_j 任意小的扰动，$j = 1, \cdots, n$。

假设函数 $f(\mathbf{x})$ 二阶连续可导，可将上式右边在 $\mathbf{x}^*$ 处进行二阶泰勒展开：

$$
\begin{aligned}
& f(\mathbf{x}^* + h\Delta\mathbf{x}) \\
& = f(\mathbf{x}^*) + h\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}\Delta\mathbf{x} + \frac{1}{2}h^2 \underbrace{(\Delta\mathbf{x})'\frac{\partial^2 f(\mathbf{x}^* + \theta h\Delta\mathbf{x})}{\partial \mathbf{x}^2}(\Delta\mathbf{x})}_{\text{二次型}}
\end{aligned}
\tag{3.47}
$$

其中，$0<\theta<1$，而二次型$(\Delta\mathbf{x})'\dfrac{\partial^2 f(\mathbf{x}^* + \theta h\Delta\mathbf{x})}{\partial \mathbf{x}^2}(\Delta\mathbf{x})$的矩阵为黑塞矩阵$\dfrac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}$在$(\mathbf{x}^* + \theta h\Delta\mathbf{x})$处的取值。

将上式代入(3.46)可得基本不等式：

$$h\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}\Delta\mathbf{x}+\frac{1}{2}h^2(\Delta\mathbf{x})'\frac{\partial^2 f(\mathbf{x}^*+\theta h\Delta\mathbf{x})}{\partial \mathbf{x}^2}(\Delta\mathbf{x})\geq 0 \tag{3.48}$$

在上式中，代入一阶条件$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}=\mathbf{0}$，并消去$\frac{1}{2}h^2$可得

$$(\Delta\mathbf{x})'\frac{\partial^2 f(\mathbf{x}^*+\theta h\Delta\mathbf{x})}{\partial \mathbf{x}^2}(\Delta\mathbf{x})\geq 0 \tag{3.49}$$

根据半正定矩阵的定义(参见下文)，上式要求黑塞矩阵 $\left[\dfrac{\partial^2 f(\mathbf{x}^* + \theta h\Delta\mathbf{x})}{\partial\mathbf{x}^2}\right]$ 半正定(positive semidefinite)。

由于 $\Delta\mathbf{x}$ 为任意的扰动，且假设二阶导函数连续，故最小化的二阶条件要求：

$$(\Delta\mathbf{x})'\frac{\partial^2 f(\mathbf{x}^*)}{\partial\mathbf{x}^2}(\Delta\mathbf{x}) \geq 0 \quad (3.50)$$

故在最小值 $\mathbf{x}^*$ 处，黑塞矩阵 $\left[\dfrac{\partial^2 f(\mathbf{x}^*)}{\partial\mathbf{x}^2}\right]$ 半正定。

几何上，这要求在 $\mathbf{x}^*$ 处，函数 $f(\mathbf{x})$ 为凸函数(convex function)。

反之，对于最大化问题，其二阶条件则要求黑塞矩阵 $\left[\frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}^2}\right]$ 半负定(negative semidefinite)，即函数 $f(\mathbf{x})$ 在 $\mathbf{x}^*$ 处为凹函数(concave function)。

3.2.3 约束极值问题：等式约束

有时我们还会遇到如下“约束极值”(constrained optimization) 问题，比如

$$\begin{aligned} &\min_{x_1,x_2} \; f(x_1,x_2) \\ &s.t.\; g(x_1,x_2)=b \end{aligned} \tag{3.51}$$

其中，“$s.t.$”表示“subject to”，即“可行解”(feasible solutions) 受到非线性等式“$g(x_1,x_2)=b$”的约束。

求解方法之一为“消元法”(elimination)，即根据约束条件，将x_2写为x_1的函数，然后代入目标函数，将其变为无约束的一元极值问题。

在一定条件下，上述约束条件定义了一个隐函数(implicit function)$x_2 = h(x_1)$。

对约束条件“$g(x_1, x_2) = b$”进行全微分可得：

$$dg = \frac{\partial g}{\partial x_1} dx_1 + \frac{\partial g}{\partial x_2} dx_2 = 0 \qquad (3.52)$$

由此可得隐函数$x_2 = h(x_1)$的导数：

$$\frac{dh}{dx_1} = \frac{dx_2}{dx_1} = -\frac{\partial g / \partial x_1}{\partial g / \partial x_2} \tag{3.53}$$

将隐函数$x_2 = h(x_1)$代入目标函数，可将此二元约束极值问题，转化为一元无约束极值问题：

$$\min_{x_1} F(x_1) \equiv f(x_1, h(x_1)) \tag{3.54}$$

根据无约束极值的一阶条件可得：

$$\frac{dF(x_1)}{dx_1} = \frac{\partial f}{\partial x_1} + \frac{\partial f}{\partial x_2}\frac{dh}{dx_1} = 0 \qquad (3.55)$$

将表达式(3.53)代入上式可得：

$$\frac{dF(x_1)}{dx_1} = \frac{\partial f}{\partial x_1} - \left(\frac{\partial f / \partial x_2}{\partial g / \partial x_2}\right)\frac{\partial g}{\partial x_1} = 0 \qquad (3.56)$$

另一方面，下式显然也成立：

$$\frac{\partial f}{\partial x_2} - \underbrace{\left(\frac{\partial f / \partial x_2}{\partial g / \partial x_2}\right)}_{\equiv \lambda} \frac{\partial g}{\partial x_2} = 0 \qquad (3.57)$$

定义$\lambda \equiv \dfrac{\partial f / \partial x_2}{\partial g / \partial x_2}$，则方程(3.56)与(3.57)意味着，约束极值问题的一阶条件为

$$\frac{\partial f}{\partial x_j} - \lambda \frac{\partial g}{\partial x_j} = 0 \quad (j = 1, 2) \qquad (3.58)$$

可通过定义如下“拉格朗日函数”(Lagrangian)来得到上述一阶条件：

$$\min_{x_1,x_2,\lambda} \; L(x_1,x_2,\lambda)=f(x_1,x_2)+\lambda\left[b-g(x_1,x_2)\right] \quad (3.59)$$

其中，λ为“拉格朗日乘子”(Lagrange multiplier)，也是优化的变量。

上式可视为针对(x_1, x_2, λ)的无约束优化问题，其一阶条件为

$$\frac{\partial L}{\partial x_j} = \frac{\partial f(\mathbf{x}^*)}{\partial x_j} - \lambda^* \frac{\partial g(\mathbf{x}^*)}{\partial x_j} = 0 \quad (j = 1, 2) \tag{3.60}$$

$$\frac{\partial L}{\partial \lambda} = b - g(x_1^*, x_2^*) = 0 \tag{3.61}$$

其中，方程(3.61)正是原来的约束条件“$g(x_1, x_2) = b$”。

方程(3.60)意味着，在最优解$\mathbf{x}^*$处，目标函数$f(\mathbf{x})$的梯度向量与约束条件$g(\mathbf{x})$的梯度向量平行(二者可能方向相同或相反)，二者仅相差一个倍数λ^*：

$$\begin{pmatrix} \dfrac{\partial f(\mathbf{x}^*)}{\partial x_1} \\ \dfrac{\partial f(\mathbf{x}^*)}{\partial x_2} \end{pmatrix} = \lambda^* \begin{pmatrix} \dfrac{\partial g(\mathbf{x}^*)}{\partial x_1} \\ \dfrac{\partial g(\mathbf{x}^*)}{\partial x_2} \end{pmatrix} \tag{3.62}$$

由于梯度向量与水平集(等值线)正交，故在最优解$\mathbf{x}^*$处，

目标函数 $f(\mathbf{x})$ 的等值线正好与函数 $g(\mathbf{x})$ 的等值线“$g(\mathbf{x})=b$”相切，参见图 3.5。

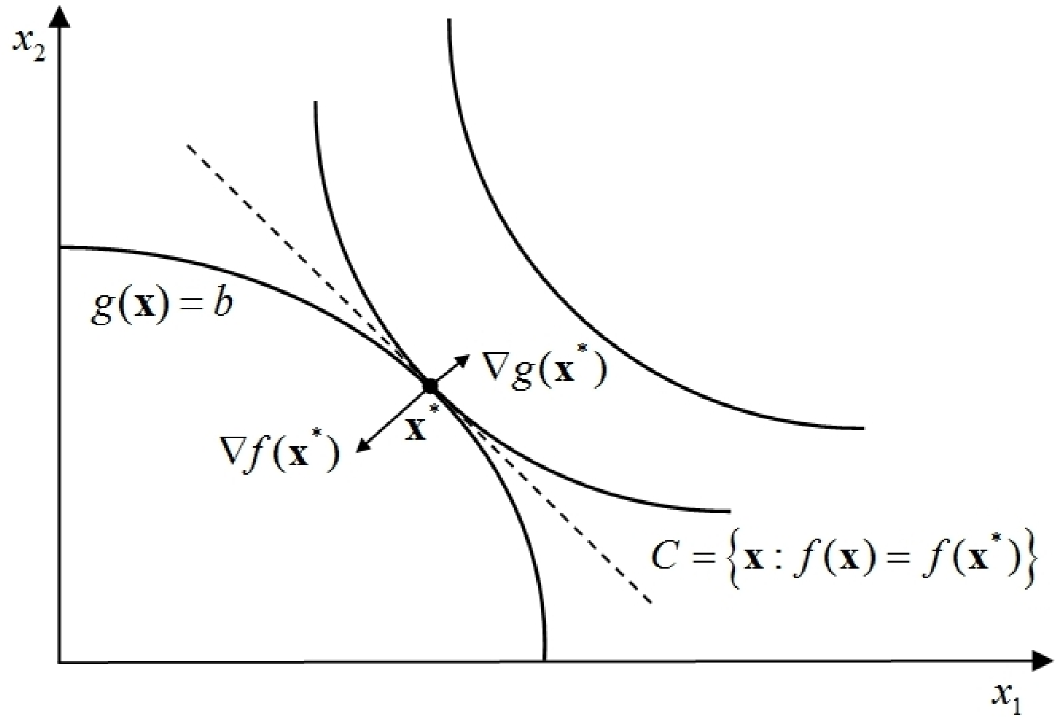

图 3.5 约束极值问题的一阶条件图示

拉格朗日乘子λ有何意义？显然，最优解$(x_1^*, x_2^*, \lambda^*)$为参数$b$的函数，可写为

$$x_1^* = x_1(b), \quad x_1^* = x_1(b), \quad \lambda^* = \lambda(b) \qquad (3.63)$$

将上式代入拉格朗日函数(3.59)可得：

$$L(b) = f\left(x_1(b), x_2(b)\right) + \lambda(b)\left[b - g\left(x_1(b), x_2(b)\right)\right] \qquad (3.64)$$

把上式对b求导数可得：

$$
\begin{aligned}
\frac{\partial L(b)}{\partial b} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial b} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial b} + \lambda'(b)\underbrace{\left[b - g(x_1, x_2)\right]}_{=0\ (\text{约束条件})} + \lambda^*\left[1 - \frac{\partial g}{\partial x_1}\frac{\partial x_1}{\partial b} - \frac{\partial g}{\partial x_2}\frac{\partial x_2}{\partial b}\right] \\
&= \underbrace{\left(\frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial b} - \lambda^*\frac{\partial g}{\partial x_1}\frac{\partial x_1}{\partial b}\right)}_{=0\ (\text{一阶条件})} + \underbrace{\left(\frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial b} - \lambda^*\frac{\partial g}{\partial x_2}\frac{\partial x_2}{\partial b}\right)}_{=0\ (\text{一阶条件})} + \lambda^* \\
&= \lambda^*
\end{aligned}
\tag{3.65}
$$

由于在最优解处，$L(b) = f(x_1(b), x_2(b))$，因此

$$
\frac{\partial L(b)}{\partial b} = \frac{\partial f(x_1(b), x_1(b))}{\partial b} = \lambda^* \tag{3.66}
$$

故最优拉格朗日乘子$\lambda^* = \lambda(b)$，等于放松约束条件b对目标函数最优值$f(x_1(b), x_2(b))$的边际作用(marginal effect)。

如果将约束条件“$g(x_1, x_2) = b$”视为资源约束(可用资源总量为b)，则在经济学上可将$\lambda^* = \lambda(b)$解释为资源的“影子价格”(shadow prices)，反映此资源的重要性或价值。

更一般地，考虑受到m个非线性等式约束的n元最小化问题：

$$\begin{aligned} &\min_{\mathbf{x}} \ f(\mathbf{x}) = f(x_1, \cdots, x_n) \\ &s.t. \ \ g_1(\mathbf{x}) = b_1, \cdots, g_m(\mathbf{x}) = b_m \end{aligned} \tag{3.67}$$

此时，由于有m个约束条件，故须在拉格朗日函数中引入m个拉格朗日乘子$(\lambda_1 \cdots \lambda_m)$：

$$\min_{\mathbf{x},\,\boldsymbol{\lambda}} \ L(\mathbf{x}, \lambda_1, \cdots, \lambda_m) = f(\mathbf{x}) + \sum_{k=1}^{m} \lambda_k \left[b_k - g_k(\mathbf{x}) \right] \tag{3.68}$$

相应的一阶条件为

$$\frac{\partial L(\mathbf{x},\lambda_1,\cdots,\lambda_m)}{\partial \mathbf{x}}=\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}-\sum_{k=1}^{m}\lambda_k\frac{\partial g_k(\mathbf{x})}{\partial \mathbf{x}}=\mathbf{0} \quad (3.69)$$

$$\frac{\partial L(\mathbf{x},\lambda_1,\cdots,\lambda_m)}{\partial \lambda_k}=b_k-g_k(\mathbf{x})=0 \quad (k=1,\cdots,m) \quad (3.70)$$

其中，等式(3.69)包含n方程(因为$\mathbf{x}$为n维向量)，而等式(3.70)包括原来的m个约束条件。

更简洁地，可以定义

$$\boldsymbol{\lambda} \equiv \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}, \quad \mathbf{b} \equiv \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix} \tag{3.71}$$

则拉格朗日函数可写为

$$\min_{\mathbf{x},\, \boldsymbol{\lambda}} \; L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}' \left[\mathbf{b} - \mathbf{g}(\mathbf{x}) \right] \tag{3.72}$$

使用向量微分的规则，一阶条件可写为

$$\frac{\partial L(\mathbf{x},\boldsymbol{\lambda})}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - \left[\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\right]' \boldsymbol{\lambda} = \mathbf{0} \quad (3.73)$$

$$\frac{\partial L(\mathbf{x},\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathbf{b} - \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.74)$$

其中，方程(3.73)的$\dfrac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}$为$\mathbf{g}(\mathbf{x})$的雅各比矩阵，并使用了复合函数的向量微分规则(将$\boldsymbol{\lambda}'\mathbf{g}(\mathbf{x})$视为复合函数)。

最优解$(\mathbf{x}^*, \boldsymbol{\lambda}^*)$包含$(n+m)$个变量，须同时满足等式(3.73)与(3.74)，共$(n+m)$个方程。

在几何上，一阶条件(3.73)意味着，目标函数的梯度向量$\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$为各约束条件梯度向量$\dfrac{\partial g_k(\mathbf{x})}{\partial \mathbf{x}}$的线性组合，而组合权重即为相应的拉格朗日乘子$\lambda_k$(反映每个约束条件的重要性)：

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial g_1(\mathbf{x})}{\partial \mathbf{x}} \cdots \frac{\partial g_m(\mathbf{x})}{\partial \mathbf{x}} \right) \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} = \sum_{k=1}^{m} \lambda_k \frac{\partial g_k(\mathbf{x})}{\partial \mathbf{x}} \tag{3.75}$$

拉格朗日乘子向量$\boldsymbol{\lambda}$依然可解释为资源约束的影子价格。

例如，拉格朗日乘子λ_1可解释为放松约束条件“$g_1(\mathbf{x}) = b_1$”对目标函数最优值的边际作用，以此类推。

二阶条件则要求，在最小值$\mathbf{x}^*$处，目标函数$f(\mathbf{x})$的黑塞矩阵$\left[\dfrac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}^2}\right]$在约束集$\{\mathbf{x}:\ \mathbf{g}(\mathbf{x})=\mathbf{b}\}$中半正定。

如果不限制在约束集$\{\mathbf{x}:\ \mathbf{g}(\mathbf{x})=\mathbf{b}\}$内，则黑塞矩阵也可以是“不定的” (indefinite)。例如，考虑以下目标函数：

$$y = f(\mathbf{x}) = x_1^2 - x_2^2 \tag{3.76}$$

显然，此函数$f(\mathbf{x})$是不定的，其几何形状为鞍形(saddle)，参见图 3.6。

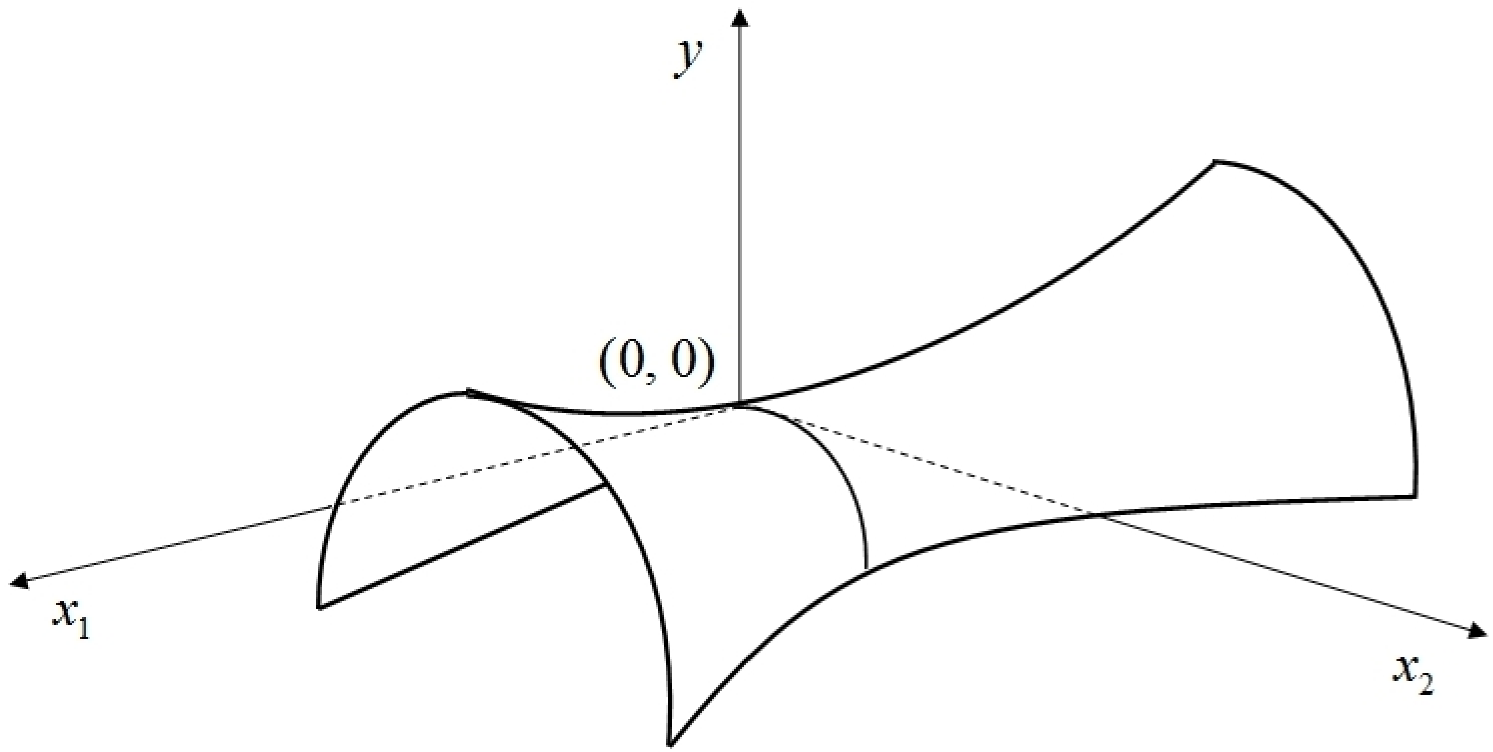

图 3.6 函数 $y = x_1^2 - x_2^2$ 的鞍形图像

从图 3.6 可见，由于函数$y = x_1^2 - x_2^2$是不定的，故并无最大值或最小值。

但如果加上约束条件“$x_2 = 0$”，则函数$y = x_1^2 - x_2^2 = x_1^2$，故在约束集$\{(x_1, x_2):\ x_2 = 0\}$(即$x_1$轴)中正定，并在“$x_1 = 0$”处达到最小值。

反之，如果加上约束条件“$x_1 = 0$”，则函数$y = x_1^2 - x_2^2 = -x_2^2$，故在约束集$\{(x_1, x_2):\ x_1 = 0\}$(即$x_2$轴)中负定，并在“$x_2 = 0$”处达到最大值。

对于函数$y = x_1^2 - x_2^2$，原点$(0, 0)$为其“鞍点”(saddle point)。在此鞍点$(0, 0)$处，沿着x_1的方向，函数$y = x_1^2 - x_2^2$达到最小值；而沿着x_2的方向，则函数$y = x_1^2 - x_2^2$达到最大值。

在一定条件下，可以证明，约束极值问题(3.67)的最优解$(\mathbf{x}^*, \boldsymbol{\lambda}^*)$，正是拉格朗日函数$L(\mathbf{x}, \boldsymbol{\lambda})$的鞍点。具体来说，在鞍点$(\mathbf{x}^*, \boldsymbol{\lambda}^*)$处，沿着$\mathbf{x}$的方向，拉格朗日函数$L(\mathbf{x}, \boldsymbol{\lambda})$在$\mathbf{x}^*$处达到最大值；而沿着$\boldsymbol{\lambda}$的方向，则拉格朗日函数$L(\mathbf{x}, \boldsymbol{\lambda})$在$\boldsymbol{\lambda}^*$处达到最小值。

3.2.4 约束极值问题：非负约束

在有些优化问题中，要求优化变量$\mathbf{x}$只能取非负值。此时，最小化问题可写为

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) = f(x_1, \cdots, x_n) \\ s.t. \quad & \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{3.77}$$

对目标函数在最优解$\mathbf{x}^*$处进行二阶泰勒展开，依然可得到同样的基本不等式：

$$h\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}\Delta\mathbf{x}+\frac{1}{2}h^2(\Delta\mathbf{x})'\frac{\partial^2 f(\mathbf{x}^*+\theta h\Delta\mathbf{x})}{\partial \mathbf{x}^2}(\Delta\mathbf{x})\geq 0 \tag{3.78}$$

如果$\mathbf{x}^*$为“内点解”(interior solution)，即$\mathbf{x}^* > \mathbf{0}$，则上式对于任意方向的$\Delta\mathbf{x}$都成立，故一阶条件依然要求梯度向量为$\mathbf{0}$，即$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}} = \mathbf{0}$。

如果最优解$\mathbf{x}^*$的某个分量x_j^*发生于“边界”(boundary)，即$x_j^* = 0$，则为了满足约束条件“$\mathbf{x} \geq \mathbf{0}$”，必有$\Delta x_j > 0$(增量必须为正)。假设其他分量的变动均为 0，即$\Delta x_i = 0 \ (i \neq j)$，则上式可写为

$$h \frac{\partial f(\mathbf{x}^*)}{\partial x_j} \Delta x_j + \frac{1}{2} h^2 \frac{\partial^2 f(\mathbf{x}^* + \theta h \Delta \mathbf{x})}{\partial x_j^2} (\Delta x_j)^2 \geq 0 \quad (3.79)$$

其中，$\Delta \mathbf{x} = (0 \cdots 0 \ \Delta x_j \ 0 \cdots 0)'$。

上式两边同除以$\Delta x_j > 0$，并让$\Delta x_j \to 0^+$可得

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_j} \geq 0 \quad (如果\ x_j^* = 0) \qquad (3.80)$$

总之　，对于内点解$x_j^* > 0$，一阶条件为$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j} = 0$；而对于边界解$x_j^* = 0$，则一阶条件为$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j} \geq 0$，参见图 3.7。

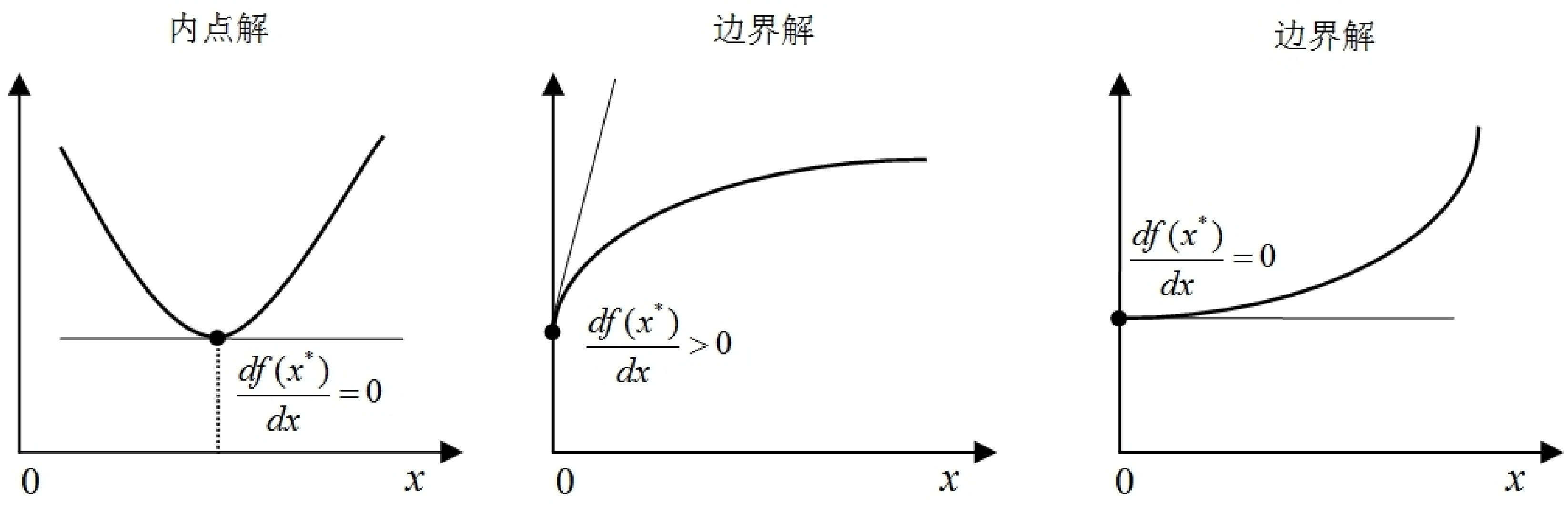

图 3.7 非负约束极值问题的一阶条件

综上所述，要么$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j}=0$(内点解)，要么$x_j^*=0$(边界解)，

故二者的乘积必然为 0：

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_j}\cdot x_j^*=0 \quad (j=1,\cdots,n) \tag{3.81}$$

上式称为“互补松弛条件”(complementary slackness conditions)，它意味着，如果$x_j^* > 0$(不等式“$x_j^* \geq 0$”取严格大于号，即所谓“松弛”)，则不等式$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j} \geq 0$就必须取等号，即$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j} = 0$，故此不等式是“紧的”(tight 或 binding)。

反之，如果$\dfrac{\partial f(\mathbf{x}^*)}{\partial x_j} > 0$(不等式取严格大于号，处于“松弛”状态)，则不等式$x_j^* \geq 0$就必须是紧的，即$x_j^* = 0$。

由于方程(3.81)对于 $j=1,\cdots,n$ 均成立，加总这 n 个方程可得：

$$\sum_{j=1}^{n}\frac{\partial f(\mathbf{x}^*)}{\partial x_j}\cdot x_j^*=\left[\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}\right]'\mathbf{x}^*=0 \qquad (3.82)$$

由于 $\frac{\partial f(\mathbf{x}^*)}{\partial x_j}\geq 0$ 且 $x_j^*\geq 0$，故求和式(3.82)意味着，每一项 $\frac{\partial f(\mathbf{x}^*)}{\partial x_j}\cdot x_j^*$ 均为 0，即方程(3.81)对所有 $j=1,\cdots,n$ 都成立。

因此，对于非负约束的极值问题，其一阶条件包括以下 $(2n+1)$ 个方程：

$$\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}} \geq \mathbf{0}, \quad \left[\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}\right]' \mathbf{x}^* = 0, \quad \mathbf{x}^* \geq 0 \tag{3.83}$$

3.2.5 约束极值问题：不等式约束

考虑以下更一般的不等式约束极值问题：

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) = f(x_1, \cdots, x_n) \\ s.t. \quad & \mathbf{g}(\mathbf{x}) \leq \mathbf{b}, \ \ \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{3.84}$$

其中，$\mathbf{g}(\mathbf{x}) = \left(g_1(\mathbf{x}) \cdots g_m(\mathbf{x})\right)'$，而$\mathbf{b} = (b_1 \cdots b_m)'$。求解方法之一为，引入 m 个“松弛变量”(slack variables) $\mathbf{s} \equiv (s_1, \cdots, s_m)'$：

$$\mathbf{s} \equiv \mathbf{b} - \mathbf{g}(\mathbf{x}) \equiv (s_1 \cdots s_m)' \tag{3.85}$$

将上式代入目标函数，可将不等式约束变为等式约束：

$$\begin{aligned}&\min_{\mathbf{x}} \quad f(\mathbf{x}) = f(x_1, \cdots, x_n) \\ &s.t. \;\; \mathbf{g}(\mathbf{x}) + \mathbf{s} = \mathbf{b}, \;\; \mathbf{x} \geq \mathbf{0}, \;\; \mathbf{s} \geq \mathbf{0}\end{aligned} \tag{3.86}$$

相应的拉格朗日函数为

$$\min_{\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}} \; L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}'\left[\mathbf{b} - \mathbf{g}(\mathbf{x}) - \mathbf{s}\right] \tag{3.87}$$

由于存在非负约束$\mathbf{x} \geq \mathbf{0}$，故有关$\mathbf{x}$的一阶条件为

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - \boldsymbol{\lambda}' \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \geq \mathbf{0}, \quad \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - \boldsymbol{\lambda}' \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\right]' \mathbf{x} = 0, \tag{3.88}$$

有关拉格朗日乘子的一阶条件依然为约束条件：

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{b} - \mathbf{g}(\mathbf{x}) - \mathbf{s} = \mathbf{0} \tag{3.89}$$

由于松弛变量$\mathbf{s}$受到非负约束$\mathbf{s} \geq \mathbf{0}$，故有关$\mathbf{s}$的一阶条件为

$$\frac{\partial L}{\partial \mathbf{s}} = -\boldsymbol{\lambda} \geq \mathbf{0}, \qquad \left[\frac{\partial L}{\partial \mathbf{s}}\right]' \mathbf{s} = -\boldsymbol{\lambda}'\mathbf{s} = 0, \qquad \mathbf{s} \geq \mathbf{0} \quad (3.90)$$

在一阶条件(3.89)与(3.90)中，代入松弛变量的表达式$\mathbf{s} = \mathbf{b} - \mathbf{g}(\mathbf{x})$，可将一阶条件简化(消去$\mathbf{s}$)：

$$\mathbf{b} - \mathbf{g}(\mathbf{x}) \geq \mathbf{0} \qquad (3.91)$$

$$\boldsymbol{\lambda} \leq \mathbf{0} \qquad (3.92)$$

$$\boldsymbol{\lambda}'(\mathbf{b} - \mathbf{g}(\mathbf{x})) = 0 \qquad (3.93)$$

其中，表达式(3.91)就是原来的不等式约束$\mathbf{g}(\mathbf{x}) \leq \mathbf{b}$。

表达式(3.92)表明，拉格朗日乘子$\boldsymbol{\lambda} \leq \mathbf{0}$，这意味着放松约束条件“$\mathbf{g}(\mathbf{x}) \leq \mathbf{b}$”(即增加$\mathbf{b}$)，可行解的集合变大，只会使得最优的目标函数下降或不变(这是最小化问题)。

方程(3.93)也是“互补松弛条件”，它意味着，要么拉格朗日乘子$\boldsymbol{\lambda} = \mathbf{0}$(影子价格为 0)，此时可允许“$\mathbf{g}(\mathbf{x}) < \mathbf{b}$”(资源不必用尽)；要么拉格朗日乘子$\boldsymbol{\lambda} < \mathbf{0}$(影子价格为负，有助于降低目标函数)，此时必须有“$\mathbf{g}(\mathbf{x}) = \mathbf{b}$”(资源完全用尽)。

在上述推导中，松弛变量$\mathbf{s}$仅起到桥梁作用，并不出现于最终的表达式，故也可仍使用原来(不包含$\mathbf{s}$)的拉格朗日函数：

$$\min_{\mathbf{x},\,\boldsymbol{\lambda}} \; L(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}'\left[\mathbf{b} - \mathbf{g}(\mathbf{x})\right] \qquad (3.94)$$

然后直接得到相应的一阶条件：

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - \boldsymbol{\lambda}'\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \geq \mathbf{0} \qquad (3.95)$$

$$\left[\frac{\partial L}{\partial \mathbf{x}}\right]' \mathbf{x} = \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - \boldsymbol{\lambda}'\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\right]' \mathbf{x} = 0 \quad (3.96)$$

$$\mathbf{x} \geq \mathbf{0} \qquad (3.97)$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{b} - \mathbf{g}(\mathbf{x}) \geq \mathbf{0} \tag{3.98}$$

$$\boldsymbol{\lambda}' \frac{\partial L}{\partial \boldsymbol{\lambda}} = \boldsymbol{\lambda}'(\mathbf{b} - \mathbf{g}(\mathbf{x})) = 0 \tag{3.99}$$

$$\boldsymbol{\lambda} \leq \mathbf{0} \tag{3.100}$$

表达式(3.95)-(3.100)即所谓“Karush-Kuhn-Tucker conditions”，简记“KKT 条件”。这些条件刻画了(characterize)不等式约束极值问题(3.84)的最优解$(\mathbf{x}^*, \boldsymbol{\lambda}^*)$。在具体求解时，一般仍需使用某种“迭代算法”(iterative algorithm)。

3.2.6 最优化算法

机器学习的最优化问题通常没有解析解，故一般需使用迭代算法来逼近最优解。对于最小化问题$\min_{\mathbf{x}} f(\mathbf{x})$，迭代算法的最一般公式为

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta\mathbf{x}_t \tag{3.101}$$

其中，$\mathbf{x}_t$为第t步迭代的取值，$\mathbf{x}_{t+1}$为第$t+1$步迭代的取值，而$\Delta\mathbf{x}_t$为该步迭代的变动幅度。

具体来说，变动幅度$\Delta\mathbf{x}_t$一般可写为梯度向量的线性函数：

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \cdot \mathbf{A}_t \nabla f(\mathbf{x}_t) \quad (3.102)$$

其中，$\nabla f(\mathbf{x}_t)$为在$\mathbf{x}_t$处的梯度向量，$\mathbf{A}_t$为$n \times n$矩阵(可随着t而更新)；而$0 < \eta < 1$是一个很小的正数，称为“步长”(step size)或“学习率”(learning rate)。

作为最简单的算法，令$\mathbf{A}_t = \mathbf{I}_n$(单位矩阵)，即可得到**梯度下降法**(gradient descent)，最早由法国数学家柯西(Cauchy)于 1847 年提出：

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) \quad (3.103)$$

此时，函数的变动方向为负梯度向量，即 $\mathbf{x}_{t+1}-\mathbf{x}_t=-\eta\nabla f(\mathbf{x}_t)$。

根据命题 3.1，负梯度方向$-\nabla f(\mathbf{x}_t)$是函数$f(\mathbf{x})$下降最快的方向。

很小的学习率η(比如 0.1 或 0.01)，可防止沿着负梯度方向走得太远，以至于“过犹不及” (overshoot)。这是因为$-\nabla f(\mathbf{x}_t)$只是在$\mathbf{x}_t$处函数下降最快的方向，离开$\mathbf{x}_t$之后就没有保证了。

另一方面，如果学习率η太小，则迭代收敛的速度可能很慢。

使用公式(3.103)反复迭代，直至梯度向量等于$\mathbf{0}$(或十分接近于$\mathbf{0}$)，即可停止迭代，认为达到局部最小值。

图 3.8 为通过梯度下降法寻找函数$z = x^2 + 2y^2$最小值的示意图。

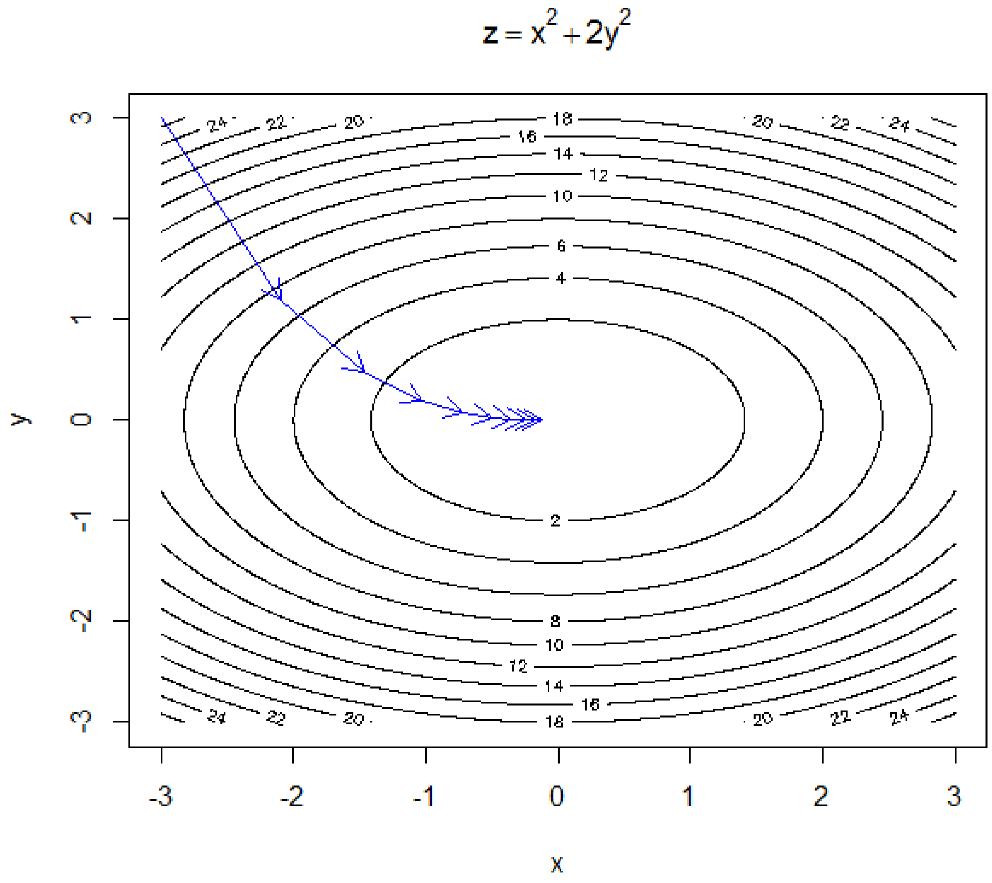

图 3.8 梯度下降法的示意图

在标准的梯度下降法中，步长η一般是固定的。另一方法是，在每步迭代时，均搜索最优的步长，即求解如下一元最小化问题：

$$\eta_t = \underset{\eta}{\operatorname{argmin}}\ f\left(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)\right) \qquad (3.104)$$

上式意味着，从$\mathbf{x}_t$出发，沿着$-\eta \nabla f(\mathbf{x}_t)$的直线方向(给定$\nabla f(\mathbf{x}_t)$，而变化$\eta$)，搜索能使函数$f\left(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)\right)$最小化的$\eta$值(比如进行网格搜索，即 grid search)，故称为“直线搜索”(line search)。由于在进行梯度下降时，选择最优的步长η_t，使得

函数的下降幅度最大，故此法称为**最速下降法**(steepest descent)。

在迭代算法的最一般公式(3.102)中，如果令$\mathbf{A}_t = \left[\mathbf{H}(\mathbf{x}_t)\right]^{-1}$(黑塞矩阵的逆矩阵)，则为**牛顿法**(Newton's method)或**牛顿−拉夫森法**(Newton-Raphson method)：

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\left[\mathbf{H}(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) \qquad (3.105)$$

为推导牛顿法的表达式，将函数$f(\mathbf{x})$在$\mathbf{x}_t$处进行二阶泰勒展开，并忽略高阶项可得：

$$f(\mathbf{x}) \approx f(\mathbf{x}_t) + \left[\nabla f(\mathbf{x}_t)\right]'(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_t)'\mathbf{H}(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t) \tag{3.106}$$

上式右边为二次(型)函数。为求上式右边的最小值，对$\mathbf{x}$求导(使用向量微分规则)，并令其梯度向量为$\mathbf{0}$：

$$\nabla f(\mathbf{x}_t) + \mathbf{H}(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t) = \mathbf{0} \tag{3.107}$$

由此可得最优解为

$$\mathbf{x}^* = \mathbf{x}_t - \left[\mathbf{H}(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) \tag{3.108}$$

在上式加上学习率η，即可得到牛顿法的表达式(3.105)。

其中，$-\left[\mathbf{H}(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)$称为“牛顿方向”。

之所以需将学习率η乘以牛顿方向$-\left[\mathbf{H}(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)$，是因为忽略高阶项的二阶泰勒展开式(3.106)，仅在$\mathbf{x}_t$附近才成立。

作为一种二阶方法，牛顿法一般比梯度下降法更有效率，收敛速度更快。如果目标函数就是二次函数，则牛顿法可一步收敛到最优值(将学习率η设为 1)。

但对于一般的函数 $f(\mathbf{x})$，如果初始值选择不恰当，则牛顿法也可能不收敛。

牛顿法需要计算黑塞矩阵，并求其逆矩阵，在很多情况下过于复杂而费时，故在机器学习中经常使用梯度下降的一阶方法。

机器学习还用到一些其他优化算法，比如“坐标下降法”(coordinate descent)、“随机梯度下降”(stochastic gradient descent)等，将在后续章节介绍。

3.3　线性代数

3.3.1 矩阵

将$m \times n$个实数排列成如下矩形的阵列，

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{3.109}$$

称$\mathbf{A}$为$m \times n$级**矩阵**(matrix)，其中m为矩阵$\mathbf{A}$的行数(row dimension)，而n为矩阵$\mathbf{A}$的列数(column dimension)。$\mathbf{A}$中元素a_{ij}表示矩阵$\mathbf{A}$的第i行、第j列元素。

矩阵$\mathbf{A}$有时也记为$\mathbf{A}_{m\times n}$(以下标强调矩阵的维度)，或$(a_{ij})_{m\times n}$(以代表性元素a_{ij}表示此矩阵)。

如果$\mathbf{A}$中所有元素都为 0，则为**零矩阵**(zero matrix)，记为$\mathbf{0}$。

零矩阵在矩阵运算中的作用，相当于 0 在数的运算中的作用。

3.3.2 方阵

如果$m = n$，则称$\mathbf{A}$为n级**方阵**(square matrix)，即

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{3.110}$$

称$a_{11}, a_{22}, \cdots, a_{nn}$为主对角线上的元素(diagonal elements)，而$\mathbf{A}$中其他元素为非主对角线元素(off-diagonal elements)。

如果方阵$\mathbf{A}$中的元素满足$a_{ij}=a_{ji}$(任意$i, j=1,\cdots,n$)，则称矩阵$\mathbf{A}$为**对称矩阵**(symmetric matrix)。

如果方阵$\mathbf{A}$的非主对角线元素全部为 0，则称为**对角矩阵**(diagonal matrix)：

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \tag{3.111}$$

如果一个n级对角矩阵的主对角线元素都为 1，则称为n级**单位矩阵**(identity matrix)，记为$\mathbf{I}$或$\mathbf{I}_n$(以下标n强调其维度)：

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n\times n} \tag{3.112}$$

单位矩阵在矩阵运算中的作用，相当于 1 在数的运算中的作用。

3.3.3 矩阵的转置

如果将矩阵$\mathbf{A} = (a_{ij})_{m\times n}$的第 1 行变为第 1 列，第 2 行变为第 2 列,……,第m行变为第m列,则可得到其**转置矩阵**(transpose)，记为$\mathbf{A}'$(读为$\mathbf{A}$ *prime*)，其维度为$n \times m$。

矩阵$\mathbf{A}'$的(i, j)元素正好是矩阵$\mathbf{A}$的(j, i)元素，即

$$(\mathbf{A}')_{ij} \equiv (\mathbf{A})_{ji} \qquad (3.113)$$

如果$\mathbf{A}$为对称矩阵，则$\mathbf{A}$的转置还是它本身，即$\mathbf{A}' = \mathbf{A}$。显然，矩阵转置的转置仍是它本身，即$(\mathbf{A}')' = \mathbf{A}$。

3.3.4. 向量

如果$m=1$，则矩阵$\mathbf{A}_{1\times n}$为n维**行向量**(row vector)；如果$n=1$，则矩阵$\mathbf{A}_{m\times 1}$为m维**列向量**(column vector)。

向量是矩阵的特例。

考察n维列向量$\mathbf{a}=(a_1\ \ a_2\ \cdots\ \ a_n)'$与$\mathbf{b}=(b_1\ \ b_2\ \cdots\ \ b_n)'$。

向量$\mathbf{a}$与$\mathbf{b}$的**内积**(inner product)或**点乘**(dot product)可定义为

$$\mathbf{a}'\mathbf{b} \equiv \mathbf{a} \cdot \mathbf{b} \equiv \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \equiv a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \tag{3.114}$$

如果$\mathbf{a}'\mathbf{b} = 0$，则称向量$\mathbf{a}$与$\mathbf{b}$**正交**(orthogonal)，这意味着两个向量在n维向量空间中相互垂直(夹角为 90 度)，参见图 3.9。

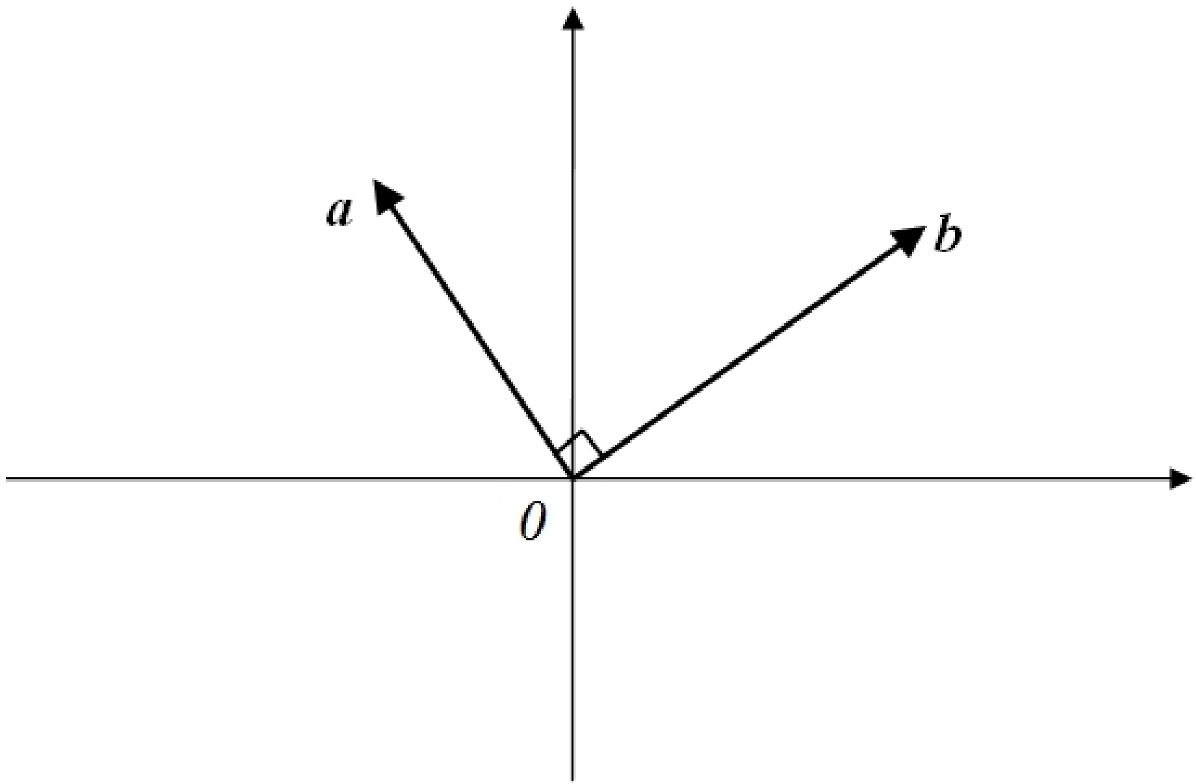

图 3.9 正交的向量

方程(3.114)提示我们，任何形如$\sum_{i=1}^{n} a_i b_i$的乘积求和，都可写为向量内积$\mathbf{a}'\mathbf{b}$的形式。特别地，平方和$\sum_{i=1}^{n} a_i^2$可写为$\mathbf{a}'\mathbf{a}$：

$$\mathbf{a}'\mathbf{a} \equiv \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \equiv a_1^2 + a_2^2 + \cdots + a_n^2 = \sum_{i=1}^{n} a_i^2 \tag{3.115}$$

可定义向量$\mathbf{a}$的“长度”(length)，也称为**范数**(norm)或“模长”，即向量$\mathbf{a}$与原点($\mathbf{0}$向量)之间的欧几里得距离：

$$\|\mathbf{a}\|_2 \equiv \|\mathbf{a}\| \equiv \sqrt{\mathbf{a}'\mathbf{a}} = \left(a_1^2 + a_2^2 + \cdots + a_n^2\right)^{\frac{1}{2}} \tag{3.116}$$

由于使用平方，故也称为 2-范数(L_2norm)。

更一般地，记向量$\mathbf{a}$与$\mathbf{b}$的夹角为θ，则此夹角的余弦为

$$\cos\theta = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{3.117}$$

如果将向量$\mathbf{b}$的长度标准化为 1，即$\|\mathbf{b}\| = 1$，则有

$$\mathbf{a}'\mathbf{b} = \|\mathbf{a}\| \cos\theta \qquad (3.118)$$

向量内积$\mathbf{a}'\mathbf{b}$的几何意义为向量$\mathbf{a}$对向量$\mathbf{b}$所做的投影之长度(可带正负号)，参见图 3.10。

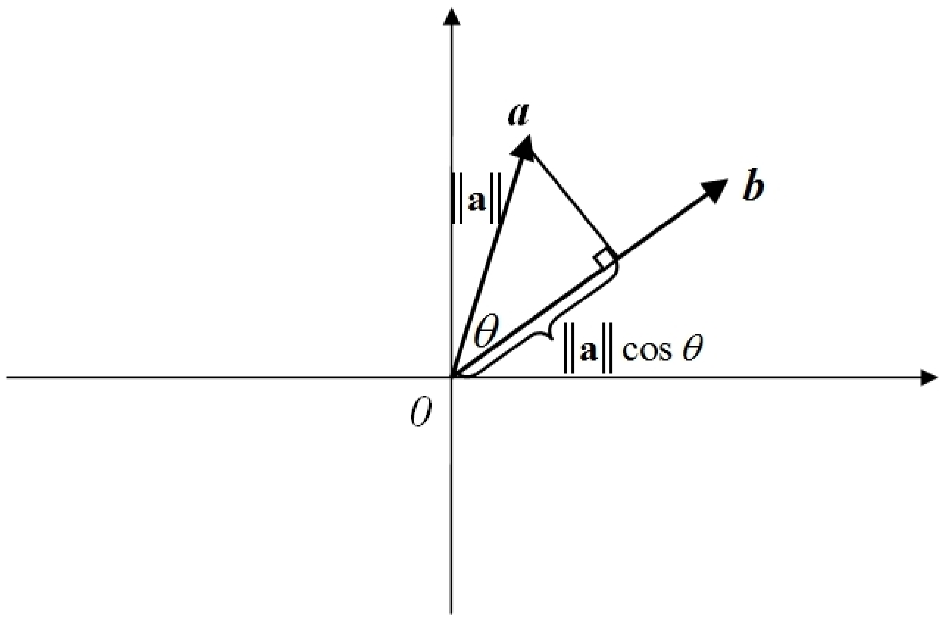

图 3.10 向量内积的几何意义为投影

将 2-范数推广，可定义更一般的 p–范数(L_p norm)：

$$\|\mathbf{a}\|_p \equiv \left(|a_1|^p + |a_2|^p + \cdots + |a_n|^p\right)^{\frac{1}{p}} \quad (3.119)$$

其中，$p \geq 1$。除了 2-范数外，较常用的范数包括 1-范数：

$$\|\mathbf{a}\|_1 \equiv |a_1| + |a_2| + \cdots + |a_n| \quad (3.120)$$

向量$\mathbf{a}$的 1-范数即为其各分量的绝对值之和。1-范数也称为“曼哈顿距离”(Manhattan distance)，因为纽约曼哈顿的街区均为东西-南北的网格状，故只能直角行走，参见图 3.11。

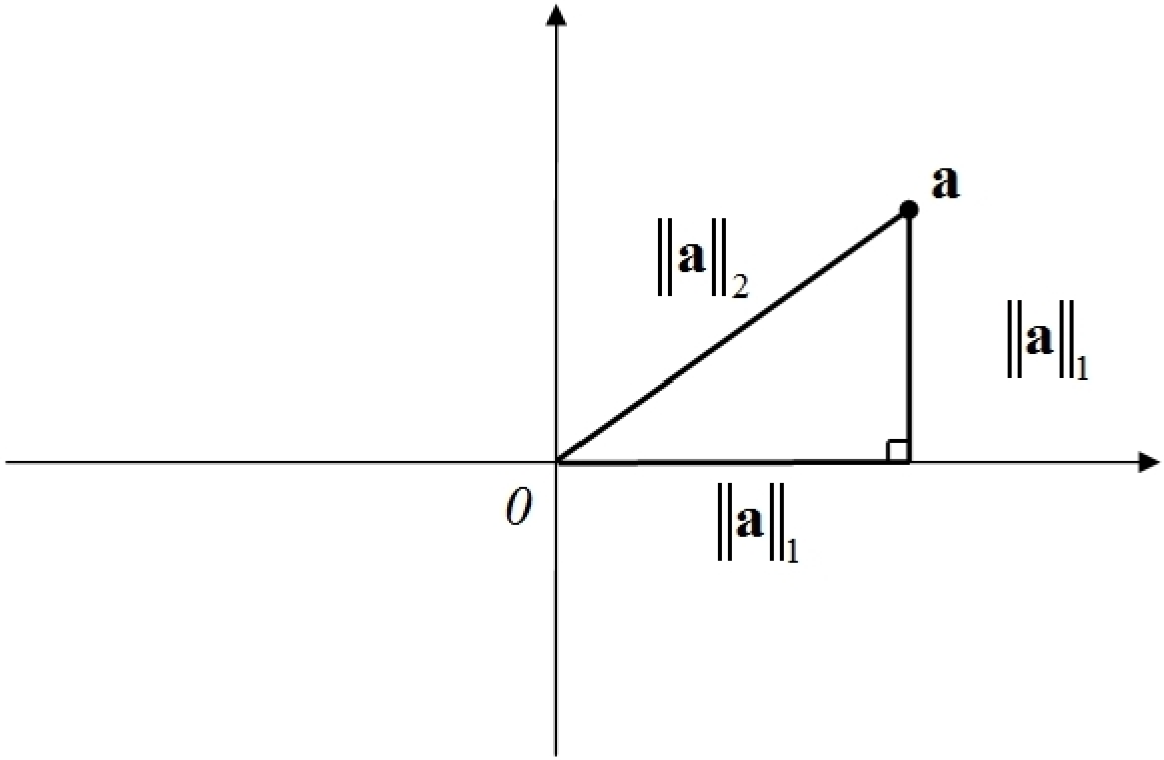

图 3.11 L_2范数与L_1范数

当 $p \to \infty$ 时，公式(3.119)右边将由向量 $\mathbf{a}$ 中分量绝对值的最大者所主导，故可定义**无穷范数**(infinity norm)：

$$\|\mathbf{a}\|_{\infty} \equiv \max\left(|a_1|, |a_2|, \cdots, |a_n|\right) \qquad (3.121)$$

在一般情况下，如无特别说明，则默认向量的范数为 2-范数，即 $\|\mathbf{a}\| = \|\mathbf{a}\|_2$。

对于n维列向量$\mathbf{a} = (a_1 \;\; a_2 \; \cdots \; a_n)'$与$\mathbf{b} = (b_1 \;\; b_2 \; \cdots \; b_n)'$，还可考虑向量$\mathbf{a}$与$\mathbf{b}$的外积(outer product)：

$$\mathbf{ab}' = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} (b_1 \;\; b_2 \; \cdots \; b_n) = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_n \end{pmatrix} \tag{3.122}$$

向量$\mathbf{a}$与$\mathbf{b}$的外积为$n \times n$矩阵(而内积为一个数)。

在上式右边的外积矩阵中，第 1 列为向量$\mathbf{a}$的b_1倍，而第 2 列为向量$\mathbf{a}$的b_2倍，以此类推。

因此，外积矩阵的所有列向量均为向量$\mathbf{a}$的某个倍数，故外积矩阵$\mathbf{ab}'$的秩(rank)为 1，称为“秩一矩阵”(rank one matrix)。

3.3.5 矩阵的加法

如果两个矩阵的维度相同(即行数与列数都分别相同)，则可以相加。对于$m\times n$级矩阵$\mathbf{A}=(a_{ij})_{m\times n}$，$\mathbf{B}=(b_{ij})_{m\times n}$，矩阵$\mathbf{A}$与$\mathbf{B}$之和定义为两个矩阵相应元素之和，即

$$\mathbf{A}+\mathbf{B}\equiv(a_{ij})_{m\times n}+(b_{ij})_{m\times n}\equiv(a_{ij}+b_{ij})_{m\times n} \tag{3.123}$$

容易证明，矩阵的加法满足以下规则：

(1)$\mathbf{A}+\mathbf{0}=\mathbf{A}$ (加上零矩阵不改变矩阵)

(2)$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$ (加法交换律)

(3)$(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$ (加法结合律)

(4)$(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$ (转置为线性运算)

3.3.6 矩阵的数乘

矩阵$\mathbf{A} = (a_{ij})_{m\times n}$与实数$k$的**数乘**(scalar multiplication)定义为此实数k与矩阵$\mathbf{A} = (a_{ij})_{m\times n}$每个元素的乘积：

$$k\mathbf{A} \equiv k(a_{ij})_{m\times n} \equiv (ka_{ij})_{m\times n} \qquad (3.124)$$

3.3.7 矩阵的乘法

如果矩阵$\mathbf{A}$的列数与矩阵$\mathbf{B}$的行数相同，则可定义**矩阵乘积**(matrix multiplication)$\mathbf{A}\times\mathbf{B}$，简记$\mathbf{AB}$。

假设矩阵$\mathbf{A}=(a_{ij})_{m\times n}$，矩阵$\mathbf{B}=(b_{ij})_{n\times q}$，则矩阵乘积$\mathbf{AB}$的$(i,j)$元素为矩阵$\mathbf{A}$第$i$行与矩阵$\mathbf{B}$的第$j$列的内积：

$$(\mathbf{AB})_{ij}\equiv\begin{pmatrix}a_{i1} & a_{i2} & \cdots & a_{in}\end{pmatrix}\begin{pmatrix}b_{1j}\\ b_{2j}\\ \vdots\\ b_{nj}\end{pmatrix}=\sum_{k=1}^{n}a_{ik}b_{kj} \tag{3.125}$$

矩阵乘法不满足交换律，一般来说，$\mathbf{AB} \neq \mathbf{BA}$。只有当矩阵 $\mathbf{B}$ 的列数 q 等于矩阵 $\mathbf{A}$ 的行数 m 时，$\mathbf{B}_{n\times q}\mathbf{A}_{m\times n}$ 才有定义。因此，在做矩阵乘法时，需区分左乘 (premultiplication) 与右乘 (postmultiplication)，即 $\mathbf{A}$ 左乘 $\mathbf{B}$ 为 $\mathbf{AB}$，而 $\mathbf{A}$ 右乘 $\mathbf{B}$ 为 $\mathbf{BA}$。

矩阵的乘法满足以下规则：

(1) $\mathbf{IA} = \mathbf{A}$，$\mathbf{AI} = \mathbf{A}$ (乘以单位矩阵不改变矩阵)

(2) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ (乘法结合律)

(3) $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ (乘法分配律)

(4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$ (转置与乘积的混合运算)

3.3.8 线性方程组

考虑以下由n个方程，n个未知数构成的线性方程组：

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{3.126}$$

其中，$(x_1 \quad x_2 \cdots \quad x_n)$为未知数。根据矩阵乘法的定义，可将上式写为

$$
\underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{21} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}}_{\mathbf{b}} \tag{3.127}
$$

记上式中的相应矩阵分别为$\mathbf{A}$，$\mathbf{x}$与$\mathbf{b}$，可得：

$$
\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.128}
$$

直观上，如果可将此方程左边的方阵$\mathbf{A}$“除”到右边去，则可得到$\mathbf{x}$的解。为此，引入逆矩阵的概念。

3.3.9 逆矩阵

对于n级方阵$\mathbf{A}$，如果存在n级方阵$\mathbf{B}$，使得$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$ (n 级单位矩阵)，则称$\mathbf{A}$为**可逆矩阵**(invertible matrix)或**非退化矩阵**(nonsingular matrix)，而$\mathbf{B}$为$\mathbf{A}$的**逆矩阵**(inverse matrix)，记为$\mathbf{A}^{-1}$。

逆矩阵的逆矩阵还是矩阵本身，即$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$。

方阵$\mathbf{A}$可逆的充分必要条件为行列式$|\mathbf{A}| \neq 0$。而且，如果$\mathbf{A}$可逆，则其逆矩阵$\mathbf{A}^{-1}$唯一。

假设方程(3.128)中的矩阵$\mathbf{A}$可逆，则在该方程两边同时左乘逆矩阵$\mathbf{A}^{-1}$可得：

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x}=\mathbf{A}^{-1}\mathbf{b} \quad \Rightarrow \quad \mathbf{I}\,\mathbf{x}=\mathbf{A}^{-1}\mathbf{b} \quad \Rightarrow \quad \mathbf{x}=\mathbf{A}^{-1}\mathbf{b} \tag{3.129}$$

矩阵求逆满足以下规则：

(1)$(\mathbf{A}^{-1})'=(\mathbf{A}')^{-1}$ (求逆与转置可交换次序)

(2)$(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$，$(\mathbf{ABC})^{-1}=\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ (求逆与乘积的混合运算)

3.3.10 矩阵的秩

考虑两个n维列向量$\mathbf{a}_1$与$\mathbf{a}_2$。如果$\mathbf{a}_1$正好是$\mathbf{a}_2$的固定倍数，则在向量组$\{\mathbf{a}_1, \mathbf{a}_2\}$中，真正包含信息的只是其中一个向量。

更一般地，考虑由K个n维向量构成的向量组$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$，如果存在$c_1, c_2, \cdots, c_K$不全为零，使得

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_K\mathbf{a}_K = \mathbf{0} \quad (3.130)$$

则称向量组$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$线性相关(linearly dependent)。

如果$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$线性相关，则其中至少有一个向量可写为其他向量的**线性组合**(linear combination), 也称**线性表出**。

反之，如果方程(3.130)必然意味着$c_1 = c_2 = \cdots = c_K = 0$，则称$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$**线性无关**(linearly independent)。

如果$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$线性相关，但从中去掉一个向量后，就变得线性无关，则$\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\}$中正好有$(K-1)$个向量真正含有信息，称$(K-1)$为此向量组的秩。

更一般地，向量组$\left\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\right\}$的极大线性无关部分组所包含的向量个数，称为该**向量组的秩**(rank)。

对于$m \times n$级矩阵$\mathbf{A}$，可将其n个列向量看成一个向量组，称此列向量组的秩为矩阵$\mathbf{A}$的**列秩**(column rank)。如果矩阵$\mathbf{A}_{m\times n}$的列秩正好等于n，则称矩阵$\mathbf{A}$**满列秩**(full column rank)。类似地，

可将矩阵$\mathbf{A}_{m\times n}$的 m 个行向量看成一个向量组，称此行向量组的秩为矩阵$\mathbf{A}$的**行秩**(row rank)。可以证明，任何矩阵$\mathbf{A}$的行秩与列秩一定相等，称为**矩阵的秩**(matrix rank)，记为$rank(\mathbf{A})$。

另外，n级可逆矩阵$\mathbf{A}$一定满秩，即$rank(\mathbf{A}) = n$。

进一步，可定义矩阵$\mathbf{A}$所对应的**列空间**(column space)与**行空间**(row space)：

$$\begin{aligned} \text{矩阵}\mathbf{A}\text{的列空间} &\equiv \text{col}(\mathbf{A}) \\ &\equiv \left\{\text{矩阵}\mathbf{A}\text{列向量的所有线性组合之集合}\right\} \end{aligned} \tag{3.131}$$

$$\text{矩阵}\mathbf{A}\text{的行空间} \equiv \text{row}(\mathbf{A})$$
$$\equiv \left\{\text{矩阵}\mathbf{A}\text{行向量的所有线性组合之集合}\right\} \tag{3.132}$$

更一般地，对于任意向量组$\left\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\right\}$，称其所有线性组合之集合为由该向量组**张成**(spanned)的空间，记为$Span\left\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K\right\}$。

矩阵$\mathbf{A}$的列空间由其列向量所张成，而矩阵$\mathbf{A}$的行空间由其行向量所张成。

3.3.11 正交矩阵

正交矩阵(orthogonal matrix)是一类性质非常好的方阵，它的每一列都相互正交，而且每个列向量的长度均被标准化为 1。

正交矩阵的转置就是其逆矩阵，即$\mathbf{A}^{-1}=\mathbf{A}'$。

记正交矩阵的列向量为$\mathbf{A}=\left(\mathbf{a}_1\cdots\mathbf{a}_n\right)$，则

$$\mathbf{A}'\mathbf{A}=\begin{pmatrix}\mathbf{a}_1'\\ \mathbf{a}_2'\\ \vdots\\ \mathbf{a}_n'\end{pmatrix}(\mathbf{a}_1\ \mathbf{a}_2\cdots\mathbf{a}_n)=\begin{pmatrix}\mathbf{a}_1'\mathbf{a}_1 & \mathbf{a}_1'\mathbf{a}_2 & \cdots & \mathbf{a}_1'\mathbf{a}_n\\ \mathbf{a}_2'\mathbf{a}_1 & \mathbf{a}_2'\mathbf{a}_2 & \cdots & \mathbf{a}_2'\mathbf{a}_n\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{a}_n'\mathbf{a}_1 & \mathbf{a}_n'\mathbf{a}_2 & \cdots & \mathbf{a}_n'\mathbf{a}_n\end{pmatrix}$$

$$=\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}=\mathbf{I}_n \tag{3.133}$$

3.3.12 矩阵的特征值与特征向量

考虑用n级方阵$\mathbf{A}$左乘n维非零列向量$\mathbf{x}$，则$\mathbf{Ax}$仍为n维列向量。一般来说，$\mathbf{Ax}$在n维空间的方向不一定与$\mathbf{x}$相同。

但如果$\mathbf{x}$为$\mathbf{A}$的特征向量，则$\mathbf{Ax}$与$\mathbf{x}$依然在同一条线上(方向相同或相反)，只是长度可能伸缩；而此伸缩比例即为特征值。

定义　如果存在数λ(可以是实数或复数)与n维非零列向量$\mathbf{x}$，满足

$$\mathbf{Ax} = \lambda \mathbf{x} \tag{3.134}$$

则称数λ为矩阵$\mathbf{A}$的**特征值**(eigenvalue)，而非零向量$\mathbf{x}$称为$\mathbf{A}$的对应于特征值λ的**特征向量**(eigenvector)。

特征值揭示了方阵的本质特征；例如，对角矩阵的特征值就是其主对角线上的各元素。

特征值与特征向量使得矩阵乘法“$\mathbf{Ax}$”变为简单的数乘“$\lambda\mathbf{x}$”，可简化矩阵运算，例如

$$\mathbf{A}^2\mathbf{x} = \mathbf{A}(\mathbf{Ax}) = \mathbf{A}(\lambda\mathbf{x}) = \lambda(\mathbf{Ax}) = \lambda^2\mathbf{x} \quad (3.135)$$

为求解特征值，将定义式(3.134)移项可得线性方程组：

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{x} = \mathbf{0} \quad (3.136)$$

其中，$\mathbf{I}_n$为n级单位矩阵。

如果矩阵 $(\mathbf{A}-\lambda\mathbf{I}_n)$ 可逆，则 $\mathbf{x}$ 存在唯一解 $\mathbf{x}=(\mathbf{A}-\lambda\mathbf{I}_n)^{-1}\mathbf{0}=\mathbf{0}$。

为得到非零的特征向量，$(\mathbf{A}-\lambda\mathbf{I}_n)$ 必须不可逆，故其行列式为 0：

$$\left|\mathbf{A}-\lambda\mathbf{I}_n\right|=0 \tag{3.137}$$

上式为关于 λ 的 n 次多项式方程，称为矩阵 $\mathbf{A}$ 的特征方程 (characteristic equation)。特征方程在复数范围内一定有 n 个解(包括重根)，即为矩阵 $\mathbf{A}$ 的特征值。得到具体的特征值 λ 后，代入方程(3.136)，即可通过求解线性方程组，得到相应的特征向量。

3.3.13 实对称矩阵的对角化与谱分解

虽然一般方阵的特征值可能为复数，但实对称矩阵(real symmetric matrix)的特征值一定为实数。进一步，n级实对称矩阵$\mathbf{A}$一定存在n个相互正交的特征向量。

这意味着，实对称矩阵$\mathbf{A}$一定可以对角化。

记实对称矩阵$\mathbf{A}$的特征值为$\lambda_1,\cdots,\lambda_n$(可以重复), 而相应的正交特征向量为$\mathbf{x}_1,\cdots,\mathbf{x}_n$。不失一般性，可将所有特征向量的长度标准化为 1，则实对称矩阵$\mathbf{A}$的特征向量矩阵$\mathbf{X}=(\mathbf{x}_1\cdots\mathbf{x}_n)$为正交矩阵。

根据特征值与特征向量的定义可知：

$$\mathbf{A}\mathbf{x}_k = \lambda_k \mathbf{x}_k \quad (k = 1, \cdots, n) \qquad (3.138)$$

上式共有n个方程$(k = 1, \cdots, n)$，更简洁地用矩阵表达：

$$\mathbf{A}(\mathbf{x}_1\ \mathbf{x}_2 \cdots \mathbf{x}_n) = (\mathbf{x}_1\ \mathbf{x}_2 \cdots \mathbf{x}_n)\underbrace{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}}_{\equiv \mathbf{\Lambda}}$$

$$(3.139)$$

由此可得，

$$\mathbf{AX} = \mathbf{X\Lambda} \qquad (3.140)$$

由于特征向量矩阵$\mathbf{X}$为正交矩阵，故$\mathbf{X}^{-1} = \mathbf{X}'$。在上式两边同时左乘$\mathbf{X}'$可得，

$$\mathbf{X}'\mathbf{AX} = \mathbf{\Lambda} \qquad (3.141)$$

在此，矩阵$\mathbf{A}$被**对角化**(diagonalized)为对角矩阵$\mathbf{\Lambda}$，而$\mathbf{\Lambda}$的主对角线元素即为$\mathbf{A}$的特征值。

另一方面，如果在方程(3.140)两边同时右乘$\mathbf{X}'$，则

$$\mathbf{A} = \mathbf{X\Lambda X}' = (\mathbf{x}_1\ \mathbf{x}_2 \cdots \mathbf{x}_n)\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}\begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

$$= \lambda_1 \mathbf{x}_1 \mathbf{x}_1' + \cdots + \lambda_n \mathbf{x}_n \mathbf{x}_n' = \sum_{k=1}^{n} \lambda_k \mathbf{x}_k \mathbf{x}_k' \tag{3.142}$$

上式将矩阵$\mathbf{A}$分解为$\sum_{k=1}^{n}\lambda_k \mathbf{x}_k \mathbf{x}_k'$，即$n$个外积$\mathbf{x}_k \mathbf{x}_k'$的加权之和，而权重为相应的特征值$\lambda_k$，这称为**谱分解**(spectral decomposition)。

注意到每个外积矩阵$\mathbf{x}_k \mathbf{x}_k'$均为简单的秩一矩阵，即$\mathbf{x}_k \mathbf{x}_k'$的秩为 1。

3.3.14 二次型

对于n维列向量$\mathbf{x}=(x_1\cdots x_n)'$，其 2-范数的平方就是内积：

$$\|\mathbf{x}\|_2^2=x_1^2+x_2^2+\cdots+x_n^2=\begin{pmatrix}x_1 & x_2 & \cdots & x_n\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix}=\mathbf{x}'\mathbf{x} \tag{3.143}$$

注意到上式也可写为

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{\mathbf{I}_n} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{x}'\mathbf{I}_n\mathbf{x} \tag{3.144}$$

方程(3.144)中的单位矩阵$\mathbf{I}_n$相当于给予此内积的每项相同权重。如允许不同的权重(比如$a_{11}x_1^2$)，并引入交叉项(比如$2a_{12}x_1x_2$)，则可使用任意对称矩阵$\mathbf{A}$，构成**二次型**(quadratic form)：

$$f(x_1, x_2, \cdots, x_n) = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{3.145}$$

其中，对称矩阵$\mathbf{A}$称为此二次型的矩阵。

所谓二次型，就是$x_1, x_2, \cdots, x_n$的二次齐次多项式函数：

$$
\begin{aligned}
f(x_1, x_2, \cdots, x_n) = a_{11}x_1^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1n}x_1x_n \\
+ \ a_{22}x_2^2 + \ \cdots \ + 2a_{2n}x_2x_n \\
+ \cdots \ \cdots \ \cdots \ \cdots \ \cdots \\
+ a_{nn}x_n^2
\end{aligned}
\tag{3.146}
$$

反之，任意二次型(3.146)，都可写为$\mathbf{x}'\mathbf{A}\mathbf{x}$的形式，其中$\mathbf{A}$为对称矩阵。

例如，考虑二维的二次型：

$$f(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \quad (3.147)$$

则此二次型可写为

$$f(x_1, x_2) = (x_1 \ \ x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (3.148)$$

其中，$a_{21} = a_{12}$(对称矩阵)。

如果$\mathbf{x}=\mathbf{0}$(零向量)，则二次型$\mathbf{x}'\mathbf{A}\mathbf{x}=0$。更有趣的情形是，当$\mathbf{x}\neq\mathbf{0}$时，二次型$\mathbf{x}'\mathbf{A}\mathbf{x}$如何取值？

首先，考虑一维的二次型：

$$f(x_1)=a_{11}x_1^2=x_1'a_{11}x_1 \qquad (3.149)$$

如果$a_{11}>0$，则只要$x_1\neq 0$，就有$f(x_1)=a_{11}x_1^2>0$。称此二次型为**正定**(positive definite)，其图形为开口向上的抛物线。

反之，如果$a_{11}<0$，则只要$x_1\neq 0$，就有$f(x_1)=a_{11}x_1^2<0$。此时，称此二次型为**负定**(negative definite)，其图形为开口向下的抛物线。

对于二维的二次型，其取值的确定性则更为复杂。例如，对于 x_1, x_2 不全为 0，二次型 $\left(x_1^2 + x_2^2\right)$ 一定为正，故为正定；二次型 $\left(-x_1^2 - x_2^2\right)$ 一定为负，故为负定；而二次型 $\left(x_1^2 - x_2^2\right)$ 则可正可负，称为**不定**(indefinite)。

但这依然没有穷尽所有的情形。考虑以下二次型：

$$f(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2 \qquad (3.150)$$

二次型$(x_1 + x_2)^2 \geq 0$(必然非负)；但即使x_1, x_2不全为 0，也可能出现$(x_1 + x_2)^2 = 0$，只要$x_1 = -x_2$；比如，$x_1 = 1$而$x_2 = -1$。此时，称此二次型为**半正定**(positive semidefinite)。

另一方面，二次型$-(x_1 + x_2)^2 \leq 0$(必然非正)；但即使x_1, x_2不全为 0，也可能出现$(x_1 + x_2)^2 = 0$，只要$x_1 = -x_2$。此时，称此二次型为**半负定**(negative semidefinite)。

在一般的n维情况下，给定对称矩阵$\mathbf{A}$，针对二次型$\mathbf{x}'\mathbf{A}\mathbf{x}$的取值确定性，可引入以下定义。

(1) 对于任意非零列向量$\mathbf{x}$，都有$\mathbf{x}'\mathbf{A}\mathbf{x} > 0$，则对称矩阵$\mathbf{A}$为**正定矩阵**(positive definite)。

(2) 对于任意非零列向量$\mathbf{x}$，都有$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$，则对称矩阵$\mathbf{A}$为**半正定矩阵**(positive semidefinite)。

(3) 对于任意非零列向量$\mathbf{x}$，都有$\mathbf{x}'\mathbf{A}\mathbf{x} < 0$，则对称矩阵$\mathbf{A}$为**负定矩阵**(negative definite)。

(4) 对于任意非零列向量$\mathbf{x}$，都有$\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$，则对称矩阵$\mathbf{A}$为**半负定矩阵**(negative semidefinite)。

正定矩阵一定半正定，而负定矩阵也一定半负定。

直观上，如果对称矩阵$\mathbf{A}$为正定矩阵，则该矩阵可通过线性变换(转换坐标系)，变为一个主对角线元素全为正数的对角矩阵；而这些主对角线元素正好是矩阵$\mathbf{A}$的特征值。

因此，正定矩阵$\mathbf{A}$一定可逆，即逆矩阵$\mathbf{A}^{-1}$存在。线性变换后的正定二次型可写为

$$f(x_1, x_2, \cdots, x_n) = \alpha_{11} x_1^2 + \alpha_{22} x_2^2 + \cdots + \alpha_{nn} x_n^2 \quad (3.151)$$

其中，$\alpha_{11}, \cdots, \alpha_{nn}$全部为正数，故当$x_1, \cdots, x_n$不全为 0 时，

$f(x_1, x_2, \cdots, x_n)$必然大于 0。

如果$\alpha_{11}, \cdots, \alpha_{nn}$全部为正数或 0，则$\mathbf{A}$为半正定矩阵。

如果$\alpha_{11}, \cdots, \alpha_{nn}$全部为负数，则$\mathbf{A}$为负定矩阵。如果$\alpha_{11}, \cdots, \alpha_{nn}$全部为负数或 0，则$\mathbf{A}$为半负定矩阵。

反之，如果$\alpha_{11}, \cdots, \alpha_{nn}$有正有负，则$\mathbf{A}$为不定的(indefinite)，其二次型$\mathbf{x}'\mathbf{A}\mathbf{x}$的取值可正可负。

命题 3.3 对于任意矩阵$\mathbf{A}$，$\mathbf{A}'\mathbf{A}$为半正定矩阵。

证明： 首先，由于$\mathbf{A}'\mathbf{A} = (\mathbf{A}'\mathbf{A})'$，故$\mathbf{A}'\mathbf{A}$为对称矩阵。

不失一般性，假设$\mathbf{A}'\mathbf{A}$为n级矩阵。其次，对于任意n维非零列向量$\mathbf{x}$，二次型

$$\mathbf{x}'(\mathbf{A}'\mathbf{A})\mathbf{x} = (\mathbf{x}'\mathbf{A}')(\mathbf{A}\mathbf{x}) = \underbrace{(\mathbf{A}\mathbf{x})'\mathbf{A}\mathbf{x}}_{\text{平方和}} \geq 0 \qquad (3.152)$$

因此，$\mathbf{A}'\mathbf{A}$为半正定矩阵。

3.4 概率统计

3.4.1 概率

如果天气预报说“明天 70%的概率下雨。”这意味着什么？

直观地，可将“概率”理解为在大量重复试验下，事件发生的频率趋向的某个稳定值。记事件“下雨”为A，其发生的**概率**(probability)记为$\mathrm{P}(A)$。

称随机试验的所有可能结果为“样本空间”(sample space)，记为S；称其中的每个可能结果为“样本点”(sample point)。

称样本空间S的某个子集为“随机事件”(random event)，简称“事件”(event)。

例 在一个掷骰子的随机试验中，样本空间为 $S=\{1,2,3,4,5,6\}$，而“偶数朝上”的随机事件可写为 $A=\{2,4,6\}$。如果骰子是公平的(a fair dice)，则每个数字朝上的概率相等，故 $\mathrm{P}(A)=0.5$。

3.4.2 条件概率

例 已知明天会出太阳，则下雨的概率有多大？

记事件“出太阳”为 B，则在出太阳的前提条件下，降雨的**条件概率**(conditional probability)为

$$\mathrm{P}(A|B) \equiv \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \qquad (3.153)$$

其中，AB表示事件A与B同时发生(即交集，也记为$A \cap B$)，故$\mathrm{P}(AB)$为“太阳雨”的概率，参见图 3.12 (其中最大的矩形区域表示样本空间 S)。

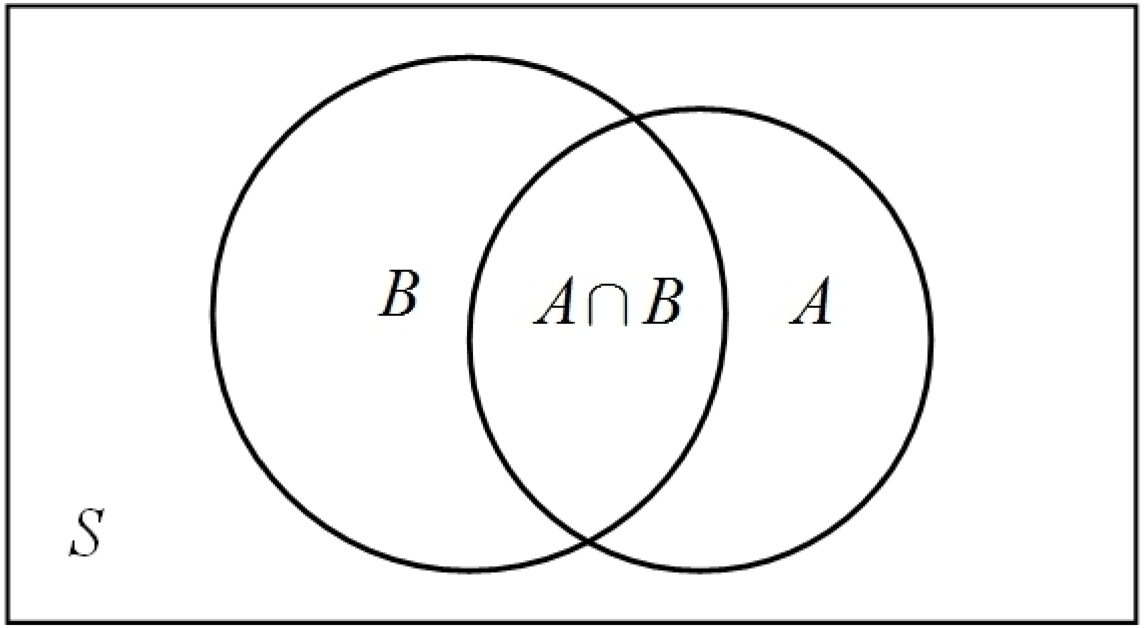

图 3.12　条件概率的示意图

例 股市崩盘的可能性为无条件概率；在已知经济已陷入严重衰退的情况下，股市崩盘的可能性则为条件概率。

根据条件概率公式(3.153)，可得如下乘法公式：

$$\mathrm{P}(AB) = \mathrm{P}(A|B)\,\mathrm{P}(B) \qquad (3.154)$$

这表明，A 与 B 同时发生的概率 $\mathrm{P}(AB)$，等于 B 发生的概率 $\mathrm{P}(B)$，乘以在 B 发生情况下，A 发生的条件概率 $\mathrm{P}(A|B)$。

3.4.3 独立事件

如果条件概率等于无条件概率，$\mathrm{P}(A|B)=\mathrm{P}(A)$，即$B$是否发生不影响$A$的发生，则称$A, B$为相互独立的随机事件。

此时，$\mathrm{P}(A|B) \equiv \dfrac{\mathrm{P}(AB)}{\mathrm{P}(B)} = \mathrm{P}(A)$，故

$$\mathrm{P}(AB)=\mathrm{P}(A)\,\mathrm{P}(B) \quad (3.155)$$

也可将此式作为独立事件的定义。

3.4.4 全概率公式

我们经常需要考虑将世界分为若干个互相排斥的状态。

定义 设 S 为随机试验的样本空间。如果事件组 $\{B_1, B_2, \cdots, B_n\}$ 两两互不相容($B_i \cap B_j = \varnothing, i \neq j$)，但必有某个事件 B_i 发生($B_1 \cup B_2 \cup \cdots \cup B_n = S$，其中"$\cup$"表示事件的并集)，则称 $\{B_1, B_2, \cdots, B_n\}$ 为样本空间 S 的一个**划分** (partition)，也称**完备事件组**。

命题 3.4（全概率公式） 设事件组$\{B_1, B_2, \cdots, B_n\}$为样本空间 S 的一个划分，且每个事件的发生概率均为正数（$P(B_i) > 0, i = 1, \cdots, n$），则对任何事件 A（无论 A 与$\{B_1, B_2, \cdots, B_n\}$是否有任何关系)，都有

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(A|B_i)\,\mathrm{P}(B_i) \qquad (3.156)$$

全概率公式(Law of Total Probability)把世界分成n种可能的情形$\{B_1, B_2, \cdots, B_n\}$，再把每种情况下的条件概率$\mathrm{P}(A|B_i)$“加权平均”而汇总成无条件概率$\mathrm{P}(A)$，而权重就是每种情形发生的概率$\mathrm{P}(B_i)$。

证明

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(AS) \qquad (S\text{ 为全集}) \\
&= \mathrm{P}\left[A(B_1 \cup B_2 \cup \cdots \cup B_n)\right] \quad (\{B_1, B_2, \cdots, B_n\}\text{为划分}) \\
&= \mathrm{P}\left[AB_1 \cup AB_2 \cup \cdots \cup AB_n\right] \quad (\text{乘法分配律}) \\
&= \mathrm{P}(AB_1) + \mathrm{P}(AB_2) + \cdots + \mathrm{P}(AB_n) \quad (AB_i\text{ 与 }AB_j\text{ 互斥})
\end{aligned}$$

$$= \mathrm{P}(A \mid B_1)P(B_1) + \mathrm{P}(A \mid B_2)\mathrm{P}(B_2) + \cdots + \mathrm{P}(A \mid B_n)\mathrm{P}(B_n)$$

(乘法公式)

$$= \sum_{i=1}^{n} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i)$$

3.4.5 贝叶斯公式

从条件概率公式(3.153)出发，再使用乘法公式(3.154)，即可得**贝叶斯公式**(Bayes Theorem)：

$$\mathrm{P}(A|B)=\frac{\mathrm{P}(AB)}{\mathrm{P}(B)}=\frac{\mathrm{P}(B\,|\,A)\,\mathrm{P}(A)}{\mathrm{P}(B)} \tag{3.157}$$

其中，$\mathrm{P}(A)$可理解为“先验概率”(prior probability)，B可理解为“数据”(data)，而$\mathrm{P}(A|B)$则为“后验概率”(posterior probability)。

贝叶斯公式给出在看到数据B后，将先验概率$\mathrm{P}(A)$更新为后验概率$\mathrm{P}(A|B)$的“贝叶斯规则”(Bayes rule)。

进一步，如果设事件组$\{B_1, B_2, \cdots, B_n\}$为样本空间$S$的一个划分，可使用贝叶斯公式计算$\mathrm{P}(B_i|A)$：

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\sum_{j=1}^{n}\mathrm{P}(A|B_j)\mathrm{P}(B_j)} \tag{3.158}$$

其中，根据全概率公式(3.156)，$\mathrm{P}(A) = \sum_{j=1}^{n}\mathrm{P}(A|B_j)\mathrm{P}(B_j)$。

3.4.6 离散型概率分布

假设随机变量X的可能取值为$\{x_1, x_2, \cdots, x_k, \cdots\}$，其对应概率为$\{p_1, p_2, \cdots, p_k, \cdots\}$，即$p_k \equiv \mathrm{P}(X = x_k)$，则称$X$为离散型随机变量，其分布律可表示为

$$\begin{array}{cccccc} X & x_1 & x_2 & \cdots & x_k & \cdots \\ p & p_1 & p_2 & \cdots & p_k & \cdots \end{array} \tag{3.159}$$

其中，$p_k \geq 0$，且$\sum_k p_k = 1$。常见的离散分布有两点分布(Bernoulli distribution)、二项分布(Binomial distribution)、泊松分布(Poisson distribution)等。

3.4.7 连续型概率分布

连续型随机变量可以取任意实数，其**概率密度函数**(probability density function，简记 pdf) $f(x)$满足，

(1) $f(x) \geq 0,\ \forall x$；

(2) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$；

(3) X落入区间$[a,b]$的概率为$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$。

定义**累积分布函数**(cumulative distribution function，简记 cdf)：

$$F(x) \equiv \mathrm{P}(-\infty < X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \quad (3.160)$$

其中，t 为积分变量，参见图 3.13。

常见的连续型概率分布包括均匀分布(uniform distribution)、正态分布(normal distribution)、t分布、χ^2分布与F分布等。

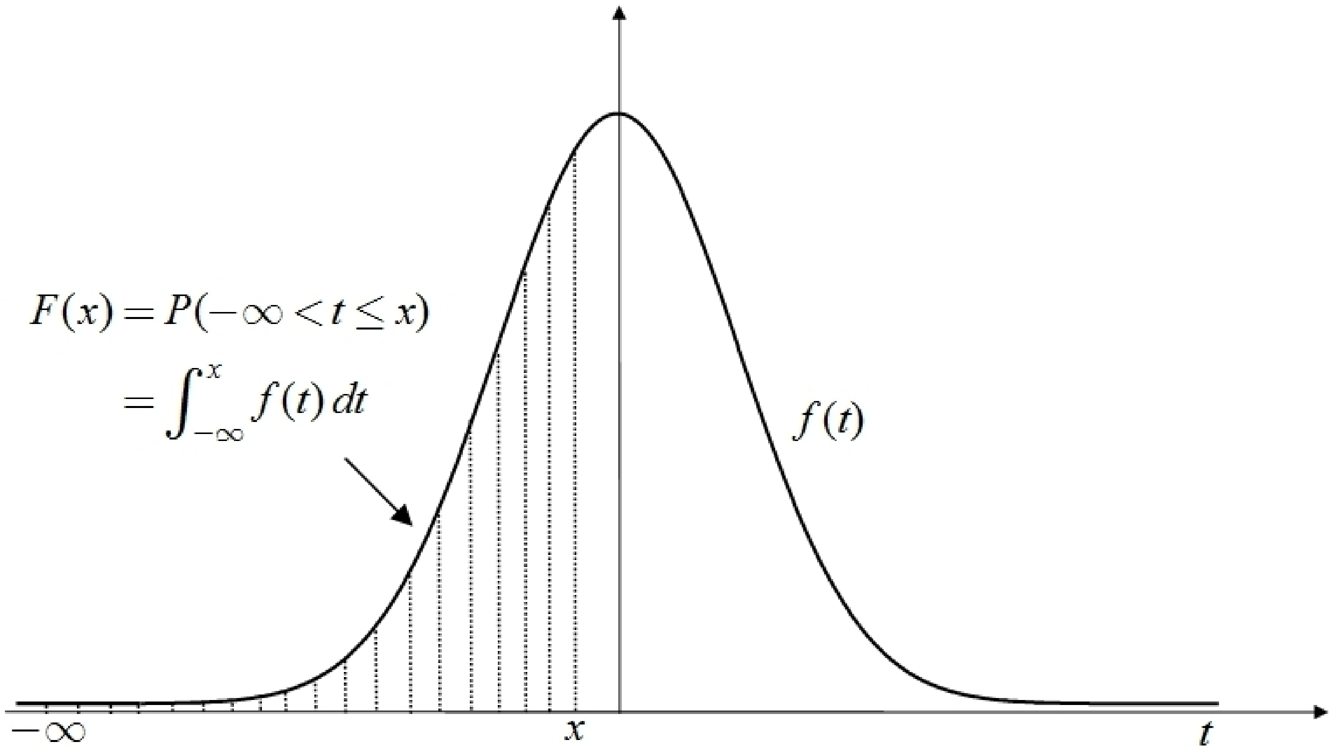

图 3.13 累积分布函数的示意图

3.4.8 多维随机向量的概率分布

为研究变量间的关系，常同时考虑两个或多个随机变量，即**随机向量**(random vector)。二维连续型随机向量(X, Y)的**联合密度函数**(joint pdf) $f(x,y)$满足，

(i) $f(x,y) \geq 0,\ \forall x,\ y$；

(ii) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$；

(iii) (X, Y) 落入平面某区域 D 的概率为 $\mathrm{P}\left\{(X,Y)\in D\right\} = \iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y$。

二维随机向量的联合密度函数就像倒扣的草帽。落入平面某区域D的概率就是此草帽下在区域D之上的体积，参见图 3.14。

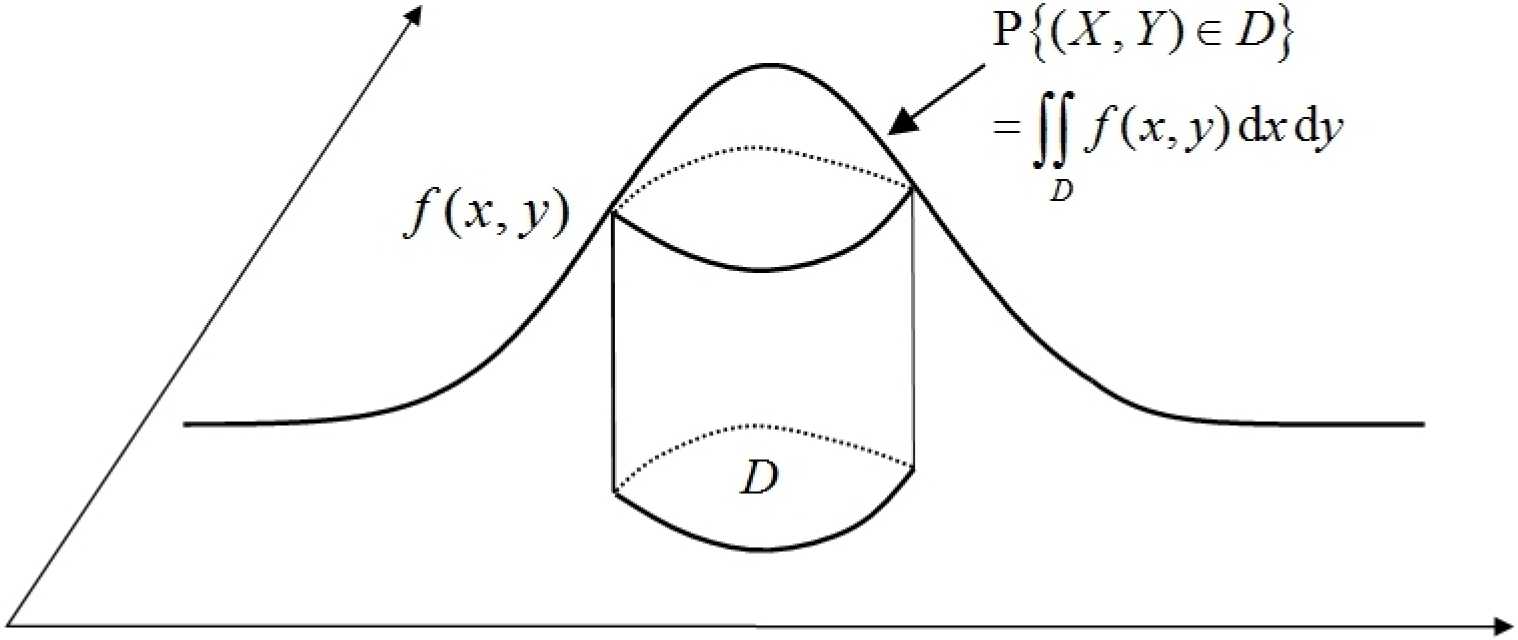

图 3.14 二维联合密度函数的示意图

更一般地，n维连续型随机向量$(X_1, X_2, \cdots, X_n)$可由联合密度函数 $f(x_1, x_2, \cdots, x_n)$来描述。

从二维联合密度 $f(x, y)$，可计算 X 的(一维)**边缘密度函数**(marginal pdf)：

$$f_x(x) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}y \qquad (3.161)$$

即给定 $X = x$，把所有 Y 取值的可能性都“加总”起来(积分的本质就是加总)。

类似地，可以计算Y的(一维)边缘密度函数：

$$f_y(y) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}x \qquad (3.162)$$

即给定$Y = y$，把所有X取值的可能性都“加总”起来。

二维随机向量(X, Y)的**累积分布函数**定义为

$$F(x, y) \equiv \mathrm{P}(-\infty < X \leq x; -\infty < Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t, s)\,\mathrm{d}t\,\mathrm{d}s \qquad (3.163)$$

其中，t与s为积分变量。

3.4.9 条件分布

考虑在$X=x$条件下Y的条件分布，记为$Y \big| X=x$或$Y \big| x$。

对于连续型分布，此条件分布相当于在“草帽”(联合密度函数)上$X=x$的位置垂直地切一刀所得的截面。

由于X为连续型随机变量，事件$\{X=x\}$发生的概率为 0。应如何计算$Y \big| X=x$的条件概率密度(conditional pdf)？

为此，考虑 x 附近的小邻域 $\left[x-\varepsilon, x+\varepsilon\right]$，计算在 $X \in \left[x-\varepsilon, x+\varepsilon\right]$ 条件下 Y 的累积分布函数，即 $\mathrm{P}\left\{Y \leq y \mid X \in \left[x-\varepsilon, x+\varepsilon\right]\right\}$(参见图 3.15)，然后让 $\varepsilon \to 0^{+}$，则可证明条件密度函数为

$$f(y \mid x)=\frac{f(x, y)}{f_{x}(x)} \tag{3.164}$$

直观上，此公式与条件概率公式(3.153)类似。

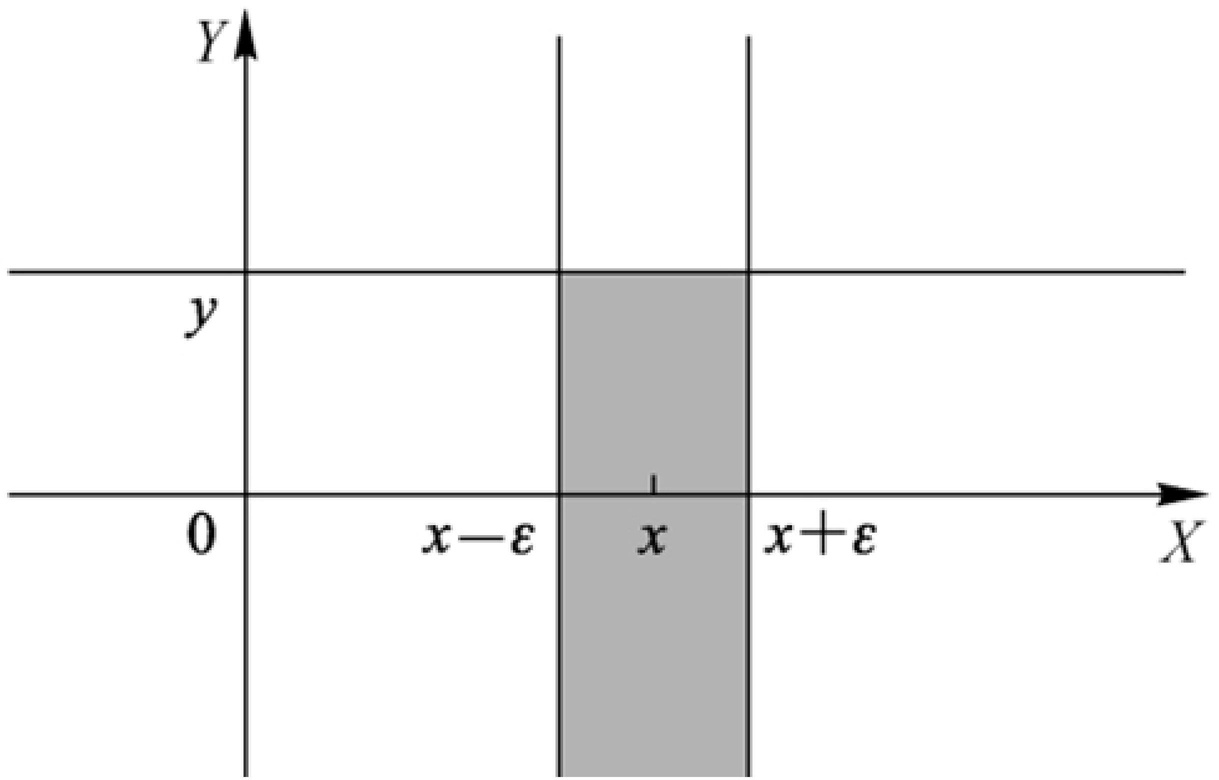

图 3.15　条件密度函数的计算

3.4.10 随机变量的数字特征

定义 对于分布律为$p_k \equiv \mathrm{P}(X = x_k)$的离散型随机变量$X$，其**期望**(expectation)为

$$\mathrm{E}(X) \equiv \mu \equiv \sum_{k=1}^{\infty} x_k p_k \qquad (3.165)$$

由上式可知，期望的直观含义是对x_k进行加权平均，而权重为概率p_k。

定义　对于概率密度函数为$f(x)$的连续型随机变量X，其**期望**为

$$\mathrm{E}(X) \equiv \mu \equiv \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \qquad (3.166)$$

直观上，上式也是对x进行加权平均，而权重为概率密度$f(x)$。

有时称求期望这种运算为**期望算子**(expectation operator)。容易证明，期望算子满足**线性性**(linearity)，即对于任意常数k都有

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y), \quad \mathrm{E}(kX) = k\,\mathrm{E}(X) \qquad (3.167)$$

定义 随机变量 X 的**方差**(variance)为

$$\mathrm{Var}(X) \equiv \sigma^2 \equiv \mathrm{E}\left[X - \mathrm{E}(X)\right]^2 \quad (3.168)$$

方差越大，则随机变量取值的波动幅度越大。称方差的平方根为**标准差**(standard deviation)，通常记为σ。

在计算方差时，可利用以下简便公式：

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2 \quad (3.169)$$

我们常需考虑两个变量之间的相关性，即一个随机变量的取值会对另一随机变量的取值有多大影响。

定义 随机变量 X 与 Y 的**协方差**(covariance)为

$$\mathrm{Cov}(X,Y) \equiv \sigma_{XY} \equiv \mathrm{E}\Big[\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big] \tag{3.170}$$

如果当随机变量 X 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时，随机变量 Y 的取值也倾向于大于(小于)其期望值 $\mathrm{E}(Y)$，则 $\mathrm{Cov}(X,Y)>0$，二者存在**正相关**；反之，如果当随机变量 X 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时，随机变量 Y 的取值反而倾向于小于(大于)其期望值 $\mathrm{E}(Y)$，则 $\mathrm{Cov}(X,Y)<0$，二者存在**负相关**。

如果$\mathrm{Cov}(X,Y)=0$，则说明二者**线性不相关**(uncorrelated)；但不一定**相互独立**(independent)，因为二者还可能存在非线性的相关关系。

在计算协方差时，可使用以下简便公式：

$$\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y) \quad (3.171)$$

协方差的运算也满足线性性，可以证明：

$$\mathrm{Cov}(X,Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z) \quad (3.172)$$

协方差的缺点是，它受X与Y度量单位的影响。为将其标准化，引入相关系数的定义。

定义 随机变量X与Y的**相关系数**(correlation)为

$$\rho \equiv \mathrm{Corr}(X, Y) \equiv \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{3.173}$$

相关系数一定介于-1与 1 之间，即$-1 \le \rho \le 1$。

若以上各定义式中的积分不收敛，则随机变量的数字特征可能不存在。比如，自由度为 1 的t分布变量，其期望与方差都不存在。

定义 条件期望(conditional expectation)就是条件分布$Y \mid x$的期望，即

$$\mathrm{E}(Y \mid X = x) \equiv \mathrm{E}(Y \mid x) = \int_{-\infty}^{+\infty} y f(y \mid x) \mathrm{d}y \quad (3.174)$$

在上式中，由于y已被积分积掉，故$\mathrm{E}(Y \mid x)$只是x的函数，参见图 3.16。

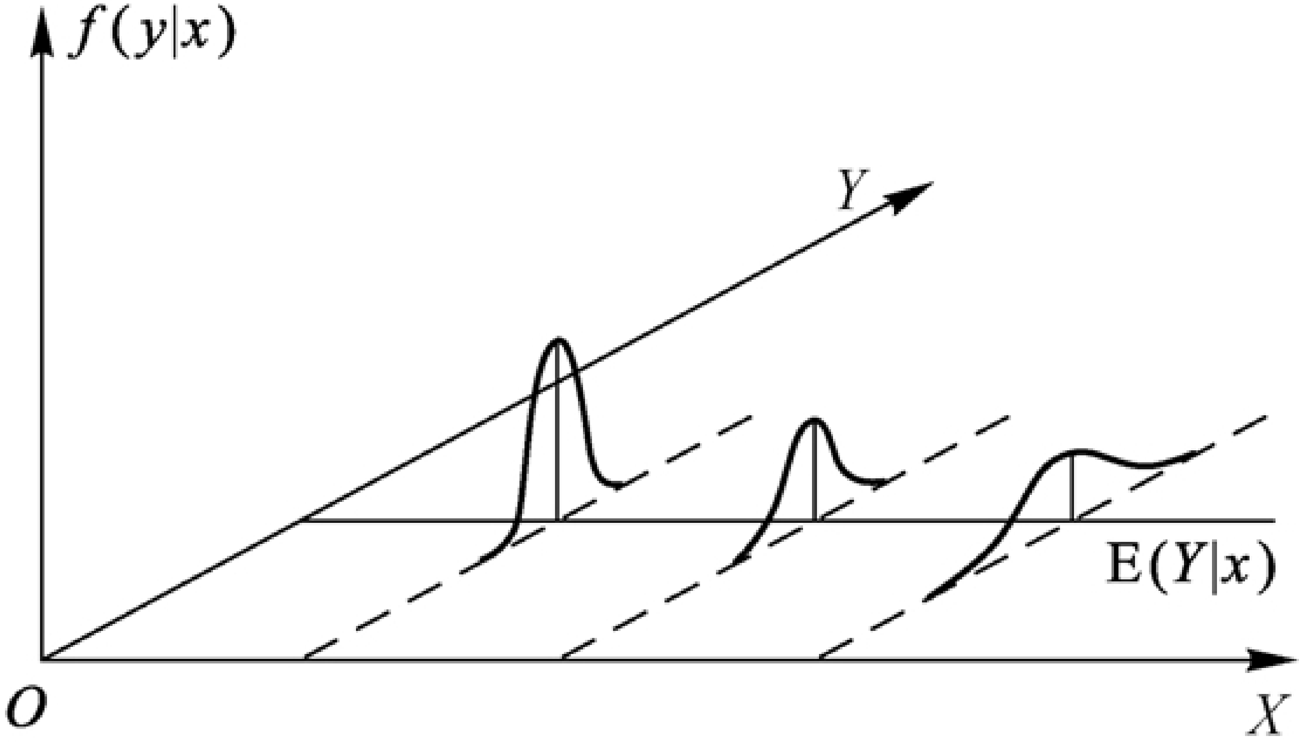

图 3.16　条件期望与条件方差的示意图

定义　条件方差(conditional variance)就是条件分布$Y \mid x$的方

差，即

$$\begin{aligned}\operatorname{Var}(Y \mid X=x) &\equiv \operatorname{Var}(Y \mid x) \\ &= \int_{-\infty}^{+\infty}\left[y-\mathrm{E}(Y \mid x)\right]^{2} f(y \mid x)\,\mathrm{d}y\end{aligned} \tag{3.175}$$

在上式中，y 已被积分积掉，故 $\operatorname{Var}(Y \mid x)$ 也只是 x 的函数，参见图 3.16。

定义 设$\mathbf{X} = (X_1 \; X_2 \; \cdots \; X_n)'$为$n$维随机向量，则其**协方差矩阵**(covariance matrix)为$n \times n$的对称矩阵：

$$
\begin{aligned}
\mathrm{Var}(\mathbf{X}) &\equiv \mathrm{E}\left[\big(\mathbf{X}-\mathrm{E}(\mathbf{X})\big)\big(\mathbf{X}-\mathrm{E}(\mathbf{X})\big)'\right] \\
&= \mathrm{E}\left[\begin{pmatrix} X_1-\mathrm{E}(X_1) \\ \vdots \\ X_n-\mathrm{E}(X_n) \end{pmatrix}\big(X_1-\mathrm{E}(X_1) \quad \cdots \quad X_n-\mathrm{E}(X_n)\big)\right] \\
&= \mathrm{E}\begin{pmatrix} \big[X_1-\mathrm{E}(X_1)\big]^2 & \cdots & \big[X_1-\mathrm{E}(X_1)\big]\big[X_n-\mathrm{E}(X_n)\big] \\ \vdots & \ddots & \vdots \\ \big[X_1-\mathrm{E}(X_1)\big]\big[X_n-\mathrm{E}(X_n)\big] & \cdots & \big[X_n-\mathrm{E}(X_n)\big]^2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}
\end{aligned}
\tag{3.176}
$$

其中，主对角线元素$\sigma_{ii} \equiv \mathrm{Var}(X_i)$，非主对角线元素$\sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j)$。

对于随机向量$\mathbf{X}$的期望与协方差矩阵的运算，有如下重要法则。假设$\mathbf{A}$为$m \times n$常数矩阵(不含随机变量)，则可证明：

(1)$\mathrm{E}(\mathbf{AX}) = \mathbf{A}\,\mathrm{E}(\mathbf{X})$ (期望为线性算子)

(2)$\mathrm{Var}(\mathbf{X}) = \mathrm{E}(\mathbf{XX}') - \mathrm{E}(\mathbf{X})\left[\mathrm{E}(\mathbf{X})\right]'$ (一维公式的推广)

(3)$\mathrm{Var}(\mathbf{AX}) = \mathbf{A}\,\mathrm{Var}(\mathbf{X})\mathbf{A}'$ (夹心估计量)

如果$\mathbf{A}$为对称矩阵，则$\mathrm{Var}(\mathbf{AX}) = \mathbf{A}\,\mathrm{Var}(\mathbf{X})\mathbf{A}$，称为**夹心估计量** (sandwich estimator)，其中两边的$\mathbf{A}$为“面包”，而夹在中间的$\mathrm{Var}(\mathbf{X})$为“菜”，在形式上类似于三明治。

使用夹心估计量的公式可以证明，协方差矩阵必然为半正定矩阵(positive semidefinite)。

在一维情况下，这意味着随机变量的方差必然为非负。

命题 3.5 n维随机向量$\mathbf{X}$的协方差矩阵$\mathrm{Var}(\mathbf{X})$为半正定矩阵。

证明：根据协方差矩阵的定义，$\mathrm{Var}(\mathbf{X})$为$n \times n$对称矩阵。对于n维非零列向量$\mathbf{c}$，随机变量$\mathbf{c}'\mathbf{X}$(即$\mathbf{X}$各分量的线性组合)的方差必然非负。因此，

$$\mathrm{Var}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\,\mathrm{Var}(\mathbf{X})\mathbf{c} \geq 0 \tag{3.177}$$

根据半正定矩阵的定义，$\mathrm{Var}(\mathbf{X})$为半正定矩阵。

进一步，如果假定$\mathbf{X}$的各分量之间不存在线性依赖的关系，则任意线性组合$\mathbf{c}'\mathbf{X}$的方差均为正数，故$\mathrm{Var}(\mathbf{X})$为正定矩阵。

3.4.11 迭代期望定律

命题 3.6（迭代期望定律） 对于条件期望的运算，有以下重要的“迭代期望定律”(Law of Iterated Expectation)：

$$\mathrm{E}(Y)=\mathrm{E}_X\left[\mathrm{E}(Y\,|\,x)\right] \quad (3.178)$$

上式表明，无条件期望$\mathrm{E}(Y)$，等于给定$X=x$情况下Y的条件期望$\mathrm{E}(Y\,|\,x)$，再对X求期望。

下面以连续型变量为例证明。

证明： 等式(3.178)的右边可写为

$$\mathrm{E}_X\left[\mathrm{E}(Y\mid x)\right]=\mathrm{E}_X\left[\int_{-\infty}^{+\infty} y\frac{f(x,y)}{f_x(x)}\mathrm{d}y\right] \quad \text{(条件期望的定义)}$$

$$=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} y\frac{f(x,y)}{f_x(x)}\mathrm{d}y\right]f_x(x)\,\mathrm{d}x \quad \text{(期望的定义)}$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y \qquad \text{(消去 } f_x(x)\text{)}$$

$$=\int_{-\infty}^{+\infty} y\left[\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\right]\mathrm{d}y \quad \text{(边缘密度 } f_y(y)\text{ 的定义)}$$

$$=\int_{-\infty}^{+\infty} y\,f_y(y)\,\mathrm{d}y=\mathrm{E}(Y) \quad \text{(期望 } \mathrm{E}(Y)\text{ 的定义)}$$

在精神上，迭代期望定律类似于全概率公式。

直观地，无条件期望等于条件期望之加权平均，而权重为条件“$X = x$”的概率密度(取值可能性)。

在离散随机变量的情形下，可看得更为清楚：

$$\mathrm{E}(Y) = \sum_{x_i} \mathrm{P}(X = x_i)\,\mathrm{E}(Y \mid x_i) \quad (3.179)$$

例 全部同学的平均成绩，等于男生的平均成绩与女生的平均成绩之加权平均，而权重则为男女生在全班人数中的比重。

推而广之，对于任意函数$g(\cdot)$，均可得到：

$$\mathrm{E}\big[g(Y)\big] = \mathrm{E}_X\, \mathrm{E}[g(Y)\big|x] \quad (3.180)$$

有时期望算子E_X的下标被省去，此时需注意对什么变量求期望。

3.4.12 随机变量无关的三个层次概念

定义 对于连续型随机变量X与Y，如果其联合密度等于边缘密度的乘积，即$f(x, y) = f_x(x) f_y(y)$，则称X与Y相互独立。

直观上，如果X与Y相互独立，则X的取值不对Y的取值产生任何影响，反之亦然。这是有关随机变量“无关”的最强概念。

线性不相关的概念则更弱，仅要求协方差为 0，即 $\mathrm{Cov}(X,Y)=0$。“相互独立”意味着“线性不相关”，但反之不然。

在二者之间还有一个中间层次的无关概念，即“均值独立”(mean-independent)，在统计学中很有用。

定义 假设条件期望 $\mathrm{E}(Y\,|\,x)$ 存在。如果 $\mathrm{E}(Y\,|\,x)$ 不依赖于 X，则称“Y **均值独立**于 X”(Y is mean-independent of X)。

均值独立不是一种对称的关系，即“Y 均值独立于 X”并不意味着“X 均值独立于 Y”。

命题 3.7 “Y 均值独立于 X” 当且仅当 $\mathrm{E}(Y \mid x) = \mathrm{E}(Y)$(即条件期望等于无条件期望)。

证明：

(1)假设“Y 均值独立于 X”，则 $\mathrm{E}(Y \mid x)$ 不依赖于 X，故 $\mathrm{E}_X\left[\mathrm{E}(Y \mid x)\right] = \mathrm{E}(Y \mid x)$。根据迭代期望定律，$\mathrm{E}(Y) = \mathrm{E}_X\left[\mathrm{E}(Y \mid x)\right] = \mathrm{E}(Y \mid x)$。

(2) 假设 $\mathrm{E}(Y \mid x) = \mathrm{E}(Y)$，则显然 $\mathrm{E}(Y \mid x)$ 不依赖于 X，故 Y 均值独立于 X。

命题 3.8 如果X与Y相互独立，则Y均值独立于X，且X均值独立于Y。

X与Y相互独立意味着，X与Y一点关系也没有，故条件期望$\mathrm{E}(Y \mid x)$也不会依赖于X。证明参见习题。

命题 3.9（均值独立意味着不相关） 如果Y均值独立于X或X均值独立于Y，则$\mathrm{Cov}(X,Y)=0$。

证明：

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{E}\Big[\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big] && \text{(协方差的定义)} \\
&= \mathrm{E}_X\,\mathrm{E}_Y\Big[\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big| x\Big] && \text{(迭代期望定律)}
\end{aligned}
$$

$$= \mathrm{E}_X\left[\left(X - \mathrm{E}(X)\right)\mathrm{E}_Y\left(Y - \mathrm{E}(Y)\middle| x\right)\right]$$

(将 $X - \mathrm{E}(X)$ 视为常数提出)

$$= \mathrm{E}_X\left[\left(X - \mathrm{E}(X)\right)\left(\mathrm{E}(Y \mid x) - \mathrm{E}(Y)\right)\right]$$

(期望算子的线性性)

$$= \mathrm{E}_X\left[\left(X - \mathrm{E}(X)\right)\cdot 0\right] = 0$$

(均值独立的定义)

总之，“相互独立” $\Rightarrow$ “均值独立” $\Rightarrow$ “线性不相关”。

3.4.13 正态分布

最常用的连续型概率分布为正态分布。如果随机变量X的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \tag{3.181}$$

则称X服从**正态分布**(normal distribution)或**高斯分布**(Gaussian distribution)，记为$X \sim N(\mu,\sigma^2)$，其中μ为期望，而σ^2为方差。将X进行标准化，定义$Z\equiv\dfrac{X-\mu}{\sigma}$，则$Z$服从**标准正态分布**(standard normal distribution)，记为$Z \sim N(0,1)$，其概率密度函数为

$$\phi(x)=\frac{1}{\sqrt{2\pi}}\exp\left\{-x^2/2\right\} \tag{3.182}$$

标准正态分布的概率密度以原点为对称，呈钟形(bell-shaped)，通常记为$\phi(x)$；其累积分布函数则记为$\Phi(x)$。在 Python 中画标准正态分布的密度图，可输入如下命令：

```
In [1]: from scipy.stats import norm
   ...: x= np.arange(-4, 4, 0.01)
   ...: plt.plot(x, norm.pdf(x, 0, 1))
   ...: plt.axvline(0, linewidth=1)
   ...: plt.title('Standard Normal Density')
```

其中，第 1 个命令从 Scipy 模块的 stats 子模块导入 `norm` 类(class)，第 2 个命令设定画图区间与网格，第 3 个命令使用方法

`norm.pdf(x, 0, 1)`得到标准正态的密度函数，而第 4 个命令在 0 的位置画一条垂直线(vertical line)，且线宽(line width)为 1，结果参见图 3.17。

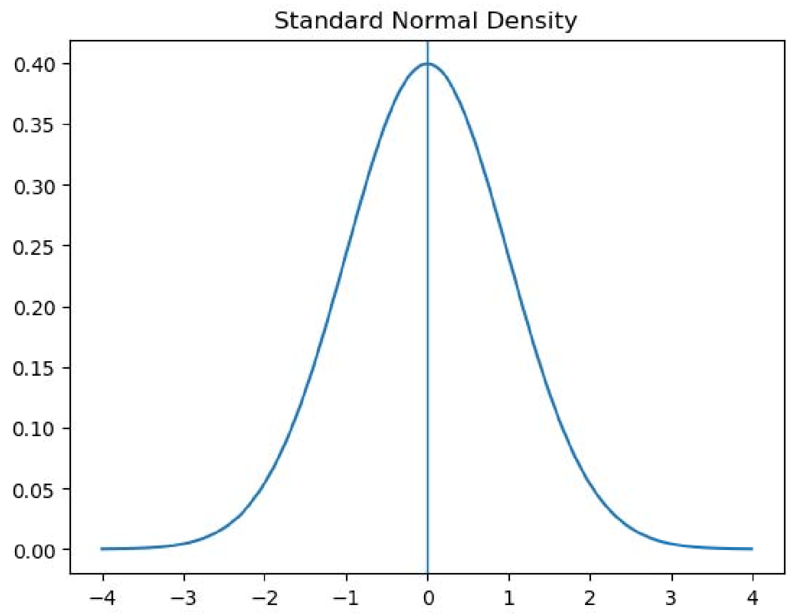

图 3.17 标准正态分布的密度图

如果n维随机向量$\mathbf{X} = (X_1 \; X_2 \cdots X_n)'$的联合密度函数为

$$f(x_1, \cdots, x_n) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\} \tag{3.183}$$

则称$\mathbf{X}$服从期望为$\boldsymbol{\mu}$、协方差矩阵为$\boldsymbol{\Sigma}$的n**维正态分布**，记为$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。在上式中，$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$为$(\mathbf{X} - \boldsymbol{\mu})$的二次型，其二次型矩阵为协方差矩阵的逆矩阵$\boldsymbol{\Sigma}^{-1}$；而$|\boldsymbol{\Sigma}|$为协方差矩阵$\boldsymbol{\Sigma}$的行列式。

3.4.14 最大似然估计

最大似然估计法(Maximum Likelihood Estimation，简记 MLE)是统计学的常用方法。

假设随机变量Y的概率密度函数为$f(Y;\boldsymbol{\theta})$，其中$\boldsymbol{\theta}$为未知参数向量。

为估计$\boldsymbol{\theta}$，从Y的总体中抽取样本容量为n的随机样本$\{Y_1, \cdots, Y_n\}$。

假设$\left\{y_1, \cdots, y_n\right\}$为独立同分布(independently and identically distributed，简记 iid)，则样本数据的联合密度函数为

$$f(Y_1;\boldsymbol{\theta})\, f(Y_2;\boldsymbol{\theta}) \cdots f(Y_n;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta}) \tag{3.184}$$

其中，$\prod_{i=1}^{n}$ 表示连乘。

在抽样之前，$\left\{Y_1, \cdots, Y_n\right\}$为随机向量。

抽样之后，$\left\{Y_1, \cdots, Y_n\right\}$则有特定的样本观测值。因此，可将样本的联合密度函数视为在给定$\left\{Y_1, \cdots, Y_n\right\}$情况下，未知参数$\boldsymbol{\theta}$的函数。

定义**似然函数**(likelihood function)为

$$L(\boldsymbol{\theta}; Y_1, \cdots, Y_n) = \prod_{i=1}^{n} f(Y_i; \boldsymbol{\theta}) \qquad (3.185)$$

似然函数与联合密度函数完全相等，只是$\boldsymbol{\theta}$与$\left\{Y_1, \cdots, Y_n\right\}$的角色互换，即把$\boldsymbol{\theta}$作为自变量，而视$\left\{Y_1, \cdots, Y_n\right\}$为给定。

为运算方便，常把似然函数取对数，将乘积的形式转化为求和的形式：

$$\ln L(\boldsymbol{\theta}; Y_1, \cdots, Y_n) = \sum_{i=1}^{n} \ln f(Y_i; \boldsymbol{\theta}) \quad (3.186)$$

最大似然估计法的思想：给定样本取值后，该样本最有可能来自参数$\boldsymbol{\theta}$为何值的总体。换言之，寻找$\hat{\boldsymbol{\theta}}_{\text{ML}}$，使得观测到样本数据的可能性最大，即最大化**对数似然函数**(loglikelihood function)：

$$\max_{\boldsymbol{\theta}} \ \ln L(\boldsymbol{\theta}; Y_1, \cdots, Y_n) \quad (3.187)$$

假设存在唯一内点解，则此无约束极值问题的一阶条件为

$$\frac{\partial \ln L(\boldsymbol{\theta}; Y_1, \cdots, Y_n)}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{3.188}$$

求解此一阶条件，即可得到最大似然估计量$\hat{\boldsymbol{\theta}}_{\text{ML}}$。

例　假设$y \sim N(\mu, \sigma^2)$，其中σ^2已知，得到一个样本容量为 1 的样本$y_1 = 2$，求对μ的最大似然估计。根据正态分布的密度函数可知，此样本的似然函数为

$$L(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(2-\mu)^2}{2\sigma^2} \right\} \qquad (3.189)$$

此似然函数在$\hat{\mu} = 2$处取最大值，参见图 3.18。

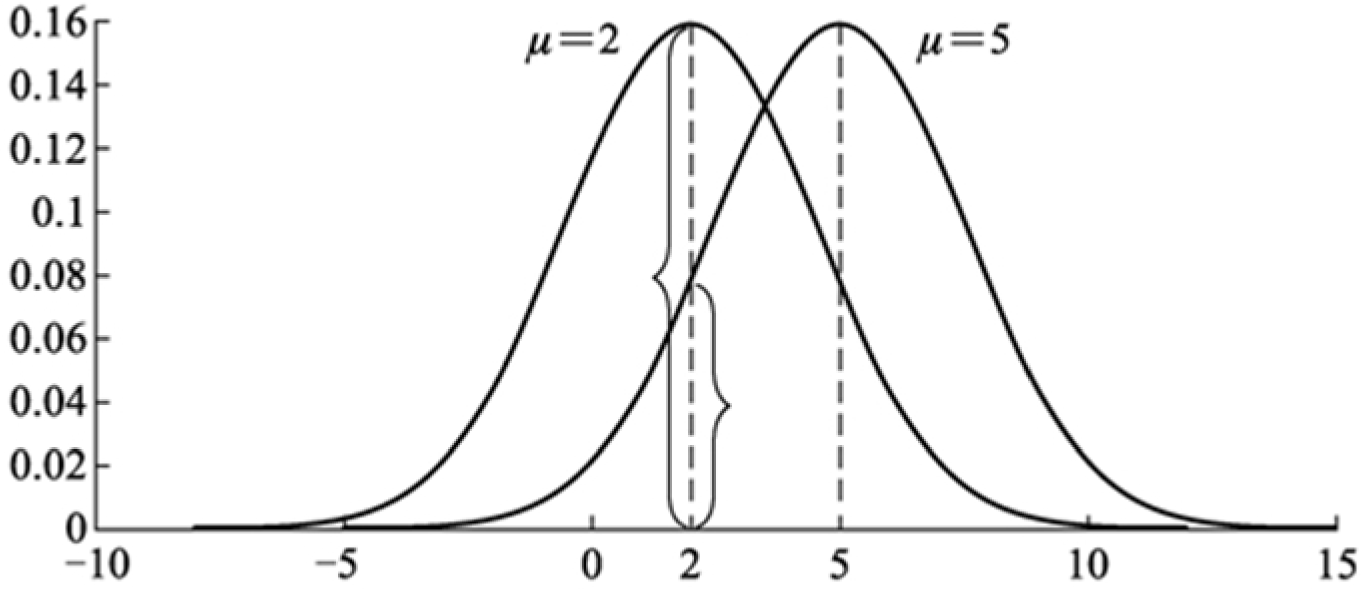

图 3.18　选择参数使观测到样本的可能性最大