第 4 章 超越梯度下降

第 3 章介绍了梯度下降法。梯度下降法基于函数局部一阶特性。一阶近似是粗糙的，这种粗糙带来了一些问题。本章将介绍函数在局部的二阶特性。基于二阶特性分析函数在局部的性质。

本章首先回顾一些矩阵的相关知识，之后介绍如何在局部对函数进行二阶近似。有了函数的二阶近似就可以确定驻点的类型：极小点、极大点或者鞍点。之后本章介绍对原始梯度下降法的一些改进，这些改进有助于提高收敛速度，防止震荡或发散，规避局部极小。

最后，本章介绍两个基于函数二阶特性的优化算法：牛顿法和共轭方向法。然后介绍用牛顿法训练逻辑回归模型。二阶算法虽然不常用在神经网络和深度学习的训练中。阅读完本章，读者应该对函数的局部形态有更深刻的理解。

4.1 矩阵

首先回顾一下矩阵。这不是一个关于矩阵的全面介绍，例如行列式这个概念就没有出现。本节只介绍一下后文讨论中用得上的相关知识。

4.1.1 矩阵基础

矩阵是实数构成的 2 维阵列。以一个3×3矩阵为例：

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = (\boldsymbol{a}_{*1} \quad \boldsymbol{a}_{*2} \quad \boldsymbol{a}_{*3}) = \begin{pmatrix} \boldsymbol{a}_{1*}^{\mathrm{T}} \\ \boldsymbol{a}_{2*}^{\mathrm{T}} \\ \boldsymbol{a}_{3*}^{\mathrm{T}} \end{pmatrix} \tag{4.1}$$

式（4.1）囊括了本书用到的对矩阵的各种表示。本书用大写粗斜体字母表示矩阵，例如$\boldsymbol{A}$。a_{ij}是实数，是矩阵$\boldsymbol{A}$的第i行、第j列元素。$\boldsymbol{a}_{*j}$是矩阵的第j列，它是一个列向量：

$$\boldsymbol{a}_{*j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{pmatrix} \tag{4.2}$$

$\boldsymbol{a}_{i*}$是矩阵的第i行，它是一个列向量：

$$\boldsymbol{a}_{i*}=\begin{pmatrix}a_{i1}\\a_{i2}\\a_{i3}\end{pmatrix} \qquad (4.3)$$

式（4.1）中对$\boldsymbol{a}_{i*}$进行了转置，以表示一行。矩阵的行数和列数不一定相等，可以是$m\times n$，$m\neq n$。表示成：

$$\boldsymbol{A}_{m\times n}=\begin{pmatrix}a_{11} & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1} & \cdots & a_{mn}\end{pmatrix} \qquad (4.4)$$

一般可省略下标$m\times n$。两个相同形状的矩阵可以相加：

$$\boldsymbol{A}+\boldsymbol{B}=\begin{pmatrix}a_{11}+b_{11} & \cdots & a_{1n}+b_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn}\end{pmatrix} \qquad (4.5)$$

矩阵相加就是把相应元素相加。可以用实数（标量）乘一个矩阵：

$$k\boldsymbol{A}=\begin{pmatrix}ka_{11} & \cdots & ka_{1n}\\ \vdots & \ddots & \vdots\\ ka_{m1} & \cdots & ka_{mn}\end{pmatrix} \qquad (4.6)$$

$-\boldsymbol{A}$就是$(-1)\boldsymbol{A}$。显然有$\boldsymbol{A}-\boldsymbol{A}=\boldsymbol{A}+(-\boldsymbol{A})=\boldsymbol{O}$。$\boldsymbol{O}$是所有元素都为 0 的矩阵——零矩阵。矩阵$\boldsymbol{A}$的转置定义为：

$$\boldsymbol{A}^{\mathrm{T}}=\begin{pmatrix}a_{11} & \cdots & a_{m1}\\ \vdots & \ddots & \vdots\\ a_{1n} & \cdots & a_{mn}\end{pmatrix}=(\boldsymbol{a}_{1*}\quad \boldsymbol{a}_{2*}\quad \boldsymbol{a}_{3*})=\begin{pmatrix}\boldsymbol{a}_{*1}^{\mathrm{T}}\\ \boldsymbol{a}_{*2}^{\mathrm{T}}\\ \boldsymbol{a}_{*3}^{\mathrm{T}}\end{pmatrix} \qquad (4.7)$$

$\boldsymbol{A}^{\mathrm{T}}$把$\boldsymbol{A}$的行当做列，列当做行。如果$\boldsymbol{A}$是$m\times n$的，那么$\boldsymbol{A}^{\mathrm{T}}$就是$n\times m$的。如果矩阵$\boldsymbol{A}$是$m\times n$的，它可以与一个$n$维向量$\boldsymbol{x}$相乘：

$$\boldsymbol{A}\boldsymbol{x}=\sum_{j=1}^{n}x_j\,\boldsymbol{a}_{*j} \qquad (4.8)$$

矩阵$\boldsymbol{A}$乘向量$\boldsymbol{x}$，使用$\boldsymbol{x}$的元素对矩阵的列进行线性组合。所以$\boldsymbol{A}$的列数和$\boldsymbol{x}$的维数必须相同。得到的结果是一个m维向量。容易看出$\boldsymbol{Ax}$的第i个元素是$\sum_{j=1}^{n} x_j a_{ij} = \boldsymbol{a}_{i*}^{\mathrm{T}}\boldsymbol{x}$，即$\boldsymbol{A}$的第$i$行与$\boldsymbol{x}$的内积。

有了矩阵和向量相乘的定义，就可以定义矩阵与矩阵相乘：

$$\boldsymbol{AB} = (\boldsymbol{Ab}_{*1} \quad \boldsymbol{Ab}_{*2} \quad \cdots \quad \boldsymbol{Ab}_{*k}) \tag{4.9}$$

$\boldsymbol{A}$与$\boldsymbol{B}$的乘积是矩阵$\boldsymbol{AB}$。$\boldsymbol{AB}$的第j列是$\boldsymbol{A}$与$\boldsymbol{B}$的第j列$\boldsymbol{b}_{*j}$的乘积。如果$\boldsymbol{A}$是$m \times n$的，那么$\boldsymbol{b}_{*j}$必须是n维向量，即$\boldsymbol{B}$必须为n行。$\boldsymbol{B}$的列数任意，例如k。所以要能够与$m \times n$的$\boldsymbol{A}$相乘，$\boldsymbol{B}$的形状必须是$n \times k$，k任意。结果$\boldsymbol{AB}$的形状是$m \times k$。$\boldsymbol{AB}$的第i行、第j列元素是：

$$\boldsymbol{a}_{i*}^{\mathrm{T}}\boldsymbol{b}_{*j} = \sum_{s=1}^{n} a_{is} b_{sj} \tag{4.10}$$

仅从形状上看$\boldsymbol{B}$与$\boldsymbol{A}$不一定能够相乘，因为k不一定等于m。就算$k = m$，$\boldsymbol{BA}$也不一定等于$\boldsymbol{AB}$。即矩阵乘法不满足交换律。一个反例就可以证明这一点。这里不再赘述。

矩阵的乘法满足结合率：

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \tag{4.11}$$

矩阵乘法对加法满足结合律：

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}, \quad (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC} \tag{4.12}$$

矩阵乘法对数乘满足：

$$\boldsymbol{A}(k\boldsymbol{B}) = (k\boldsymbol{A})\boldsymbol{B} = k(\boldsymbol{AB}) \tag{4.13}$$

矩阵的数乘满足分配率：

$$k(\boldsymbol{A} + \boldsymbol{B}) = k\boldsymbol{A} + k\boldsymbol{B}, \quad (k + h)\boldsymbol{A} = k\boldsymbol{A} + h\boldsymbol{A} \tag{4.14}$$

矩阵乘积的转置是：

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T \qquad (4.15)$$

上述几个规则的证明很简单，只需要检查一下矩阵元素的表达式。向量$\boldsymbol{x}$的转置$\boldsymbol{x}^T$可以乘一个矩阵$\boldsymbol{A}$：

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}} \qquad (4.16)$$

把列向量和行向量分别看作$n \times 1$和$1 \times n$的矩阵，则矩阵与向量的乘法也满足上述几个规则。

行数和列数相同的矩阵是方阵。方阵$\boldsymbol{A}$的对角线元素之和称为它的迹（trace）：

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \qquad (4.17)$$

方阵$\boldsymbol{A}$和$\boldsymbol{B}$的乘积$\boldsymbol{AB}$的迹等于$\boldsymbol{BA}$的迹，因为：

$$tr(\boldsymbol{AB}) = \sum_{i=1}^{n} \boldsymbol{a}_{i*}\boldsymbol{b}_{*i} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \sum_{j=1}^{n} \boldsymbol{b}_{j*}\boldsymbol{a}_{*j} = tr(\boldsymbol{BA}) \qquad (4.18)$$

如果一个$n \times n$的方阵的对角线元素为 1，其余元素都是 0，那么它是单位阵：

$$\boldsymbol{I}_{n\times n} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \qquad (4.19)$$

容易验证对于任何矩阵$\boldsymbol{A}_{m\times n}$，$\boldsymbol{I}_{m\times m}\boldsymbol{A}_{m\times n} = \boldsymbol{A}_{m\times n}\boldsymbol{I}_{n\times n} = \boldsymbol{A}_{m\times n}$。在上下文很清晰时一般省略$\boldsymbol{I}$的下标。

4.1.2 矩阵的逆

令$\boldsymbol{A}$是$n \times n$方阵，如果存在$n \times n$方阵$\boldsymbol{A}^{-1}$满足：

$$\boldsymbol{A}\boldsymbol{A}^{-1}=\boldsymbol{A}^{-1}\boldsymbol{A}=\boldsymbol{I} \qquad (4.20)$$

则称$\boldsymbol{A}$是可逆的。$\boldsymbol{A}^{-1}$是$\boldsymbol{A}$的逆矩阵。$\boldsymbol{A}$的逆矩阵是唯一的。因为假如任何一个矩阵$\boldsymbol{B}$是$\boldsymbol{A}$的逆矩阵，根据定义有：

$$\boldsymbol{B}=\boldsymbol{I}\boldsymbol{B}=\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{B}=\boldsymbol{A}^{-1} \qquad (4.21)$$

如果$\boldsymbol{A}$可逆则$\boldsymbol{A}$的列线性独立。因为假如$\boldsymbol{a}_{*j=1\ldots n}$线性相关，则存在一组不全为 0 的系数$w_1,w_2,\cdots,w_n$，使得$\sum_{i=1}^{n}w_i\boldsymbol{a}_{*i}=\boldsymbol{0}$。即存在向量$\boldsymbol{w}=(w_1 \quad \cdots \quad w_n)^{\mathrm{T}}\neq\boldsymbol{0}$使：

$$\boldsymbol{A}\boldsymbol{w}=\boldsymbol{0} \qquad (4.22)$$

因为$\boldsymbol{A}$可逆，存在$\boldsymbol{A}^{-1}$：

$$\boldsymbol{w}=\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{w}=\boldsymbol{0} \qquad (4.23)$$

这与$\boldsymbol{w}\neq\boldsymbol{0}$矛盾。所以可逆矩阵$\boldsymbol{A}$的列$\boldsymbol{a}_{*j=1\ldots n}$一定线性独立。如果方阵$\boldsymbol{A}$的逆矩阵是$\boldsymbol{A}^{\mathrm{T}}$，则称$\boldsymbol{A}$为正交矩阵：

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I} \qquad (4.24)$$

从（4.24）可以看出$\boldsymbol{A}$的列$\boldsymbol{a}_{*j=1\ldots n}$是单位向量且两两正交：

$$\begin{cases}\boldsymbol{a}_{*i}{}^{T}\boldsymbol{a}_{*j}=0,\ i\neq j\\ \boldsymbol{a}_{*i}{}^{T}\boldsymbol{a}_{*j}=1,\ i=j\end{cases} \qquad (4.25)$$

也就是说，正交矩阵$\boldsymbol{A}$的列都是单位向量，$\left\|\boldsymbol{a}_{*j=1\ldots n}\right\|=1$,。任意两列是正交的（夹角为$\pi/2$）。因为$\boldsymbol{A}$可逆，所以$\boldsymbol{a}_{*j=1\ldots n}$线性独立，是$n$维线性空间$\mathbb{R}^n$的一组基。因为$\boldsymbol{a}_{*j=1\ldots n}$两两正交，还是单位向量，所以它们被称为$\mathbb{R}^n$的一组标准正交基。

如果一组向量$\boldsymbol{x}_{i=1\ldots n}$是线性独立的，可以通过施密特正交化过程构造一组正交的向量$\boldsymbol{x}'_{i=1\ldots n}$。

构造过程是，首先令$\boldsymbol{x'}_1 = \boldsymbol{x}_1$。然后令：

$$\boldsymbol{x'}_2 = \boldsymbol{x}_2 - \frac{\boldsymbol{x}_2{}^{\mathrm{T}}\boldsymbol{x'}_1}{\boldsymbol{x'}_1{}^{\mathrm{T}}\boldsymbol{x'}_1}\boldsymbol{x'}_1 = \boldsymbol{x}_2 - \frac{\|\boldsymbol{x}_2\|\cdot\cos\theta_{21}}{\|\boldsymbol{x'}_1\|}\boldsymbol{x'}_1 \qquad (4.26)$$

θ_{21}是$\boldsymbol{x}_2$与$\boldsymbol{x'}_1$的夹角。$\boldsymbol{x'}_2$是$\boldsymbol{x}_2$减去$\boldsymbol{x}_2$向$\boldsymbol{x'}_1$的投影。$\boldsymbol{x}_2$与$\boldsymbol{x'}_1 = \boldsymbol{x}_1$线性独立，它不是$\boldsymbol{x'}_1$的数乘，所以$\boldsymbol{x'}_2$不是零向量。而且容易验证$\boldsymbol{x'}_2$与$\boldsymbol{x'}_1$正交。再令：

$$\boldsymbol{x'}_3 = \boldsymbol{x}_3 - \frac{\boldsymbol{x}_3{}^{\mathrm{T}}\boldsymbol{x'}_1}{\boldsymbol{x'}_1{}^{\mathrm{T}}\boldsymbol{x'}_1}\boldsymbol{x'}_1 - \frac{\boldsymbol{x}_3{}^{\mathrm{T}}\boldsymbol{x'}_2}{\boldsymbol{x'}_2{}^{\mathrm{T}}\boldsymbol{x'}_2}\boldsymbol{x'}_2 = \boldsymbol{x}_3 - \frac{\|\boldsymbol{x}_3\|\cdot\cos\theta_{31}}{\|\boldsymbol{x'}_1\|}\boldsymbol{x'}_1 - \frac{\|\boldsymbol{x}_3\|\cdot\cos\theta_{32}}{\|\boldsymbol{x'}_2\|}\boldsymbol{x'}_2 \qquad (4.27)$$

θ_{31}是$\boldsymbol{x}_3$与$\boldsymbol{x'}_1$的夹角，θ_{32}是$\boldsymbol{x}_3$与$\boldsymbol{x'}_2$的夹角。$\boldsymbol{x'}_3$是$\boldsymbol{x}_3$减去$\boldsymbol{x}_3$向$\boldsymbol{x'}_1, \boldsymbol{x'}_2$张成空间的投影。如果$\boldsymbol{x'}_3 = 0$，那么$\boldsymbol{x}_3$可以被$\boldsymbol{x'}_1, \boldsymbol{x'}_2$线性表出，也就可以被$\boldsymbol{x}_1, \boldsymbol{x}_2$线性表出，这与$\boldsymbol{x}_{i=1\ldots n}$线性独立矛盾。故$\boldsymbol{x'}_3 \neq 0$。容易验证$\boldsymbol{x'}_3$正交于$\boldsymbol{x'}_1, \boldsymbol{x'}_2$。此过程继续下去，最终可构造一组正交向量$\boldsymbol{x'}_{i=1\ldots n}$。这就是施密特正交化过程。将$\boldsymbol{x'}_{i=1\ldots n}$的每一个向量除以各自的模，缩放到长度为 1，就得到了一组正交的单位向量。

4.1.3 特征值与特征向量

特征值和特征向量的概念不局限于方阵，但本书主要关注方阵。用方阵$\boldsymbol{A}$乘向量$\boldsymbol{x}$是在$\mathbb{R}^n$中进行一个变换，将$\boldsymbol{x}$变换成$\boldsymbol{Ax}$。例如矩阵：

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \qquad (4.28)$$

用$\boldsymbol{R}$乘向量$\boldsymbol{x}$等于将$\boldsymbol{x}$逆时针旋转θ度。这可以自行验证。任何方阵$\boldsymbol{A}$也改变不了零向量$\boldsymbol{0}$，因为$\boldsymbol{A0} \equiv \boldsymbol{0}$。如果对于某非零向量$\boldsymbol{v} \neq \boldsymbol{0}$，$\boldsymbol{A}$只能改变$\boldsymbol{v}$的长度而不能改变其方向，即存在某个标量（可以为 0）λ，有：

$$\boldsymbol{Av} = \lambda\boldsymbol{v} \qquad (4.29)$$

则称λ是$\boldsymbol{A}$的特征值，$\boldsymbol{v}$是$\boldsymbol{A}$的对于λ的特征向量。同一个特征向量不可能对应两个特征值。假如$\lambda_1 \neq \lambda_2$都有$\boldsymbol{v}$是其特征向量：

$$(\lambda_1 - \lambda_2)\boldsymbol{v} = \lambda_1\boldsymbol{v} - \lambda_2\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} - \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0} \qquad (4.30)$$

$\lambda_1 - \lambda_2 \neq 0$且$\boldsymbol{v} \neq \boldsymbol{0}$，所以式（4.30）是不可能的。但是同一个特征值可以对应多个特征向量。如果$\boldsymbol{v}$是λ对应的特征向量，容易验证$k\boldsymbol{v}$也是λ对应的特征向量。线性独立的两个向量也有可能是同一个特征值对应的特征向量。

假如λ对应的特征向量$\boldsymbol{v}$和$\boldsymbol{w}$是线性独立的，即谁也不是另一个的数乘。那么$k\boldsymbol{v} + l\boldsymbol{w}$也是$\lambda$对应的特征向量。这也很容易验证。如果特征值$\lambda$共有$k$个线性独立的特征向量，由它们线性组合而得的向量也是λ的特征向量。这k个线性独立的特征向量张成的k维线性空间称为λ对应的特征空间，其中所有向量都是λ的特征向量。

将式（4.29）变形。如果λ是$\boldsymbol{A}$的特征值，它必须满足对某个$\boldsymbol{v} \neq \boldsymbol{0}$，有：

$$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0} \qquad (4.31)$$

因为$\boldsymbol{v} \neq \boldsymbol{0}$，所以$\boldsymbol{A} - \lambda\boldsymbol{I}$的列线性相关。求$\boldsymbol{A}$的特征值和特征向量，就是求满足方程（4.31）的$\lambda$和$\boldsymbol{v}$。若要$\boldsymbol{A} - \lambda\boldsymbol{I}$的列线性相关，则$\boldsymbol{A} - \lambda\boldsymbol{I}$的行列式$|\boldsymbol{A} - \lambda\boldsymbol{I}|$等于 0。本书没有涉及行列式，因为行列式与本书主线关系不大，加进来会影响流畅性。读者可以查阅任何一种线性代数教材。$|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$是$\lambda$的$n$次方程。它有$n$个根（包括重根和复数根）。即$\boldsymbol{A}$有$n$个特征值（包括重复的以及复特征值）。求得了$\lambda$，就可以再求它对应的特征向量。

矩阵$\boldsymbol{A}$属于不同特征值的特征向量是线性独立的。现在证明这一点。$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_k$是$k$个不同特征值$\lambda_1, \lambda_2, \cdots, \lambda_k$对应的特征向量。如果它们线性相关，则其中某一个$\boldsymbol{v}_s$可以被其他$\boldsymbol{v}_{i\neq s}$线性表出：

$$\boldsymbol{v}_s = \sum_{i\neq s} a_i \boldsymbol{v}_i \qquad (4.32)$$

因为$\boldsymbol{v}_s$是$\boldsymbol{A}$的特征向量，所以它不是零向量。那么$a_{i\neq s}$一定不全为 0。另外根据特征值和特征向量的定义：

$$\lambda_s\boldsymbol{v}_s = \boldsymbol{A}\boldsymbol{v}_s = \sum_{i\neq s} a_i \boldsymbol{A}\boldsymbol{v}_i = \sum_{i\neq s} a_i \lambda_i \boldsymbol{v}_i \qquad (4.33)$$

如果$\lambda_s = 0$，那么$\lambda_{i\neq s} \neq 0$。再加上$a_{i\neq s}$不全为 0，说明$\boldsymbol{v}_{i\neq s}$线性相关。在$\lambda_s = 0$情况下我们将问题规模减小了 1。如果$\lambda_s \neq 0$，有：

$$\boldsymbol{v}_s = \sum_{i\neq s} a_i \frac{\lambda_i}{\lambda_s} \boldsymbol{v}_i \qquad (4.34)$$

于是式（4.34）等于式（4.32），所以有：

$$\sum_{i\neq s}\left(a_i\frac{\lambda_i}{\lambda_s}\boldsymbol{v}_i-a_i\boldsymbol{v}_i\right)=\sum_{i\neq s}a_i\left(\frac{\lambda_i}{\lambda_s}-1\right)\boldsymbol{v}_i=\boldsymbol{0} \qquad (4.35)$$

因为都是不同的特征值，所以$\lambda_{i\neq s}/\lambda_s-1\neq 0$。再加上$a_{i\neq s}$不全为 0，说明$\boldsymbol{v}_{i\neq s}$线性相关。在$\lambda_s\neq 0$情况下我们也将问题规模减小了 1。这个过程持续下去，最终将只剩下两个向量$\boldsymbol{v}_i$和$\boldsymbol{v}_j$。他们分属不同的特征值λ_i和λ_j。且$\boldsymbol{v}_i$和$\boldsymbol{v}_j$线性相关，其中一个是另一个的数乘。不妨假设$\boldsymbol{v}_i=k\boldsymbol{v}_j$，则$\boldsymbol{v}_i$也是$\lambda_j$的特征向量。之前已经证明，一个向量不可能同时属于两个不同特征值。这就推翻了最早的假设，证明了$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$线性独立。

4.1.4 对称矩阵的谱分解

如果方阵$\boldsymbol{A}$满足:

$$\boldsymbol{A}=\boldsymbol{A}^{\mathrm{T}} \qquad (4.36)$$

如果$\boldsymbol{A}$的元素都是实数，则它是一个实矩阵。实矩阵的特征值都是实数，特征向量是实向量。为了证明这个结论，我们需要暂时离开实数域。

复数$\lambda=a+bi$的共轭是$\bar{\lambda}=a-bi$。

$$\lambda\bar{\lambda}=(a+bi)(a-bi)=a^2+b^2\geq 0 \qquad (4.37)$$

只有当$a=b=0$，即$\lambda=0$时，才有$\lambda\bar{\lambda}=0$。否则$\lambda\bar{\lambda}>0$。对于两个复数$\lambda=a+bi$和$\xi=c+di$，有：

$$\lambda\xi=(ac-bd)+(bc+ad)i=\bar{\lambda}\bar{\xi} \qquad (4.38)$$

把复矩阵$\boldsymbol{A}$的元素全都取共轭就得到$\boldsymbol{A}$的共轭$\overline{\boldsymbol{A}}$。如果（复数）$\lambda$和（复向量）$\boldsymbol{v}$是$\boldsymbol{A}$的特征值及对应特征向量，由式（4.38）容易看出：$\bar{\lambda}$和$\overline{\boldsymbol{v}}$是$\overline{\boldsymbol{A}}$的特征值及对应特征向量。

$$\bar{\boldsymbol{A}}\bar{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} = \bar{\lambda}\bar{\boldsymbol{v}} \qquad (4.39)$$

因为$\boldsymbol{A}$是实对称矩阵，有$\boldsymbol{A} = \bar{\boldsymbol{A}} = \boldsymbol{A}^{\mathrm{T}} = \bar{\boldsymbol{A}}^{\mathrm{T}}$，所以有：

$$\lambda\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\lambda\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v} = \bar{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{v} = (\bar{\boldsymbol{A}}\bar{\boldsymbol{v}})^{\mathrm{T}}\boldsymbol{v} = \left(\bar{\lambda}\bar{\boldsymbol{v}}\right)^{\mathrm{T}}\boldsymbol{v} = \bar{\lambda}\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} \qquad (4.40)$$

因为$\boldsymbol{x} \neq \boldsymbol{0}$，根据式（4.37）$\bar{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v} > 0$。所以$\lambda = \bar{\lambda}$，即$\lambda$是实数。$\boldsymbol{A}$是实矩阵，$\lambda$是实数，所以$\boldsymbol{v}$一定是实向量。这就证明了实对称矩阵的特征值都是实数，特征向量是实向量。后文谈到矩阵都是实矩阵。

所以实对称矩阵$\boldsymbol{A}$有n个实特征值（可重复）。有一个结论我们不加证明：如果λ是$\boldsymbol{A}$的k重特征值（方程$|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$的$k$重根），则$\lambda$对应的特征空间是$k$维，即对于$\lambda$能找到$k$个线性独立的特征向量。

对于对称矩阵$\boldsymbol{A}$，求得它的n个特征值$\lambda_1, \lambda_2, \cdots, \lambda_n$，从大到小排列。对每个特征值找到一个它的单位特征向量。如果某个特征值是k重，那么找到它的k个线性独立的特征向量，经过施密特正交化过程，再缩放到长度为 1。这样，一共找到n个单位特征向量$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$。可以证明这$n$个特征向量是线性独立的。因为假如它们线性相关，则存在一组不全为 0 的参数$a_1, a_2, \cdots, a_n$满足：

$$\sum_{i=1}^{n} a_i\boldsymbol{v}_i = \boldsymbol{0} \qquad (4.41)$$

如果n个特征值共有s种不同取值，把属于同一个取值的特征向量归到一起：

$$\sum_{i=1}^{s}\sum_{j\epsilon I(i)} a_j\boldsymbol{v}_j = \sum_{i=1}^{s}\boldsymbol{v}_i' = \boldsymbol{0} \qquad (4.42)$$

$I(i)$是属于第i个排重特征值（一共s排重特征值）的特征向量下标集合。式（4.42）中每一个$\boldsymbol{v}_i' = \sum_{j\epsilon I(i)} a_j\boldsymbol{v}_j$是对第$i$个排重特征值的特征向量的线性组合，它仍然是第$i$个排重特征值的特征向量。不可能所有$\boldsymbol{v}_i'$都是零向量。因为存在非 0 的$a_j$且$\sum_{j\epsilon I(i)} a_j\boldsymbol{v}_j = 0$，这与$\boldsymbol{v}_{j\in I(i)}$线性独立产生矛盾。那些非零的$\boldsymbol{v}_i'$加在一起是零向量，又与属于不同特征值的特征向量线性独立矛盾。所以$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$是线性独立的。

对称矩阵$\boldsymbol{A}$属于不同特征值的特征向量是正交的。如果λ_i和λ_j是$\boldsymbol{A}$的两个不同特征值，$\boldsymbol{v}_i$和$\boldsymbol{v}_j$是特们各自的特征向量，有：

$$\lambda_j \boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{v}_j = \boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{v}_j = \boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{v}_j = \boldsymbol{v}_j{}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_j{}^{\mathrm{T}} \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{v}_j \qquad (4.43)$$

式（4.43）多次用到了特征值/特征向量的定义，$\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}$以及标量的转置还是该实数本身。注意观察式（4.43）那些计算结果是标量。

根据式（4.43），有：

$$(\lambda_i - \lambda_j) \boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{v}_j = 0 \qquad (4.44)$$

因为$\lambda_i \neq \lambda_j$，所以必有$\boldsymbol{v}_i{}^{\mathrm{T}} \boldsymbol{v}_j = 0$。

$\boldsymbol{A}$的属于同一个特征值的特征向量已经经过施密特正交化，它们彼此正交。刚刚证明了属于$\boldsymbol{A}$不同特征值的特征向量是正交的，所以$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$是线性独立、两两正交的单位向量。用它们作为列，构造矩阵：

$$\boldsymbol{V}^{\mathrm{T}} = (\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_n) \qquad (4.45)$$

$\boldsymbol{V}^{\mathrm{T}}$是正交矩阵。用$\boldsymbol{A}$的特征值$\lambda_1, \lambda_2, \cdots, \lambda_n$构造对角矩阵：

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \qquad (4.46)$$

$\boldsymbol{\Lambda}$的n个对角线元素是从大到小排列的特征值。其余元素是 0。有：

$$\boldsymbol{V}^{\mathrm{T}} \boldsymbol{\Lambda} = (\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_n) \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} = (\lambda_1 \boldsymbol{v}_1 \quad \cdots \quad \lambda_n \boldsymbol{v}_n) = \boldsymbol{A} \boldsymbol{V}^{\mathrm{T}} \qquad (4.47)$$

因为$\boldsymbol{V} \boldsymbol{V}^{\mathrm{T}} = \boldsymbol{I}$，所以有：

$$\boldsymbol{A} = \boldsymbol{V}^{\mathrm{T}} \boldsymbol{\Lambda} \boldsymbol{V} \qquad (4.48)$$

这就是对称矩阵的谱分解。

4.1.5 二次型

对于某个方阵$\boldsymbol{A}$（不一定对称）：

$$q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \qquad (4.49)$$

称作一个二次型。例如：

$$q(\boldsymbol{x}) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = {x_1}^2 + 5x_1x_2 + 4{x_2}^2 \qquad (4.50)$$

二次型只包含二次项，不包含一次项、常数（零次）项以及更高次项。这种情况称为是齐次的。从（4.50）可以看出，$q(\boldsymbol{x})$也可以写成：

$$q(\boldsymbol{x}) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & \frac{5}{2} \\ \frac{5}{2} & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = {x_1}^2 + 5x_1x_2 + 4{x_2}^2 \qquad (4.51)$$

将“交叉”项（二元情况下就是x_1x_2）的系数对半分，任何一个二次型都能写成关于对称矩阵的二次型。

一个对称矩阵$\boldsymbol{A}$，如果对于任何向量$\boldsymbol{x}$都有：

$$q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \geq 0 \qquad (4.52)$$

则称$\boldsymbol{A}$是半正定的（positive semidefinite）。如果不等号是严格的（$>$），则$\boldsymbol{A}$是正定的（positive definite）。相应地还有半负定（negative semidefinite）和负定（negative definite）。半正定矩阵的所有特征值都不为负，因为假如$\boldsymbol{A}$一个负数特征值$\lambda_i < 0$，那么$\boldsymbol{A}$可以分解为：

$$\boldsymbol{A} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{V} \qquad (4.52)$$

因为$\boldsymbol{V}^{\mathrm{T}}$是正交矩阵，根据正交矩阵的定义，$\boldsymbol{V}$也是正交矩阵。$\boldsymbol{V}$的列是$\mathbb{R}^n$的基，即存在$\boldsymbol{x}$使

得$\boldsymbol{e}_i = \boldsymbol{V}\boldsymbol{x}$。那么有：

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{V}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{V}\boldsymbol{x} = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{e}_i = \lambda_i < 0 \qquad (4.53)$$

这与$\boldsymbol{A}$的半正定性矛盾。所以半正定矩阵的所有特征值都非负。同样可以证明正定矩阵的所有特征值都为正。半负定矩阵的所有特征值都非正以及负定矩阵的所有特征值都为负。

半正定矩阵可以分解成两个矩阵的乘积：

$$\boldsymbol{A} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{V} = \boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{pmatrix}\boldsymbol{V} = \boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\sqrt{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{\lambda_n}\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{\lambda_n}\end{pmatrix}\boldsymbol{V} =$$

$$\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{\mathrm{T}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V} = \left(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}\right)^{\mathrm{T}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V} \qquad (4.54)$$

如果限定$\boldsymbol{x}$是单位向量，则半正定矩阵$\boldsymbol{A}$的二次型的最大值是$\boldsymbol{A}$的最大特征值λ_1，达到这个最大值的$\boldsymbol{x}$是λ_1的某一个单位特征向量：

$$\max_{\|\boldsymbol{x}\|=1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \lambda_1 \qquad (4.55)$$

$\boldsymbol{V}$是正交矩阵，$\boldsymbol{x}$是任意向量，审视$\boldsymbol{V}\boldsymbol{x}$的模：

$$\|\boldsymbol{V}\boldsymbol{x}\|^2 = (\boldsymbol{V}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = \|\boldsymbol{x}\|^2 \qquad (4.66)$$

即用正交矩阵去变换任意向量，不改变向量的长度。从几何角度理解，正交矩阵的行是标准正交基。用正交矩阵乘一个向量相当于将该向量投影在新的标准坐标系中，所以长度不发生改变。

令$\boldsymbol{\Lambda}$是对半正定$\boldsymbol{A}$谱分解得到的对角阵，对于任意单位向量$\boldsymbol{x}' = \boldsymbol{V}\boldsymbol{x}$，有：

$$(\boldsymbol{x}')^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{x}' = \sum_{i=1}^{n}\lambda_i(x_i')^2 \qquad (4.67)$$

显然在$\sum_{i=1}^{n}(x_i')^2 = 1$限制下，当$x_1' = 1$且$x_{i\neq 1}' = 0$，也就是$\boldsymbol{x}' = \boldsymbol{e}_1$时式（4.67）达到最大。如果$\boldsymbol{V}\boldsymbol{x} = \boldsymbol{e}_1$，则$\boldsymbol{x} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{e}_1 = \boldsymbol{v}_1$。所以当$\boldsymbol{x}$是$\boldsymbol{A}$的最大特征值$\lambda_1$对应的某个单位特征向量$\boldsymbol{v}_1$时，

$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$取得最大值$\lambda_1$。

问$\|\boldsymbol{x}\|=1$时次大的$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$值没有意义。二次型是连续的，当$\boldsymbol{x}$无限接近$\boldsymbol{v}_1$，$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$可以无限接近$\lambda_1$。我们加上一个限制条件：$\boldsymbol{x}$须与$\boldsymbol{v}_1$正交。有结论：

$$\max_{\|x\|=1,x\perp v_1}\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\lambda_2 \qquad (4.68)$$

在模为 1 并正交于$\boldsymbol{v}_1$约束下，$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$的最大值是$\boldsymbol{A}$的第二大特征值$\lambda_2$。$\boldsymbol{x}$取$\lambda_2$的某个单位特征向量时达到此值。证明这个结论，回顾式（4.67），正交于$\boldsymbol{v}_1$的约束使$\boldsymbol{x}'$的第一个元素为 0。那么令式（4.67）最大就是把所有“量”都分配给$\boldsymbol{x}'$的第 2 个元素，因为它对应的系数是第二大特征值λ_2。所以$\boldsymbol{V}\boldsymbol{x}=\boldsymbol{e}_2$，即$\boldsymbol{x}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{e}_2=\boldsymbol{v}_2$。

总结一下：半正定矩阵$\boldsymbol{A}$的最大特征值λ_1对应的单位特征向量$\boldsymbol{v}_1$，在$\|\boldsymbol{x}\|=1$的约束下使$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$达到最大值$\lambda_1$；$\boldsymbol{A}$的第二大特征值$\lambda_2$对应的单位特征向量$\boldsymbol{v}_2$，在$\|\boldsymbol{x}\|=1$且正交于$\boldsymbol{v}_1$的约束下使$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$达到最大值$\lambda_2$；$\boldsymbol{A}$的第三大特征值$\lambda_2$对应的单位特征向量$\boldsymbol{v}_3$，在$\|\boldsymbol{x}\|=1$且正交于$\boldsymbol{v}_1$和$\boldsymbol{v}_2$的约束下使$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$达到最大值$\lambda_3$。依次类推。

这个结论在后文介绍主成分分析和导致梯度下降发生震荡的损失函数病态形态时都有用处。

4.2 多元函数局部二阶特性与驻点的类型

第 3 章讲到：梯度为零向量的点——驻点，是局部极小/大点的必要条件。但是紧靠一阶梯度信息，无法判断驻点的具体类型。要做到这一点，就要分析函数在驻点附近的二阶特性。

4.2.1 赫森矩阵

n元函数$f(\boldsymbol{x})$在$\boldsymbol{x}$点的赫森（Hessian）矩阵是$n\times n$矩阵$\boldsymbol{H}(\boldsymbol{x})$:

$$\boldsymbol{H}(\boldsymbol{x})=\begin{pmatrix}\frac{\partial f}{\partial x_1\partial x_1} & \cdots & \frac{\partial f}{\partial x_n\partial x_1}\\ \vdots & \ddots & \vdots\\ \frac{\partial f}{\partial x_1\partial x_n} & \cdots & \frac{\partial f}{\partial x_n\partial x_n}\end{pmatrix} \qquad (4.69)$$

$\frac{\partial f}{\partial x_j\partial x_i}$表示先求$f$对$x_i$求偏导，再对$x_j$求偏导，即$f$的二阶偏导。这些二阶导数是在$\boldsymbol{x}$点计算，应记作$\frac{\partial f}{\partial x_j\partial x_i}(\boldsymbol{x})$。文中为了清晰省略自变量$\boldsymbol{x}$。有一个结论本书不给出证明：如果$\frac{\partial f}{\partial x_j\partial x_i}$和$\frac{\partial f}{\partial x_i\partial x_j}$都是连续的，则它们相等。在本书考虑的问题中这种连续性要求都是满足的。所以$\boldsymbol{H}(\boldsymbol{x})$

是一个对称矩阵。

赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$的第一行是$\begin{pmatrix}\frac{\partial f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial f}{\partial x_n \partial x_1}\end{pmatrix}$。它的$n$个元素是$\frac{\partial f}{\partial x_1}$对$x_1, x_2 \cdots, x_n$的偏导。这个行向量是函数$\frac{\partial f}{\partial x_1}$的梯度的转置$\left(\nabla \frac{\partial f}{\partial x_1}\right)^{\mathrm{T}}$。

4.2.2 二阶泰勒展开

前文我们知道函数$\boldsymbol{f}(\boldsymbol{x})$在$\boldsymbol{x}$沿着方向$\boldsymbol{d}$的方向导数是$\nabla \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$。当限制在一个方向上，可以把$\boldsymbol{f}(\boldsymbol{x})$看做一元函数，方向导数就是该一元函数的导数。那么当然也存在沿着方向$\boldsymbol{d}$的二阶导数，即导数的导数。首先我们需要求$\nabla \boldsymbol{f}^{\mathrm{T}}\boldsymbol{d}$的梯度：

$$\frac{\partial \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}}{\partial x_i} = \frac{\partial \sum_{j=1}^{n} d_j \frac{\partial f}{\partial x_j}}{\partial x_i} = \sum_{j=1}^{n} d_j \frac{\partial f}{\partial x_i \partial x_j} = \boldsymbol{h}_{*i}{}^{T}\boldsymbol{d} \qquad (4.70)$$

其中$\boldsymbol{h}_{*i}$是$\boldsymbol{H}(\boldsymbol{x})$的第$i$列。$\nabla \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$对$x_i$的偏导数是$\boldsymbol{H}(\boldsymbol{x})$的第$i$列与$\boldsymbol{d}$的内积，所以$\nabla \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$在$\boldsymbol{x}$的梯度是：

$$\nabla(\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}) = \begin{pmatrix} \frac{\partial \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}}{\partial x_1} \\ \vdots \\ \frac{\partial \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{h}_{*1}{}^{\mathrm{T}}\boldsymbol{d} \\ \vdots \\ \boldsymbol{h}_{*n}{}^{\mathrm{T}}\boldsymbol{d} \end{pmatrix} = \boldsymbol{H}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d} \qquad (4.71)$$

所以$\nabla \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$沿$\boldsymbol{d}$方向的方向导，也就是$\boldsymbol{f}(\boldsymbol{x})$沿$\boldsymbol{d}$方向的二阶导是：

$$\left(\nabla(\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d})\right)^{\mathrm{T}}\boldsymbol{d} = (\boldsymbol{H}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d})^{\mathrm{T}}\boldsymbol{d} = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{d} \qquad (4.72)$$

结论是$\boldsymbol{f}(\boldsymbol{x})$在$\boldsymbol{x}$沿着方向$\boldsymbol{d}$的二阶导数是二次型$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{d}$。回忆一下一元函数的二阶泰勒展开（一元微积分是本书的先导知识，故涉及一元微积分的结论直接使用）：

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)h^2}{2} + \mathcal{R}(h) \qquad (4.73)$$

余项$\mathcal{R}(h)$是h^2的高阶无穷小：

$$\lim_{h\to 0}\frac{\mathcal{R}(h)}{h^2}=0 \qquad (4.74)$$

二阶泰勒展开可以扩展到多维：

$$f(\boldsymbol{x}+\boldsymbol{h})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}+\frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{h}}{2}+\mathcal{R}(\boldsymbol{h}) \qquad (4.75)$$

$\nabla f(\boldsymbol{x})$是f在$\boldsymbol{x}$的梯度，$\boldsymbol{H}(\boldsymbol{x})$是$f$在$\boldsymbol{x}$的赫森矩阵。$\mathcal{R}(\boldsymbol{h})$是$\|\boldsymbol{h}\|^2$的高阶无穷小：

$$\lim_{\|\boldsymbol{h}\|\to 0}\frac{\mathcal{R}(\boldsymbol{h})}{\|\boldsymbol{h}\|^2}=0 \qquad (4.76)$$

因为任何一个$f(\boldsymbol{x}+\boldsymbol{h})$都可以看作是自变量从$\boldsymbol{x}$出发，沿着$\boldsymbol{d}=\boldsymbol{h}/\|\boldsymbol{h}\|$方向变化了$\|\boldsymbol{h}\|$距离。将$f(\boldsymbol{x}+\boldsymbol{h})$视为关于一个标量$t$的一元函数$g(t)=f(\boldsymbol{x}+t\,\boldsymbol{h}/\|\boldsymbol{h}\|)$，利用一元函数的二阶泰勒展开，有：

$$f(\boldsymbol{x}+\boldsymbol{h})=g(\|\boldsymbol{h}\|)=g(0)+g'(0)\|\boldsymbol{h}\|+\frac{g''(0)}{2}\|\boldsymbol{h}\|^2+\mathcal{R}=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}+\frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{h}}{2}+\mathcal{R} \qquad (4.77)$$

当变化距离$\|\boldsymbol{h}\|$趋近于 0 时，余项$\mathcal{R}$也趋近于 0，且比距离的平方消失得更快（$\|\boldsymbol{h}\|^2$的高阶无穷小）。这就证明了式（4.75）。

式（4.75）右边刨除余项的部分：

$$q(\boldsymbol{h})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}+\frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{h}}{2} \qquad (4.78)$$

$q(\boldsymbol{h})$是关于向量$\boldsymbol{h}$的二次函数。注意$q(\boldsymbol{h})$不是二次型。二次型只包含二次项，而$q(\boldsymbol{h})$包含常数（零次）项、一次项以及二次项。二次函数的典型图像如图 4-1。

图 4-1　二次函数的典型图像

$q(\boldsymbol{h})$的图像过$(0 \quad f(\boldsymbol{x}))$。将$q(\boldsymbol{h})$的图像平移，使$(0 \quad f(\boldsymbol{x}))$移动到$(\boldsymbol{x} \quad f(\boldsymbol{x}))$。平移后的图像是$f(\boldsymbol{x})$在$\boldsymbol{x}$附近的二阶近似。$q(\boldsymbol{h})$的前两项$f(\boldsymbol{x}) + \nabla f^{\mathrm{T}}\boldsymbol{h}$是仿射函数。它的全部信息包含在$f$的梯度中。它是$f(\boldsymbol{x})$的一阶近似。所以，二阶泰勒展开是对原函数精细到二阶的近似。误差信息包含在余项中。函数的局部一阶特性包含在梯度中，局部二阶特性包含在赫森矩阵中。如图 4-2 所示。

图 4-2　函数的局部二阶近似

4.2.3　驻点的类型

驻点到底是局部极小点，局部极大点或鞍点，这些信息一定程度上包含在函数的局部二阶特性中。假如$\boldsymbol{x}^*$是$f(\boldsymbol{x})$的驻点。令$\boldsymbol{H}(\boldsymbol{x})$是$f(\boldsymbol{x})$在$\boldsymbol{x}^*$的赫森矩阵。因为有$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$，根据二阶泰勒展开，离开$\boldsymbol{x}^*$的变化量为$\boldsymbol{h}$的点的函数值是：

$$f(\boldsymbol{x}^* + \boldsymbol{h}) = f(\boldsymbol{x}^*) + \frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(x)\boldsymbol{h}}{2} + \mathcal{R}(\boldsymbol{h}) \qquad (4.79)$$

令$\boldsymbol{d} = \boldsymbol{h}/\|\boldsymbol{h}\|$是与$\boldsymbol{h}$同方向的单位向量。假如$\boldsymbol{H}(\boldsymbol{x})$是正定的，有$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{d} > 0$。$\mathcal{R}(\boldsymbol{h})$是$\|\boldsymbol{h}\|^2$的高阶无穷小。所以当$\|\boldsymbol{h}\| \to 0$，式（4.79）右边后两项有：

$$\lim_{\|\boldsymbol{h}\|\to 0} \frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(x)\boldsymbol{h}/2 + \mathcal{R}(\boldsymbol{h})}{\|\boldsymbol{h}\|^2} = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{d} > 0 \qquad (4.80)$$

所以当$\boldsymbol{x} + \boldsymbol{h}$足够靠近$\boldsymbol{x}$时，有$f(\boldsymbol{x}^* + \boldsymbol{h}) > f(\boldsymbol{x}^*)$。$\boldsymbol{x}^*$是局部极小点。注意，赫森矩阵必须为正定。如果仅仅是半正定，那么式（4.90）的极限只能是大于等于 0，那么自变量顺着某个方向可以从负侧接近 0，不能保证$\boldsymbol{x}^*$局部极小。类似地，如果$\boldsymbol{H}(\boldsymbol{x})$是负定的，$\boldsymbol{x}^*$是局部极大点。

正定矩阵特征值都为正，负定矩阵所有特征值都为负。如果特征值有正有负，则矩阵是不定的。如果$\boldsymbol{H}(\boldsymbol{x})$不定，则沿着正特征值的特征方向二阶导数为正，函数向上翘；沿着负特征值方向，函数向下弯。此时在任何领域内，都有函数值大于$f(\boldsymbol{x}^*)$的点，也有函数值小于$f(\boldsymbol{x}^*)$的点，$\boldsymbol{x}^*$是一个鞍点。如图 4-3 所示。

图 4-3　驻点的类型与赫森矩阵的关系

总结如下：

- 驻点$\boldsymbol{x}^*$的赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$正定，$\boldsymbol{x}^*$是局部极小点；
- 驻点$\boldsymbol{x}^*$的赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$负定，$\boldsymbol{x}^*$是局部极大点；
- 驻点$\boldsymbol{x}^*$的赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$不定，$\boldsymbol{x}^*$是鞍点；
- 驻点$\boldsymbol{x}^*$的赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$半正定但非正定，或半正定但非负定，$\boldsymbol{x}^*$的性质无法确定；

如果不考虑最后一种情况，那么只有当赫森矩阵是正定的，也就是其特征值全部大于 0 时，驻点是局部极小点。在没有任何先验知识时可以假设特征值大于 0 的概率是 0.5 且彼此独立。那么如果模型有n个参数，则损失函数的赫森矩阵的特征值全部大于 0 的概率是$1/2^n$。这是一个极小的概率。所以在例如神经网络或深度学习这种参数很多的模型中，局部极小点应该是很罕见的。

函数的局部二阶特性提供了比一阶特性更多的信息。但如果赫森矩阵只是“半定”的，则仍然没有足够的信息完全判断驻点性质，这时就需要更精确的近似。

在梯度下降法中，梯度反方向$-\nabla f$虽然是下降最快的方向，但如果知道$\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\nabla f(\boldsymbol{x})/\|\nabla f(\boldsymbol{x})\|^2 > 0$且绝对值很大，则沿着$-\nabla f(\boldsymbol{x})$方向，（方向）导数上升且升得很快。这时步长如果较大，则有可能一步之后函数值不降反升。如果在选择前进方向时不仅参考梯度$\nabla f(\boldsymbol{x})$，也参考赫森矩阵$\boldsymbol{H}(\boldsymbol{x})$，优化算法能做出更优的选择。后文将要讲解的牛顿法和共轭方向法就是如此。

4.2.4　条件数与病态峡谷

令$q(x_1, x_2)$是一个二元二次函数：

$$q(x_1, x_2) = a + b_1 x_1 + b_2 x_2 + \frac{c_{11}}{2}{x_1}^2 + \frac{c}{2}x_1 x_2 + \frac{c_{22}}{2}{x_2}^2 = a + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x} \tag{4.81}$$

其中$\boldsymbol{x} = (x_1 \quad x_2)^{\mathrm{T}}$，$\boldsymbol{b} = (b_1 \quad b_2)^{\mathrm{T}}$，$\boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$，$c_{12} = c_{21} = c/2$。考察$q(x_1, x_2)$的梯度和赫森矩阵：

$$\nabla q = \begin{pmatrix} \frac{\partial q}{\partial x_1} \\ \frac{\partial q}{\partial x_2} \end{pmatrix} = \begin{pmatrix} b_1 + c_{11}x_1 + c_{12}x_2 \\ b_2 + c_{21}x_1 + c_{22}x_2 \end{pmatrix} = \boldsymbol{b} + \boldsymbol{C}\boldsymbol{x} \tag{4.82}$$

q的赫森矩阵是：

$$\boldsymbol{H}=\begin{pmatrix}\frac{\partial q}{\partial x_1\partial x_1} & \frac{\partial q}{\partial x_2\partial x_1}\\ \frac{\partial q}{\partial x_1\partial x_2} & \frac{\partial q}{\partial x_2\partial x_2}\end{pmatrix}=\begin{pmatrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{pmatrix}=\boldsymbol{C} \tag{4.83}$$

q的二阶泰勒展开就是它自身，余项为 0。即二次函数的二阶泰勒展开就精确地是它自身。二次函数的赫森矩阵是常矩阵$\boldsymbol{C}$，它的二阶特性是处处相同的。

现在假设q的$\boldsymbol{C}$是正定的，它的所有特征值为正。容易验证$\boldsymbol{C}$可逆。$\boldsymbol{C}$的逆矩阵是：

$$\boldsymbol{C}^{-1}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}=\boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\frac{1}{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_n}\end{pmatrix}\boldsymbol{V} \tag{4.84}$$

$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{V}$是$\boldsymbol{C}$的谱分解。

因为$\boldsymbol{C}$可逆，所以 q的唯一驻点是$\boldsymbol{x}^*=-\boldsymbol{C}^{-1}\boldsymbol{b}$是一个局部极小点。$q$在驻点$\boldsymbol{x}^*$上梯度为 0，则$q$沿任意方向的方向导数为 0。但是因为$\boldsymbol{C}$正定，所以$q$在$\boldsymbol{x}^*$沿任意方向的二阶导是$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{d}>0$，也就是沿任意方向的方向导数持续增大，过$\boldsymbol{x}^*$之前为负，过$\boldsymbol{x}^*$之前为正。也就是沿任意方向，$q$都是先下降，在$\boldsymbol{x}^*$降到最低然后上升。$q$的图像呈碗状。因为$\boldsymbol{x}^*$是唯一的局部极小点，所以它是$q$的全局最小点。

q的图形沿任意方向都是两端上升的抛物线，但上升的速度不一样。二阶导较大的点，方向导变化较快，沿该方向的抛物线上升更快；二阶导较小的点，方向导变化较慢，沿该方向的抛物线上升更慢。

使二阶导$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{d}$最大的单位向量$\boldsymbol{d}$是$\boldsymbol{C}$的最大特征值$\lambda_1$对应的特征向量$\boldsymbol{v}_1$。在此方向上$q$具有最大的二阶导$\lambda_1$。垂直于$\boldsymbol{v}_1$且使二阶导$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{d}$最大的单位向量是$\boldsymbol{C}$的次大特征值$\lambda_2$（在二元函数情况下就是最下特征值）。在此方向上$q$具有最大的二阶导$\lambda_2$。

称λ_1/λ_2是$\boldsymbol{C}$的条件数。条件数越大，最大特征值与最小特征值之比越大，在最大特征值的特征方向上函数图像的弯曲程度比垂直于该方向上函数图像的弯曲程度就更加剧烈。所以二次函数的赫森矩阵的条件数越大，则函数图像成越狭长的峡谷。这时候称二次函数的图像是病态的。反之若$\lambda_1/\lambda_2=1$，则函数图像超各个方向的弯曲程度是相同的，图像成完美的碗状。如图 4-4 所示。

图 4-4　条件数与二次函数图像的狭长程度

狭长的峡谷称为病态的，它对梯度下降算法尤其不利。在第 3 章中已经看到在这种情况下，不合适的步长会引发震荡甚至不收敛。现在我们从某一个角度提供一个洞见。如果$\boldsymbol{C}$的条件数是 1，则$\lambda_1 = \lambda_2$，令它们都等于λ。则有：

$$\boldsymbol{C}^{-1} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{V} = \boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\frac{1}{\lambda} & 0\\ 0 & \frac{1}{\lambda}\end{pmatrix}\boldsymbol{V} = \frac{1}{\lambda}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \frac{1}{\lambda}\mathbf{I} \qquad (4.85)$$

那么q在的驻点$\boldsymbol{x}^*$是：

$$\boldsymbol{x}^* = -\boldsymbol{C}^{-1}\boldsymbol{b} = -\frac{1}{\lambda}\boldsymbol{b} \qquad (4.86)$$

从任意点$\boldsymbol{x}$指向驻点$\boldsymbol{x}^*$的“箭头”是：

$$\boldsymbol{x}^* - \boldsymbol{x} = -\frac{1}{\lambda}\boldsymbol{b} - \boldsymbol{x} \qquad (4.87)$$

而q在$\boldsymbol{x}$的负梯度是：

$$-\nabla q(\boldsymbol{x}) = -(\boldsymbol{b} + \boldsymbol{C}\boldsymbol{x}) = -(\boldsymbol{b} + \lambda\boldsymbol{x}) = \lambda(\boldsymbol{x}^* - \boldsymbol{x}) \qquad (4.88)$$

式（4.98）的含义是，在任一点$\boldsymbol{x}$，其梯度正指向朝着$\boldsymbol{x}^*$的方向。于是每一步梯度下降都是朝着正确的方向前进。当然如果步长过大，有可能一步跨过$\boldsymbol{x}^*$，发生震荡。步长更大时，仍有可能发生不收敛。

现实中损失函数作为参数值的函数不是二次的，但是在驻点附近二次函数是很好的近似。本节分析的二次函数的特性，可以近似地表示任何函数在局部的特性。

4.3 基于函数二阶特性的优化

如果参考函数二阶特性信息，则优化过程会更有效率。但是高维情况下赫森矩阵的计算量是巨大的，这妨碍了二阶优化算法在神经网络或深度学习中的应用。人们提出了一些节省计算量的近似二阶算法，本书不对其进行讨论。本节介绍两个基于赫森矩阵的优化算法——牛顿法和共轭方向法，并探讨它们的几何意义。

4.3.1 牛顿法

牛顿法的思想是：在某点附近对函数进行二阶泰勒展开，之后求函数的近似二次函数的驻点，以该点为下一个点，迭代地进行此步骤。根据式（4.75），离开$\boldsymbol{x}$一个变化量$\boldsymbol{h}$的函数值$f(\boldsymbol{x}+\boldsymbol{h})$的二次近似函数是：

$$f(\boldsymbol{x}+\boldsymbol{h}) \approx q(\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h} + \frac{\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\boldsymbol{h}}{2} \qquad (4.89)$$

$q(\boldsymbol{h})$的梯度是：

$$\nabla q(\boldsymbol{h}) = \nabla f(\boldsymbol{x}) + \boldsymbol{H}(\boldsymbol{x})\boldsymbol{h} \qquad (4.90)$$

如果$\boldsymbol{H}(\boldsymbol{x})$的特征值都非 0，则利用式（4.84）可以构造它的逆矩阵。特征值取某个特定值，例如 0 的概率很低，也就是绝大部分情况下都可以认为$\boldsymbol{H}(\boldsymbol{x})$是可逆的。令$\nabla q(\boldsymbol{h}) = \boldsymbol{0}$，解得$q(\boldsymbol{h})$的驻点是：

$$\boldsymbol{h}^* = -\boldsymbol{H}(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x}) \qquad (4.91)$$

所以，牛顿法对自变量$\boldsymbol{x}$的迭代更新量就是$-\boldsymbol{H}(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$。牛顿法伪代码如下：

$\boldsymbol{x} \leftarrow$ randomly initialized

while $\|\nabla f(\boldsymbol{x})\| \geq \varepsilon$:

$\quad \boldsymbol{x} \leftarrow \boldsymbol{x} - \boldsymbol{H}(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$

与梯度下降法一样，$\varepsilon > 0$是一个预设的阈值，当$\|\nabla f(\boldsymbol{x})\| < \varepsilon$时认为$\nabla f(\boldsymbol{x})$已经足够接近零向量，算法停止。也可以采用循环次数达到预设的最大值，或者函数值的下降幅度小于阈值等停止标准。

梯度下降法，向着反梯度方向$-\nabla f(\boldsymbol{x})$更新自变量。而牛顿法将$-\nabla f(\boldsymbol{x})$乘上了赫森矩阵的逆矩阵$\boldsymbol{H}(\boldsymbol{x})^{-1}$。假设$\boldsymbol{H}(\boldsymbol{x})$是正定的，根据对称矩阵的谱分解，$\boldsymbol{H}(\boldsymbol{x})^{-1}$可以分解成：

$$\boldsymbol{H}(\boldsymbol{x})^{-1}=\boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\frac{1}{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_n}\end{pmatrix}\boldsymbol{V} \qquad (4.92)$$

$\boldsymbol{V}^{\mathrm{T}}$是由$\boldsymbol{H}(\boldsymbol{x})$的特征向量构成的正交矩阵。$\lambda_1,\lambda_2,\cdots,\lambda_n$是从大到小排列的$\boldsymbol{H}(\boldsymbol{x})$的$n$个特征值（可重）。牛顿算法的更新量可以写成：

$$-\boldsymbol{H}(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})=-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{V}\nabla f(\boldsymbol{x})=-\boldsymbol{V}^{\mathrm{T}}\begin{pmatrix}\frac{1}{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_n}\end{pmatrix}\boldsymbol{V}\nabla f(\boldsymbol{x}) \qquad (4.93)$$

$\boldsymbol{V}^{\mathrm{T}}$的列（$\boldsymbol{V}$的行）是$\mathbb{R}^n$的一组标准正交基，也就是一个坐标系。该坐标系的第一个轴沿着$f(\boldsymbol{x})$二阶导最大的方向；第二个轴沿着垂直于第一个轴且$f(\boldsymbol{x})$二阶导第二大的方向。向量$\boldsymbol{V}\nabla f(\boldsymbol{x})$相当于对$\nabla f(\boldsymbol{x})$做了一次旋转。如果忽略$\boldsymbol{\Lambda}$矩阵，$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\nabla f(\boldsymbol{x})=\nabla f(\boldsymbol{x})$是将$\nabla f(\boldsymbol{x})$旋转过去又旋转回来。

加上$\boldsymbol{\Lambda}$的作用后，相当于用$1/\lambda_i$惩罚 $\boldsymbol{H}(\boldsymbol{x})$的各个特征方向上$\nabla f(\boldsymbol{x})$的分量。特征值$\lambda_i$越大，其对应的特征方向上$f(\boldsymbol{x})$的二阶导越大。$1/\lambda_i$越小，则对$\nabla f(\boldsymbol{x})$在该特征方向上的分量惩罚越严厉。还记得狭长峡谷中梯度下降法的震荡发生的原因，就是二阶导较大的方向主宰了梯度方向，导致算法没有想着理想的局部极小点方向前进。牛顿法通过对二阶导最大的方向进行适当惩罚，缓解了这种现象。如图 4-5 所示。

图 4-5　牛顿法相当于对二阶导过大的方向进行惩罚

如果函数本身是二次函数，它的二阶泰勒展开就是它自身。牛顿法可以直接定位到二次近似的驻点。也就是说对于二次函数，牛顿法一步迭代就能寻找到它的唯一驻点——全局最小点。

4.3.2　共轭方向法

对于一个二次函数$f(\boldsymbol{x})$，它在任意点的赫森矩阵是$\boldsymbol{H}$。假设$\boldsymbol{H}$正定。对于任意两个单位向量$\boldsymbol{d}_1$和$\boldsymbol{d}_2$，如果满足：

$$\boldsymbol{d}_1^{\mathrm{T}}\,\boldsymbol{H}\boldsymbol{d}_2=0 \qquad (4.94)$$

则称$\boldsymbol{d}_1$与$\boldsymbol{d}_2$关于 $\boldsymbol{H}$共轭。令自变量从某一个点$\boldsymbol{x}$开始，沿着$\boldsymbol{d}_1$运动，可以构造一个一元二次

函数：

$$g(h)=f(\boldsymbol{x}+h\boldsymbol{d}_1)=f(\boldsymbol{x})+h\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}_1+\frac{h^2}{2}\boldsymbol{d}_1^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_1 \qquad (4.95)$$

一元函数$g(h)$的导数是：

$$g'(h)=\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}_1+h\boldsymbol{d}_1^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_1 \qquad (4.96)$$

一元函数$g(h)$的二阶导数是：

$$g''(h)=\boldsymbol{d}_1^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_1>0 \qquad (4.97)$$

因为$\boldsymbol{H}$正定，所以二次函数$g(h)$有全局最小点：

$$h^*=-\frac{\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}_1}{\boldsymbol{d}_1^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_1} \qquad (4.98)$$

因为h^*是$g(h)$的全局最小点，所以$g'(h^*)=0$，即$f(\boldsymbol{x})$在$\boldsymbol{x}+h^*\boldsymbol{d}_1$沿$\boldsymbol{d}_1$的方向导数为0。

$$\nabla_{\boldsymbol{d}_1}f(\boldsymbol{x}+h^*\boldsymbol{d}_1)=\nabla f(\boldsymbol{x}+h^*\boldsymbol{d}_1)^T\boldsymbol{d}_1=0 \qquad (4.99)$$

$\boldsymbol{d}_1$与$\boldsymbol{d}_2$关于$\boldsymbol{H}$共轭。

令自变量从$\boldsymbol{x}+h^*\boldsymbol{d}_1$运动一个变化量$\boldsymbol{h}$。利用二阶泰勒展开，$f(\boldsymbol{x}+h^*\boldsymbol{d}_1+\boldsymbol{h})$作为$\boldsymbol{h}$的方程是：

$$f(\boldsymbol{x}+h^*\boldsymbol{d}_1+\boldsymbol{h})=f(\boldsymbol{x}+h^*\boldsymbol{d}_1)+\nabla f(\boldsymbol{x}+h^*\boldsymbol{d}_1)^T\boldsymbol{h}+\frac{1}{2}\boldsymbol{h}^T\boldsymbol{H}\boldsymbol{h} \qquad (4.100)$$

$f(\boldsymbol{x}+h^*\boldsymbol{d}_1+\boldsymbol{h})$对$\boldsymbol{h}$的梯度是：

$$\nabla f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1+\boldsymbol{h})=\nabla f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1)+\boldsymbol{H}\boldsymbol{h} \tag{4.101}$$

$f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1+\boldsymbol{h})$沿$\boldsymbol{d}_1$的方向导数是：

$$\nabla_{\boldsymbol{d}_1} f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1+\boldsymbol{h})=\nabla f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1+\boldsymbol{h})^{\mathrm{T}}\boldsymbol{d}_1=\nabla f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1)^{\mathrm{T}}\boldsymbol{d}_1+\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_1 \tag{4.102}$$

如果变化量$\boldsymbol{h}$是沿着$\boldsymbol{d}_2$方向，即$\boldsymbol{h}=h\boldsymbol{d}_2$。根据式（4.99）以及$\boldsymbol{d}_1$与$\boldsymbol{d}_2$关于 $\boldsymbol{H}$共轭，有：

$$\nabla_{\boldsymbol{d}_2} f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1+h\boldsymbol{d}_2)=\nabla f(\boldsymbol{x}+h^{*}\boldsymbol{d}_1)^{\mathrm{T}}\boldsymbol{d}_1+h{\boldsymbol{d}_1}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{d}_2=0 \tag{4.103}$$

也就是说，自变量从$\boldsymbol{x}$出发沿着