目录

Yuan and Song

Yuan and Song

第零章　绪言

0.1　引入

[** 本章介绍若干机器学习的引入知识 **]

Yuan and Song

第一部分

基础理论

Yuan and Song

第一章　微积分与线性代数基础

1.1　多元函数及其导数

多元函数（multivariate function）是机器学习理论的基本工具，其定义如下：

定义 1.1.1 (多元函数) 设 X 为实数集 $\mathbb{R}^n$ 中的一个子集，如果存在一个对应法则 f，使得对于任意 $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in X$，都存在唯一 $y\in\mathbb{R}$ 与之对应，则称 f 为定义在 X 上，（陪域为 $\mathbb{R}$）的 n 元函数，记作：

$$f:X\to\mathbb{R};\quad \boldsymbol{x}\mapsto y=f(\boldsymbol{x})=f(x_1,x_2,\ldots,x_n).$$

我们在本节中将讨论多元函数的偏导数、梯度、二阶导数及海瑟矩阵的定义.

1.1.1 偏导数

对于一元函数，我们可以使用导数来研究其局部的特性. 而对于多元函数，我们无法直接一次对所有变量进行求导. 取而代之的是，我们会取定某一个分量 x_i，而将其余分量视为常数，对分量 x_i 求导数. 这样求得的导数可以反映函数 f 在 x_i 方向上的局部特性. 我们将该导数称为函数 f 关于 x_i 的偏导数（partial derivative），其严格定义如下：

定义 1.1.2 (偏导数) 对于定义在 $X\subseteq\mathbb{R}^n$ 上的多元函数 $f(x_1,\ldots,x_n)$，给定一点 $\boldsymbol{x}^{(0)}=(x_1^{(0)},\ldots,x_n^{(0)})\in X$，若以下极限存在且有限

$$\ell=\lim_{\Delta x\to 0}\frac{f(x_1^{(0)},\ldots,x_i^{(0)}+\Delta x,\ldots,x_n^{(0)})-f(x_1^{(0)},\ldots,x_i^{(0)},\ldots,x_n^{(0)})}{\Delta x},$$

则称 ℓ 为多元函数 f 在 $\boldsymbol{x}^{(0)}$ 点处关于第 i 个分量的（单点）偏导数，记为 $f_i'(x_1^{(0)},\ldots,x_n^{(0)})$. 进一步，对于任意一个给定点 $(x_1^{(0)},\ldots,x_n^{(0)})$，均存在唯一的 $f_i'(x_1^{(0)},\ldots,x_n^{(0)})$ 与之对应. 因此，考虑所有 $\boldsymbol{x}\in X$，f 关于第 i 个分量的偏导数构成了一个 n 元函数，我们称其为偏导函数，简称为偏导数，记为 $f_i'(x_1,\ldots,x_n)$.

Yuan and Song

注意偏导数的定义域可能不为 X, 而是 x 的一个真子集. 关于偏导数定义域的分析需要考虑不可导情形及边缘单侧导数的情形，不在本书所讨论的范围之内. 本书中我们一般不讨论偏导数的严格定义域，或默认定义域与原函数的定义域 X 相同.

小贴士

请注意，偏导数虽然聚焦于某一分量 x_i, 但其仍然为关于整体 n 元变量 $(x_1,\dots,x_n)$ 的多元函数，而非仅关于 x_i 的一元函数，这点尤须注意.

偏导数有除 $f_i'(x_1,\dots,x_n)$ 外的多种写法. 如果自变量的符号固定，譬如 $(x,y)\in\mathbb{R}^2$, 则关于 x,y 的偏导数可分别写作 $f_x'(x,y), f_y'(x,y)$. 如果因变量的符号同样固定，譬如 $l=f(x,y)$, 则关于 x,y 的偏导数可分别写作 $\dfrac{\partial l}{\partial x}(x,y), \dfrac{\partial l}{\partial y}(x,y)$. 在不引起歧义的情况下，还可以省略自变量标记，直接简记为 $\dfrac{\partial l}{\partial x}, \dfrac{\partial l}{\partial y}$, 在此情况下，如果要强调在某一点 $(x^{(0)},y^{(0)})$ 的单点导数，可以使用角标形式，即 $\left.\dfrac{\partial l}{\partial x}\right|_{(x^{(0)},y^{(0)})}, \left.\dfrac{\partial l}{\partial y}\right|_{(x^{(0)},y^{(0)})}$. 在因变量不固定时，有时也使用函数对应法则 f 来代指因变量，即 $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}$. 在深度学习相关的数学理论中，使用偏导数记号 $\dfrac{\partial\cdot}{\partial\cdot}$ 的形式反而更为常见.

1.1.2 梯度

如果一个多元函数关于所有分量的偏导数均已知，我们可以很自然地引出梯度（gradient）的定义. 与上文相同，我们同样只考虑偏导数在整个定义域中均存在的简化情形. 此外，为确保方向导数良定义，我们进一步假设偏导数为连续函数，这是梯度存在的充分不必要条件，但这一条件在机器学习所研究的问题中一般都是满足的.

定义 1.1.3 (梯度) *对于定义在 $X\subseteq\mathbb{R}^n$ 上的多元函数 $f(x_1,\dots,x_n)$, 若其所有偏导数 $f_i'(\boldsymbol{x}), i=1,2,\dots,n$ 在定义域内均存在且为连续函数，则函数 f 的梯度定义为由所有分量对应的偏导数组成的列向量（或行向量），即：*

$$\nabla f(\boldsymbol{x}) := (f_1'(\boldsymbol{x}), f_2'(\boldsymbol{x}), \dots, f_n'(\boldsymbol{x}))^\top.$$

其中符号“∇”称为哈密顿算子，读作“Nabla”. 除上述表示方式外，梯度也可以有其他表示方法，如 $\mathrm{grad} f(\boldsymbol{x})$, 也可以直接借用一元函数导数的记号 $f'(\boldsymbol{x})$.

小贴士

梯度的列向量与行向量形式分别具有不同的用途，列向量形式用于代入梯度下降等算法中，而行向量形式则用于向量值函数求导及链式法则中. 后续内容中，为了不引起混淆，我们会分别使用不同的记号来表示行、列梯度.

在深度学习任务中，由于神经网络通常以分层的形式呈现，因此变量也通常采用分块的形式，我们考虑一个简单的两块的情形，即 $\boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{y} \in \mathbb{R}^q$, 函数 $f(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^{p+q} \to \mathbb{R}$, 我们可以求得如下的分块的偏导数：

$$f'_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{y}) := \big(f'_{x_1}(\boldsymbol{x}, \boldsymbol{y}), \ldots, f'_{x_p}(\boldsymbol{x}, \boldsymbol{y})\big)^\top, \quad f'_{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{y}) := \big(f'_{y_1}(\boldsymbol{x}, \boldsymbol{y}), \ldots, f'_{y_q}(\boldsymbol{x}, \boldsymbol{y})\big)^\top.$$

同样地，如果因变量的符号固定为 l, 分块偏导数可以写为 $\dfrac{\partial l}{\partial \boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{y}), \dfrac{\partial l}{\partial \boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{y})$, 或省略自变量标记直接写为 $\dfrac{\partial l}{\partial \boldsymbol{x}}, \dfrac{\partial l}{\partial \boldsymbol{y}}$, 后者是深度学习中最常用的形式. 注意分块导数和梯度一样，均为向量的形式，因此分块导数有时也借用梯度的符号，记为 $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$.

下面我们通过几个例子来演示函数梯度的求解方法：

例 1.1.4 设函数 $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{2}\big(x_1^2 + x_2^2 + \cdots + x_n^2\big),$$

其关于每个分量的偏导数为：

$$f'_i(\boldsymbol{x}) = x_i,$$

因此整个函数的梯度为：

$$\nabla f(\boldsymbol{x}) = (x_1, \ldots, x_n)^\top = \boldsymbol{x}.$$

例 1.1.5 设函数 $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b,$$

其中 $\boldsymbol{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$. 其关于第 i 个分量的偏导数为：

$$f'_i(\boldsymbol{x}) = a_i.$$

因此梯度为：

$$\nabla f(\boldsymbol{x}) = (a_1, \ldots, a_n)^\top = \boldsymbol{a}.$$

Yuan and Kong

例 1.1.6 设函数 $f:\mathbb{R}^n\to\mathbb{R}$,

$$f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x}+\boldsymbol{b}^\top\boldsymbol{x}+c,$$

其中 $A=(a_{ij})\in\mathbb{R}^{n\times n}$. 其关于第 i 个分量的偏导数为:

$$f_i'(\boldsymbol{x})=\frac{1}{2}\sum_{j=1}^{n}(a_{ij}+a_{ji})x_j+b_i.$$

因此梯度为:

$$\nabla f(\boldsymbol{x})=\frac{1}{2}(A+A^\top)\boldsymbol{x}+\boldsymbol{b}.$$

特别地，当 A 为对称矩阵时,

$$\nabla f(\boldsymbol{x})=A\boldsymbol{x}+\boldsymbol{b}.$$

例 1.1.7 设函数 $f:\mathbb{R}^n\to\mathbb{R}$,

$$f(x_1,x_2,\ldots,x_n)=\ln\left(x_1^2+x_2^2+\cdots+x_n^2\right),$$

其关于每个分量的偏导数为:

$$f_i'(\boldsymbol{x})=\frac{2x_i}{x_1^2+\cdots+x_n^2},$$

因此整个函数的梯度为:

$$\nabla f(\boldsymbol{x})=\frac{1}{x_1^2+\cdots+x_n^2}(2x_1,\ldots,2x_n)^\top=\frac{2}{\boldsymbol{x}^\top\boldsymbol{x}}\boldsymbol{x}.$$

例 1.1.8 设函数 $f:\mathbb{R}^n\to\mathbb{R}$,

$$f(\boldsymbol{x})=\exp\left(\boldsymbol{a}^\top\boldsymbol{x}\right),$$

其中 $\boldsymbol{a}\in\mathbb{R}^n$. 其关于第 i 个分量的偏导数为:

$$f_i'(\boldsymbol{x})=a_i\exp\left(\boldsymbol{a}^\top\boldsymbol{x}\right).$$

因此梯度为:

$$\nabla f(\boldsymbol{x})=\exp\left(\boldsymbol{a}^\top\boldsymbol{x}\right)\boldsymbol{a}.$$

例 1.1.9 设函数 $f:\mathbb{R}^n\to\mathbb{R}$,

$$f(\boldsymbol{x})=\ln\left(\sum_{j=1}^{n}e^{x_j}\right).$$

Yuan and Song

记

$$S(\boldsymbol{x}) = \sum_{j=1}^{n} e^{x_j}.$$

则其关于第 i 个分量的偏导数为：

$$f_i'(\boldsymbol{x}) = \frac{e^{x_i}}{S(\boldsymbol{x})} = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}.$$

因此梯度为：

$$\nabla f(\boldsymbol{x}) = \left(\frac{e^{x_1}}{\sum_{j=1}^{n} e^{x_j}}, \ldots, \frac{e^{x_n}}{\sum_{j=1}^{n} e^{x_j}} \right)^{\top}.$$

该梯度正是 *softmax* 函数的输出.

例 1.1.10 设样本 $(\boldsymbol{x}, y)$ 满足

$$\boldsymbol{x} \in \mathbb{R}^n, \qquad y \in \{0, 1\}.$$

定义模型输出

$$p(\boldsymbol{x}) = \sigma\left(\boldsymbol{w}^{\top}\boldsymbol{x}\right) = \frac{1}{1 + \exp\left(-\boldsymbol{w}^{\top}\boldsymbol{x}\right)}.$$

二元逻辑回归的交叉熵损失为

$$\ell(\boldsymbol{w}) = -y \ln p(\boldsymbol{x}) - (1 - y) \ln\big(1 - p(\boldsymbol{x})\big).$$

由于

$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big),$$

并且

$$\frac{\partial(\boldsymbol{w}^{\top}\boldsymbol{x})}{\partial \boldsymbol{w}} = \boldsymbol{x},$$

根据链式法则可得

$$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) = \big(p(\boldsymbol{x}) - y\big)\boldsymbol{x}.$$

也就是

$$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) = \left(\sigma\left(\boldsymbol{w}^{\top}\boldsymbol{x}\right) - y\right)\boldsymbol{x}.$$

例 1.1.11 对于数据集

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N},$$

定义平均损失

$$L(\boldsymbol{w}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \right],$$

其中

$$p_i = \sigma\left(\boldsymbol{w}^{\top}\boldsymbol{x}_i\right).$$

Yuan and Song

因此

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)\boldsymbol{x}_i$$

若将样本组成矩阵

$$X = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{pmatrix}, \qquad \boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},$$

则梯度可以写成紧凑形式：

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{1}{N} X^\top (\boldsymbol{p} - \boldsymbol{y}).$$

1.1.3 二阶导数与海瑟矩阵

对于一元函数，我们可以使用一阶导数来探究函数的增减性和斜率，同时进一步可以使用二阶导数来探究函数的凹凸性和曲率. 对于多元函数，我们也希望类似地讨论其二阶导数.

对于函数 $f : \mathbb{R}^n \to \mathbb{R}$, 我们可以求出 n 个偏导数 $f_1'(\boldsymbol{x}), \ldots, f_n'(\boldsymbol{x})$, 它们均为定义在 $\mathbb{R}^n$ 上的多元函数. 进一步，我们可以对这 n 个多元函数再分别求偏导数，并称其为二阶偏导数. 二阶偏导数可记为 $f_{ij}''; i, j = 1, \ldots, n$, 当自变量和因变量分别固定为 $\boldsymbol{x}, y$ 时，也可记为 $\dfrac{\partial^2 y}{\partial x_i \partial x_j}(\boldsymbol{x})$. 函数 f 的二阶偏导数共计有 n^2 个，我们将这 n^2 个二阶偏导数构成的 $n \times n$ 矩阵称为函数 f 的海瑟矩阵（Hessian matrix），其严格定义如下：

定义 1.1.12 (海瑟矩阵) 对于定义在 $X \subseteq \mathbb{R}^n$ 上的多元函数 $f(x_1, \ldots, x_n)$, 若其所有二阶偏导数 $f_{i,j}''(\boldsymbol{x}), i, j = 1, 2, \ldots, n$ 在定义域内均存在且为连续函数，则函数 f 的海瑟矩阵定义为由所有二阶偏导数构成的矩阵，即：

$$H_f(\boldsymbol{x}) := \begin{pmatrix} f_{11}''(\boldsymbol{x}) & \cdots & f_{1n}''(\boldsymbol{x}) \\ \vdots & & \vdots \\ f_{n1}''(\boldsymbol{x}) & \cdots & f_{nn}''(\boldsymbol{x}) \end{pmatrix}.$$

海瑟矩阵也可以有多种记法，可借用梯度符号的记号记作 $\nabla^2 f(\boldsymbol{x})$, 也可借用一元函数导数的记号记为 $f''(\boldsymbol{x})$.

Yuan and Song

小贴士

根据克莱罗定理（或施瓦茨定理），如果函数的二阶偏导数连续（深度学习中的问题基本均满足这一条件），那么混合偏导数与求导次序无关，即：

$$f''_{ij}(\boldsymbol{x}) = f''_{ji}(\boldsymbol{x}).$$

此时海瑟矩阵是一个**实对称矩阵**.

我们同样给出若干实例来演示海瑟矩阵的求法.

例 1.1.13 对于常数矩阵 $A = (a_{ij}) \in \mathbb{R}^{n\times n}$（不一定为对称矩阵），定义函数 $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x}.$$

我们先求其关于第 i 个分量的一阶偏导数. f 中关于 x_i 所有项为：

$$\frac{1}{2}a_{ii}x_i^2 + \frac{1}{2}\sum_{k=1,k\neq i}^{n}(a_{ik}+a_{ki})x_kx_i,$$

因此关于 x_i 的偏导数为

$$f'_i(\boldsymbol{x}) = a_{ii}x_i + \frac{1}{2}\sum_{k=1,k\neq i}^{n}(a_{ik}+a_{ki})x_k = \frac{1}{2}\sum_{k=1}^{n}(a_{ik}+a_{ki})x_k.$$

进一步考虑第 i 个偏导数如何求关于 x_j 的二阶偏导数，由于其为线性函数，我们可以很容易得到

$$f''_{ij} = \frac{1}{2}(a_{ij}+a_{ji}).$$

整合所有二阶偏导数，我们可以求得函数 f 的海瑟矩阵为：

$$H_f(\boldsymbol{x}) = \frac{1}{2}(A + A^\top),$$

特别地，如果 A 本身为实对称矩阵，则 $H_f(\boldsymbol{x}) = A$.

例 1.1.14 继续考虑例*1.1.7*,

$$f(\boldsymbol{x}) = \ln(x_1^2 + \cdots + x_n^2),$$

每个分量的偏导数为

$$f'_i(\boldsymbol{x}) = \frac{2x_i}{x_1^2 + \cdots + x_n^2}.$$

对于二阶偏导数，我们分别考虑 $j = i$ 与 $j \neq i$ 的情形：当 $j = i$ 时，

$$f''_{ii}(\boldsymbol{x}) = \frac{2(x_1^2 + \cdots + x_n^2) - 4x_i^2}{(x_1^2 + \cdots + x_n^2)^2};$$

Yuan and Song

当 $j \neq i$ 时，

$$f''_{ij}(\boldsymbol{x}) = \frac{-4x_i x_j}{(x_1^2 + \cdots + x_n^2)^2}.$$

因此我们可以整合得到该函数的海瑟矩阵为

$$H_f(\boldsymbol{x}) = \frac{2(\boldsymbol{x}^\top \boldsymbol{x})I - 4\boldsymbol{x}\boldsymbol{x}^\top}{(\boldsymbol{x}^\top \boldsymbol{x})^2}.$$

例 1.1.15 设训练数据为

$$X \in \mathbb{R}^{N\times d}, \qquad \boldsymbol{y} \in \mathbb{R}^N,$$

考虑线性回归的平方损失函数

$$f(\boldsymbol{w}) = \frac{1}{2N} \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2.$$

将其展开为

$$f(\boldsymbol{w}) = \frac{1}{2N} \left(\boldsymbol{w}^\top X^\top X\boldsymbol{w} - 2\boldsymbol{y}^\top X\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y}\right).$$

根据二次型函数的求导公式，函数梯度为

$$\nabla f(\boldsymbol{w}) = \frac{1}{N} X^\top \left(X\boldsymbol{w} - \boldsymbol{y}\right).$$

进一步求其二阶偏导数，得到海瑟矩阵

$$H_f(\boldsymbol{w}) = \frac{1}{N} X^\top X.$$

注意到

$$H_f(\boldsymbol{w}) \succeq 0,$$

因此线性回归的平方损失函数是凸函数. 当 X 列满秩时，

$$H_f(\boldsymbol{w}) \succ 0,$$

此时损失函数严格凸.

例 1.1.16 考虑带有 L_2 正则项的岭回归目标函数

$$f_\lambda(\boldsymbol{w}) = \frac{1}{2N} \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2,$$

其中 $\lambda \geq 0$ 为正则化参数. 根据前一个例子，第一项的梯度为

$$\frac{1}{N} X^\top \left(X\boldsymbol{w} - \boldsymbol{y}\right),$$

而正则项的梯度为

$$\nabla \left(\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right) = \lambda \boldsymbol{w}.$$

因此

$$\nabla f_\lambda(\boldsymbol{w}) = \frac{1}{N} X^\top \left(X\boldsymbol{w} - \boldsymbol{y}\right) + \lambda \boldsymbol{w}.$$

Yuan and Song

进一步得到海瑟矩阵

$$H_{f_\lambda}(\boldsymbol{w}) = \frac{1}{N} X^\top X + \lambda I.$$

当 $\lambda > 0$ 时，

$$H_{f_\lambda}(\boldsymbol{w}) \succ 0,$$

因此岭回归目标函数严格凸.

例 1.1.17 设训练数据为

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N, \qquad \boldsymbol{x}_i \in \mathbb{R}^d, \quad y_i \in \{0, 1\}.$$

逻辑回归模型定义为

$$p_i = \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}_i\right) = \frac{1}{1 + \exp\left(-\boldsymbol{w}^\top \boldsymbol{x}_i\right)}.$$

其平均交叉熵损失可以写为

$$L(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^N \left[\ln\left(1 + \exp\left(\boldsymbol{w}^\top \boldsymbol{x}_i\right)\right) - y_i \boldsymbol{w}^\top \boldsymbol{x}_i\right].$$

根据链式法则，函数梯度为

$$\nabla L(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^N \left(p_i - y_i\right) \boldsymbol{x}_i.$$

再次求导可得

$$H_L(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^N p_i(1 - p_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

令

$$D = \operatorname{Diag}\left(p_1(1 - p_1), \ldots, p_N(1 - p_N)\right),$$

并令 X 的第 i 行为 $\boldsymbol{x}_i^\top$，则海瑟矩阵可以写为

$$H_L(\boldsymbol{w}) = \frac{1}{N} X^\top D X.$$

由于 $p_i(1 - p_i) \geq 0$，因此

$$H_L(\boldsymbol{w}) \succeq 0,$$

从而逻辑回归的交叉熵损失函数是凸函数.

1.1.4* 可微性

我们下面简要介绍一下可微性（differentiability）的概念. 该概念为链式法则，以及复合函数微分求导法则做铺垫，不要求掌握. 我们给出二元函数的可微性的定义，多元函数可以以此类推.

Yuan and Song

定义 1.1.18 (可微性) 设函数 $z = f(x, y)$ 在点 (x_0, y_0) 的一个邻域中有定义，我们记 $\Delta z := f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$, 其中 $\Delta x, \Delta y$ 为两个自由变元. 如果 Δz 可表示为

$$\Delta z = A\Delta x + B\Delta y + o(\rho), \tag{1.1}$$

其中 A, B 是两个不依赖于 $\Delta x, \Delta y$ 的常数, $\rho := \sqrt{(\Delta x)^2 + (\Delta y)^2}$, 则我们称函数 $f(x, y)$ 在点 (x_0, y_0) 处可微. 增量形式(1.1)可以等价地写作如下微分形式：*[** 高阶无穷小 **]*

$$\mathrm{d}z = A\mathrm{d}x + B\mathrm{d}y$$

小贴士

使用微分符号 $\mathrm{d}(\cdot)$ 的形式为非严谨形式，通常只用于理论的快速推导，在物理等领域中应用频率更高. 如果需要严谨证明，请回退至增量形式(1.1).

可微性的定义相比偏导数的定义稍强一些. 事实上，如果函数 $z = f(x, y)$ 在 (x_0, y_0) 处可微，那么它一定在 (x_0, y_0) 处可导，且它的偏导数正是 A 与 B. 这是因为，当我们固定 $\Delta y = 0$,

$$f'_x(x_0, y_0) = \lim_{\Delta x \to 0} \frac{\Delta z}{\Delta x} = \lim_{\Delta x \to 0} \left(A + \frac{o(|\Delta x|)}{\Delta x} \right) = \lim_{\Delta x \to 0} \big(A + o(1) \big) = A.$$

同理得 $f'_y(x_0, y_0) = B$. 从而我们通常直接将函数 f 的微分形式写为：

$$\mathrm{d}z = \frac{\partial z}{\partial x}\mathrm{d}x + \frac{\partial z}{\partial y}\mathrm{d}y.$$

可微性的重要性在于，其揭示了两个自变量分量对于因变量的影响可以表达为简单的线性加和形式. 相比于偏导数一次只能聚焦于一个分量，可微性则一次性考虑了自变量的所有分量，因此可微性通常也可以叫做“全微分”（totoal differential），与偏导数形成对照. 一般地，假如自变量为 $\boldsymbol{x} \in \mathbb{R}^n$, 且函数 $z = f(\boldsymbol{x})$ 的梯度已知为 $\nabla f(\boldsymbol{x})$, 则微分形式可以紧凑地记为如下格式（注意此时梯度为行向量）：

$$\mathrm{d}z = \nabla f(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}.$$

小贴士

在一元函数中，由于微分的形式不变性，我们可以用 $\mathrm{d}z/\mathrm{d}x$ 来表示导数. 然而在多元函数中，由于我们没有向量相除的运算，$(\times)\, \mathrm{d}z/\mathrm{d}\boldsymbol{x}$ 的形式是没有意义的，不能用来表示梯度.

1.2 线性映射与矩阵相似

我们在本节考虑线性映射、线性变换以及矩阵的相似. 注意本节中考虑的域为任意一般的域，不只局限于实数域 $\mathbb{R}$.

1.2.1 线性映射

我们先考虑一般域 F（实数域、复数域或其他域）上的线性空间中定义的线性映射（linear mapping），随后说明线性映射在 n 维实数向量空间中必然具有 $\boldsymbol{y}=A\boldsymbol{x}$ 的形式.

定义 1.2.1 (线性映射) 设 U, V 为两个定义在域 F 上的线性空间，如果一个映射 $T: U \to V$ 满足线性条件：

$$T(c_1\boldsymbol{u}_1+\cdots+c_s\boldsymbol{u}_s)=c_1T(\boldsymbol{u}_1)+\cdots+c_sT(\boldsymbol{u}_s), \quad \forall c_1,\ldots,c_s\in F;\ \forall \boldsymbol{u}_1,\ldots,\boldsymbol{u}_s\in U,$$

则 T 称为一个线性映射.

我们回忆线性空间中的基的概念. n 维线性空间 U 中存在 n 个线性无关的向量 $\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_n$, 使得对于任意一个 $\boldsymbol{u}\in U$, 均可以表示为基向量线性组合的形式，即

$$\boldsymbol{u}=c_1\boldsymbol{e}_1+c_2\boldsymbol{e}_2+\cdots+c_n\boldsymbol{e}_n.$$

根据线性映射的定义，我们有

$$T(\boldsymbol{u})=c_1T(\boldsymbol{e}_1)+c_2T(\boldsymbol{e}_2)+\cdots+c_nT(\boldsymbol{e}_n).$$

因此，我们只需给出 U 中的 n 个基向量的线性映射的结果，就可以掌握该线性映射的全部信息. 我们设陪域空间 V 的维度为 m, 记 $\boldsymbol{v}_1=T(\boldsymbol{e}_1),\ldots,\boldsymbol{v}_n=T(\boldsymbol{e}_n)$ 为 U 中 n 个基向量在 V 中的线性映射的像，我们有如下的紧凑的记号：

$$T(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n):=(T(\boldsymbol{e}_1),\ldots,T(\boldsymbol{e}_n))=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n).$$

小贴士

$T(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)$ 仅为紧凑记号，并非表示对矩阵作线性变换，计算时需要将其展开.

下面我们考虑将 U,V 具体化为 n 维、m 维的实数向量空间. 如果我们取 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 为标准正交基，即 $\boldsymbol{e}_1=(1,0,\ldots,0)^\top,\ldots,\boldsymbol{e}_n=(0,0,\ldots,1)^\top\in\mathbb{R}^n$, 它们在线性映射 T 下的像为 $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\in\mathbb{R}^m$, 我们可以将其排列为一个矩阵 $A=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n)\in\mathbb{R}^{m\times n}$. 考

Yuan and Song

虑 $U=\mathbb{R}^n$ 中的任何一个向量 $\boldsymbol{x}=(x_1,\ldots,x_n)^\top$, 我们有

$$\begin{aligned}T(\boldsymbol{x})&=T(x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\cdots+x_n\boldsymbol{e}_n)\\&=x_1T(\boldsymbol{e}_1)+x_2T(\boldsymbol{e}_2)+\cdots+x_nT(\boldsymbol{e}_n)\\&=x_1\boldsymbol{v}_1+x_2\boldsymbol{v}_2+\cdots+x_n\boldsymbol{v}_n\\&=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n)\boldsymbol{x}=A\boldsymbol{x}.\end{aligned}$$

从而我们证明了，$\mathbb{R}^n$ 上的任何一个线性映射均可以写为 $\boldsymbol{y}=A\boldsymbol{x}$ 的形式，且 A 为 $\mathbb{R}^n$ 中标准正交基在该线性映射下的像构成的矩阵.

1.2.2 线性变换

线性映射考虑两个不同的线性空间，而如果我们考虑线性空间 U 到自身的映射，我们则称其为线性变换（linear transformation）.

我们同样使用基向量的视角来研究线性映射. 由于线性变换是线性空间 U 到自身的映射，我们可以使用同一套基向量来同时表示映射前与映射后的向量. 对于 U 中的基向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$, 线性变换 T 作用在每个基向量上得到的向量同样可以由基向量表示，即：

$$T(\boldsymbol{e}_j)=a_{1j}\boldsymbol{e}_1+\cdots+a_{nj}\boldsymbol{e}_n,\quad j=1,\ldots,n,$$

我们可以得到如下紧凑的格式：

$$T(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)\begin{pmatrix}a_{11}&\cdots&a_{1n}\\\vdots&&\vdots\\a_{n1}&\cdots&a_{nn}\end{pmatrix}:=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)A.$$

我们称上述的式子为线性变换 T 在基向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 下的矩阵表示.

如果 U 为 n 维实数向量空间 $\mathbb{R}^n$, 基向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 为标准正交基，那么对于任意 $\boldsymbol{x}\in\mathbb{R}^n$, 不难得到此时 $T(\boldsymbol{x})=A\boldsymbol{x}$, 这与我们上文得到的结论是一致的. 然而，如果基向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 不是 $\mathbb{R}^n$ 上的标准正交基，那么此时得到的新的矩阵 A' 与标准正交基下的矩阵 A 之间有什么关系呢？我们下面对其进行探究.

设 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 以及 $\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_n$ 为 $\mathbb{R}^n$ 上的两个基，由于它们二者之间可以互相表出，因此必存在一个可逆矩阵 $P=(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n)$ 使得

$$(\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_n)=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)P,\quad(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)=(\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_n)P^{-1}.$$

在已知 $T(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)A$ 的情况下，

$$T(\boldsymbol{e}'_j)=T((\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)\boldsymbol{p}_j)=\sum_{i=1}^{n}p_{ij}T(\boldsymbol{e}_i)$$

$$= \sum_{i=1}^{n} p_{ij}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) \sum_{i=1}^{n} p_{ij} \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix}$$

$$= (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) \begin{pmatrix} \sum_{i=1}^{n} a_{1i} p_{ij} \\ \vdots \\ \sum_{i=1}^{n} a_{ni} p_{ij} \end{pmatrix} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)(AP)_{\cdot,j},$$

其中 $(AP)_{\cdot,j}$ 表示矩阵 AP 的第 j 列构成的列向量. 从而我们有

$$T(\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_n) = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)AP = (\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_n)P^{-1}AP.$$

即我们有 A' 与 A 之间的关系为 $A' = P^{-1}AP$.

接下来我们给出一些线性映射与线性变换的实例.

例 1.2.2 (从二维空间到三维空间) 定义

$$T : \mathbb{R}^2 \to \mathbb{R}^3, \quad T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 \\ 2x_1 - x_2 \\ 3x_2 \end{pmatrix}.$$

这个映射可以写成

$$T(\boldsymbol{x}) = A\boldsymbol{x}, \qquad A = \begin{pmatrix} 1 & 2 \\ 2 & -1 \\ 0 & 3 \end{pmatrix}.$$

对于任意 $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^2$ 和 $a, b \in \mathbb{R}$，有

$$\begin{aligned} T(a\boldsymbol{u} + b\boldsymbol{v}) &= A(a\boldsymbol{u} + b\boldsymbol{v}) \\ &= aA\boldsymbol{u} + bA\boldsymbol{v} \\ &= aT(\boldsymbol{u}) + bT(\boldsymbol{v}). \end{aligned}$$

因此，T 是线性映射.

定义域的两个标准基向量的像为

$$T \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \qquad T \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}.$$

Yuan and Song

它们恰好是矩阵 A 的两列. 例如，

$$T\begin{pmatrix}3\\-2\end{pmatrix}=3\begin{pmatrix}1\\2\\0\end{pmatrix}-2\begin{pmatrix}2\\-1\\3\end{pmatrix}=\begin{pmatrix}-1\\8\\-6\end{pmatrix}.$$

这个例子说明：定义域和陪域可以具有不同的维数；矩阵的列数对应输入维数，行数对应输出维数.

例 1.2.3 (保持零向量并不足以保证线性) 考虑映射

$$S:\mathbb{R}^2\to\mathbb{R}^2,\qquad S\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}x^2\\y\end{pmatrix}.$$

它满足 $S(\mathbf{0})=\mathbf{0}$. 但是，取 $\boldsymbol{u}=(1,0)^\top$，有

$$S(2\boldsymbol{u})=S\begin{pmatrix}2\\0\end{pmatrix}=\begin{pmatrix}4\\0\end{pmatrix},$$

而

$$2S(\boldsymbol{u})=2\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}2\\0\end{pmatrix}.$$

因此 $S(2\boldsymbol{u})\neq 2S(\boldsymbol{u})$，所以 S 不是线性映射.

更值得注意的是，它在两个标准基向量上的取值与恒等映射相同：

$$S(\boldsymbol{e}_1)=\boldsymbol{e}_1,\qquad S(\boldsymbol{e}_2)=\boldsymbol{e}_2.$$

但它仍然不是恒等映射，也不是线性映射.

因此，“由基向量的像确定整个映射”以映射具有线性性为前提. 仅仅检查基向量上的取值，不能证明一个映射是线性的.

例 1.2.4 (多项式求导的矩阵表示) 设

$$\mathcal{P}_2=\{a+bt+ct^2:a,b,c\in\mathbb{R}\},\qquad \mathcal{P}_1=\{a+bt:a,b\in\mathbb{R}\}.$$

定义

$$D:\mathcal{P}_2\to\mathcal{P}_1,\qquad D(p)=p'.$$

对于任意 $p,q\in\mathcal{P}_2$ 和 $\alpha,\beta\in\mathbb{R}$，根据求导法则，

$$\begin{aligned}D(\alpha p+\beta q)&=(\alpha p+\beta q)'\\&=\alpha p'+\beta q'\\&=\alpha D(p)+\beta D(q).\end{aligned}$$

Yuan and Song

因此，求导是线性映射.

定义域选取有序基 $(1,t,t^2)$，陪域选取有序基 $(1,t)$. 由于

$$D(1)=0,\qquad D(t)=1,\qquad D(t^2)=2t,$$

对应的矩阵为

$$M_D=\begin{pmatrix}0&1&0\\0&0&2\end{pmatrix}.$$

例如，多项式 $p(t)=3-2t+5t^2$ 在定义域基下的坐标是 $(3,-2,5)^\top$，所以其导数的坐标为

$$M_D\begin{pmatrix}3\\-2\\5\end{pmatrix}=\begin{pmatrix}-2\\10\end{pmatrix}.$$

对应的多项式为 $D(p)=-2+10t$，与直接求导一致.

线性空间中的向量可以是多项式；矩阵表示作用于多项式在所选基下的坐标.

例 1.2.5 (平面上的旋转) 定义

$$T:\mathbb{R}^2\to\mathbb{R}^2,\qquad T\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}-y\\x\end{pmatrix}.$$

这是将平面中的向量绕原点逆时针旋转 90° 的变换.

两个标准基向量的像为

$$T(\boldsymbol{e}_1)=\boldsymbol{e}_2,\qquad T(\boldsymbol{e}_2)=-\boldsymbol{e}_1.$$

因此，其在标准基下的矩阵为

$$R=\begin{pmatrix}0&-1\\1&0\end{pmatrix}.$$

由于 $T(\boldsymbol{x})=R\boldsymbol{x}$，它是线性映射；又因为定义域和陪域都是 $\mathbb{R}^2$，所以按照本文约定，它是线性变换.

连续作用两次，得到

$$T(T(\boldsymbol{x}))=-\boldsymbol{x},\qquad R^2=-I_2.$$

连续作用四次，则回到原来的向量：

$$R^4=I_2.$$

Yuan and Song

其逆变换是顺时针旋转 $90°$，对应矩阵

$$R^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

这个例子也展示了：线性变换的复合对应矩阵的乘法.

例 1.2.6 (投影到横轴) 定义

$$T : \mathbb{R}^2 \to \mathbb{R}^2, \qquad T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix}.$$

这个变换将每个向量投影到横轴，其矩阵为

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

它保持横轴上的向量不变：

$$T\begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix},$$

并将纵轴上的所有向量映射为零：

$$T\begin{pmatrix} 0 \\ y \end{pmatrix} = \mathbf{0}.$$

已经投影到横轴上的向量，再投影一次仍然不变，因此

$$T(T(\boldsymbol{x})) = T(\boldsymbol{x}), \qquad A^2 = A.$$

另一方面，不同输入可能具有相同输出，例如

$$T\begin{pmatrix} 1 \\ 2 \end{pmatrix} = T\begin{pmatrix} 1 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

因此，该线性变换不可逆.

线性变换虽然从一个空间映射到自身，但不一定可逆，其矩阵也不一定是可逆矩阵.

例 1.2.7 (更换基后，矩阵表示发生变化) 考虑 $\mathbb{R}^2$ 上的线性变换

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x \\ y \end{pmatrix}.$$

它在标准基下的矩阵为

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Yuan-ma Song

现在选取新基

$$\boldsymbol{b}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{b}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

由于

$$T(\boldsymbol{b}_1) = 2\boldsymbol{b}_1, \qquad T(\boldsymbol{b}_2) = \boldsymbol{b}_1 + \boldsymbol{b}_2,$$

所以 T 在新基下的矩阵为

$$A' = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

新基向量在标准基下的坐标按列排列，得到

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad P^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

直接计算可验证

$$\begin{aligned} P^{-1}AP &= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} = A'. \end{aligned}$$

例如，$\boldsymbol{x} = (3,1)^{\top} = 2\boldsymbol{b}_1 + \boldsymbol{b}_2$ 在新基下的坐标为

$$\boldsymbol{x}' = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

输出的新坐标为

$$\boldsymbol{y}' = A'\boldsymbol{x}' = \begin{pmatrix} 5 \\ 1 \end{pmatrix}.$$

换回标准坐标，得到

$$\boldsymbol{y} = P\boldsymbol{y}' = \begin{pmatrix} 6 \\ 1 \end{pmatrix},$$

与直接计算 $A\boldsymbol{x}$ 的结果一致.

A 与 A' 描述的是同一个线性变换，只是输入、输出采用的坐标基不同.

小贴士

得到 $(AP)_{\cdot,j}$ 的过程并不直接，需要一些经验与洞察力. 有关矩阵及张量线性运算的系统理论可以参考爱因斯坦求和约定（Einstein summation convention）.

Yuan and Song

1.2.3 矩阵相似、特征值与特征向量

上一小节我们得到，一个线性空间中的一个线性变换在不同基向量下的矩阵表示具有形如 $A' = P^{-1}AP$ 的关系，事实上这是矩阵理论中十分重要的一个关系，我们称其为相似关系（similarity）：

定义 1.2.8 (相似矩阵) *给定方形矩阵 $A \in F^{n\times n}$, 对于另一相同形状的矩阵 $A' \in F^{n\times n}$, 若存在可逆矩阵 $P \in F^{n\times n}$ 使得*

$$A' = P^{-1}AP,$$

则称矩阵 A' 与矩阵 A 相似.

根据相似矩阵的定义，我们可以很容易证明其满足自反性（任意矩阵 A 与其自身相似）、对称性（若矩阵 A 与 B 相似，则矩阵 B 与 A 也相似）、传递性（若矩阵 A 与 B 相似，B 与 C 相似，则矩阵 A 与 C 相似）.

相似矩阵有以下三个重要性质：

1. 秩相等：即 $\text{rank}(A') = \text{rank}(P^{-1}AP) = \text{rank}(A)$, 因为可逆矩阵 P 相当于对原先矩阵做初等行列变换，不会改变原先的秩.
2. 迹相等：即 $\text{tr}(A') = \text{tr}(P^{-1}AP) = \text{tr}(APP^{-1}) = \text{tr}(A)$，其中 $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$ 为矩阵 A 的迹.
3. 行列式相等：即 $|A'| = |P^{-1}AP| = |P^{-1}||A||P| = |P|^{-1}|A||P| = |A|$.

在实际计算中，直接判定 A, B 两个矩阵是否相似并不容易. 一个较为可行的思路是，对于两个给定的矩阵 A, B, 我们是否可以找到某种意义下最为简单的矩阵 D, 如果矩阵 A, B 均与 D 相似，那么矩阵 A 与矩阵 B 也相似. 打一比方，在不借助地图或是导航等工具等情况下，判断两个人的家是否离得较近并不容易，然而，如果两个人的家附近都有同一个很明显的标志性建筑物，我们便可以断定两人的家离得也不会很远.

在矩阵的世界中，**对角矩阵**可以充当这些明显的标志性建筑物. 如果两个矩阵均相似于同一个对角矩阵，则这两个矩阵也相似. 而且对角矩阵的另一个重要性质是，两个对角矩阵相似当且仅当两个对角矩阵的对角元素在不考虑顺序的情况下完全相同（这一性质的证明需要用到特征多项式的代数性质，因此在这里省略），这一性质保证了对角矩阵可以作为某种程度上唯一的标志物，便可作为一种标准. 因此我们称对角矩阵为所有与其相似的矩阵的**标准形**.

Yuan and Song

小贴士

并非所有的矩阵都能相似于一个对角矩阵，譬如矩阵 $\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$, 无论如何右上角的非对角元均不能消去. 这些矩阵的标准形需要采用对角矩阵以外的其他形式. 而我们更多感兴趣的矩阵还是能够相似于对角矩阵的矩阵，我们称这些矩阵为“可对角化”的矩阵.

现假设矩阵 $A \in \mathbb{R}^{n\times n}$ 相似于一个对角矩阵 $D = \mathrm{Diag}\{\lambda_1, \ldots, \lambda_n\}$, 则存在一个可逆矩阵 $P = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$ 使得

$$A(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix} = (\lambda_1\boldsymbol{p}_1, \ldots, \lambda_n\boldsymbol{p}_n).$$

因此我们可以看到，如果 A 相似于对角矩阵，那么有 n 个实数与 n 个 n 维向量满足

$$A\boldsymbol{p}_i = \lambda_i\boldsymbol{p}_i, \quad i = 1, \ldots, n.$$

这表示，如果将 A 视为一个线性变换，那么 A 作用在向量 $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ 上，就相当于对它们分别进行 $\lambda_1, \ldots, \lambda_n$ 倍的缩放. $\lambda_i, \boldsymbol{p}_i, i = 1, \ldots, n$ 分别称为矩阵 A 的特征值（eigenvalue）和特征向量（eigenvector），其中前缀“eigen-”来自德语，表示“固有的、本征的”，读作/'aɪgən/. 特征值与特征向量的严格定义如下：

定义 1.2.9 (特征值与特征向量) *给定方形矩阵 $A \in F^{n\times n}$, 若存在 $\lambda \in F$, 及非零向量 $\boldsymbol{p} \in F^n$ 满足*

$$A\boldsymbol{p} = \lambda\boldsymbol{p},$$

则称 λ 为矩阵 A 的一个特征值，称 $\boldsymbol{p}$ 为特征值 λ 对应的一个特征向量.

对于实际问题中的一个方形矩阵 $A \in \mathbb{R}^{n\times n}$, 计算其特征值与特征向量有固定的步骤，具体分为如下两步：

1. **计算特征值**：如果 λ 为矩阵 A 的特征值, 则一定存在一个非零的特征向量 $\boldsymbol{p}$ 使得 $(\lambda I - A)\boldsymbol{p} = \boldsymbol{0}$, 根据齐次线性方程组的结论，有行列式 $|\lambda I - A| = 0$. 事实上，$|\lambda I - A|$ 构成了一个关于变量 λ 的 n 次多项式，我们称其为矩阵 A 的**特征多项式**（characteristic polynomial），求解该多项式的所有根便可求出矩阵 A 的所有特征值.
2. **计算特征向量**：对于已经求出的一个特征值 λ, 我们可以求解齐次线性方程组 $(\lambda I - A)\boldsymbol{p} = \boldsymbol{0}$ 来求出该特征值对应的所有特征向量.

Yuan and Song

小贴士

1. 不难证明，如果向量 $\boldsymbol{p}$ 为特征值 λ 的一个特征向量，则对于任意 $c \in F\backslash\{0\}$, $c\boldsymbol{p}$ 也为特征值 λ 的一个特征向量；如果一个特征值 λ 有多个线性无关的特征向量 $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r$, 则它们的任意非零线性组合 $c_1\boldsymbol{p}_1 + \cdots + c_r\boldsymbol{p}_r; c_1, \ldots, c_r \in F$ 也为特征值 λ 的一个特征向量. 因此同一个特征值的所有特征向量，连同零向量，构成了一个子空间，称为**特征子空间**（eigen-subspace）.
2. 互相相似的矩阵有相同的特征值，但通常有着不同的特征向量.
3. 特征多项式 $|\lambda I - A|$ 中的一个根 λ 可能是多重根，我们称 λ 在特征多项式中的重数为**代数重数**（algebraic multiplicity）.
4. 一个特征值 λ 对应的特征子空间的维度为特征值 λ 的**几何重数**（geometric multiplicity）.
5. 如果矩阵 A 可以对角化，那么它的每一个特征值的代数重数与几何重数必须相等.

我们最后用若干实例来结束本节的内容：

例 1.2.10 (交换对角元素的排列顺序) 考虑两个对角矩阵

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

取

$$P = (\boldsymbol{e}_3, \boldsymbol{e}_1, \boldsymbol{e}_2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

这是一个置换矩阵，其逆矩阵为

$$P^{-1} = P^{\top} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

计算得到

$$P^{-1}AP = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = B.$$

因此，A 与 B 相似.

从基的角度看，这次变换只是把基向量的顺序从 $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ 改为 $(\boldsymbol{e}_3, \boldsymbol{e}_1, \boldsymbol{e}_2)$. 两个矩阵都满足

$$\mathrm{rank}(A) = \mathrm{rank}(B) = 3, \qquad \mathrm{tr}(A) = \mathrm{tr}(B) = 6, \qquad |A| = |B| = 6.$$

这个例子具体说明了：对角矩阵的对角元素重新排列后，得到的矩阵与原矩阵相似.

例 1.2.11 (利用行列式不同排除相似) 考虑

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

两者的秩相同：

$$\mathrm{rank}(A) = \mathrm{rank}(B) = 2,$$

迹也相同：

$$\mathrm{tr}(A) = 1 + 3 = 4, \qquad \mathrm{tr}(B) = 2 + 2 = 4.$$

但是，

$$|A| = 3, \qquad |B| = 4.$$

如果存在可逆矩阵 P 使得 $B = P^{-1}AP$，那么必有

$$|B| = |P^{-1}||A||P| = |A|,$$

与上述计算矛盾. 因此，A 与 B 不相似. 秩、迹或行列式中，只要有一个不同，就足以判定两个矩阵不相似.

例 1.2.12 (利用相似关系计算矩阵的高次幂) 考虑矩阵

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

取可逆矩阵

$$P = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad P^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

直接计算得

$$\begin{aligned} P^{-1}AP &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} =: D. \end{aligned}$$

因此，A 与对角矩阵 D 相似，并且 $A = PDP^{-1}$.

Nan and Song

计算 A^2 时，中间的 $P^{-1}P$ 可以消去：

$$\begin{aligned} A^2 &= (PDP^{-1})(PDP^{-1}) \\ &= PD(P^{-1}P)DP^{-1} \\ &= PD^2P^{-1}. \end{aligned}$$

同理，对于任意正整数 n，都有

$$\boxed{A^n = PD^nP^{-1}}.$$

对角矩阵的幂很容易计算：

$$D^n = \begin{pmatrix} 2^n & 0 \\ 0 & 1 \end{pmatrix}.$$

所以

$$\begin{aligned} A^n &= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2^n & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \boxed{\begin{pmatrix} 2^n & 2^n - 1 \\ 0 & 1 \end{pmatrix}}. \end{aligned}$$

例如，

$$A^5 = \begin{pmatrix} 32 & 31 \\ 0 & 1 \end{pmatrix}.$$

这个例子说明：如果一个矩阵相似于对角矩阵，就可以将矩阵的高次幂计算，转化为对角元素的高次幂计算.

Yuan and Song

例 1.2.13 考虑 $\mathbb{R}^2$ 上的线性变换 T：

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

矩阵 A 就是线性变换 T 在标准正交基 $(1,0)^\top, (0,1)^\top$ 下的矩阵表示.

我们下面先求矩阵 A 的特征值，其特征多项式为：

$$|\lambda I - A| = \begin{vmatrix} \lambda - 3 & -1 \\ -1 & \lambda - 3 \end{vmatrix} = (\lambda - 3)^2 - 1 = (\lambda - 4)(\lambda - 2).$$

特征多项式的两个根分别为 4 和 2, 它们即为矩阵 A 的特征值.

下面我们分别求两个特征值对应的特征向量. 先求 $\lambda = 4$ 的特征向量，我们解如下

的齐次线性方程组：

$$(4I-A)\boldsymbol{p}=\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\begin{pmatrix}p_1\\ p_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix},$$

易解得其解集为 $\{c_1(1,1)^\top | c_1\in\mathbb{R}\}$. 除去零向量后就是特征值 4 对应的特征向量. 同理我们解得特征值 2 的特征向量为 $\{c_2(-1,1)^\top | c_2\in\mathbb{R}\backslash\{0\}\}$.

我们取两个易于计算的特征向量 $\boldsymbol{p}_1=(1,1)^\top, \boldsymbol{p}_2=(-1,1)^\top$ 来构造相似矩阵. 记 $P=(\boldsymbol{p}_1,\boldsymbol{p}_2)$, $D=\mathrm{Diag}\{4,2\}$, 我们有：

$$AP=\begin{pmatrix}3 & 1\\ 1 & 3\end{pmatrix}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}=PD=\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}4 & 0\\ 0 & 2\end{pmatrix},$$

即矩阵 A 与矩阵 D 相似.

我们下面考虑使用上述的特征向量构建另一个基：

$$(\boldsymbol{e}_1',\boldsymbol{e}_2')=(\boldsymbol{e}_1,\boldsymbol{e}_2)\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}=\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}=P,$$

新基与原先的基下坐标表示关系满足

$$\begin{pmatrix}x_1\\ x_2\end{pmatrix}=(\boldsymbol{e}_1,\boldsymbol{e}_2)\begin{pmatrix}x_1\\ x_2\end{pmatrix}=(\boldsymbol{e}_1',\boldsymbol{e}_2')\begin{pmatrix}x_1'\\ x_2'\end{pmatrix}=P\begin{pmatrix}x_1'\\ x_2'\end{pmatrix},$$

则原线性变换在新的基下的坐标表示为

$$\begin{pmatrix}y_1'\\ y_2'\end{pmatrix}=P^{-1}\begin{pmatrix}y_1\\ y_2\end{pmatrix}=P^{-1}A\begin{pmatrix}x_1\\ x_2\end{pmatrix}=P^{-1}AP\begin{pmatrix}x_1'\\ x_2'\end{pmatrix}=D\begin{pmatrix}x_1'\\ x_2'\end{pmatrix}.$$

因此，从几何角度来看，矩阵 A 对平面中向量的变换，就相当于分别沿着特征向量 $(1,1)^\top,(-1,1)^\top$ 方向放大特征值 $4,2$ 倍. 如果将坐标系变换为以 $(1,1)^\top,(-1,1)^\top$ 为两个坐标轴，则矩阵 A 就相当于沿着两个轴的方向做缩放变换. 图1.1为其示意图.



图 1.1: 矩阵 A 作为线性变换矩阵示意图

Yuan and Song

例 1.2.14 (重特征值、几何重数与不可对角化) 考虑

$$A=\begin{pmatrix}2&0&0\\0&2&0\\0&0&3\end{pmatrix},\qquad B=\begin{pmatrix}2&1&0\\0&2&0\\0&0&3\end{pmatrix}.$$

两者具有相同的特征多项式：

$$|\lambda I-A|=|\lambda I-B|=(\lambda-2)^2(\lambda-3).$$

因此，特征值 2 的代数重数都是 2，特征值 3 的代数重数都是 1.

对于矩阵 A，方程

$$(A-2I)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\mathbf{0}$$

只要求 $x_3=0$. 所以

$$E_2(A)=\left\{\begin{pmatrix}x_1\\x_2\\0\end{pmatrix}:x_1,x_2\in\mathbb{R}\right\}=\operatorname{span}(\boldsymbol{e}_1,\boldsymbol{e}_2).$$

特征值 2 的几何重数为 2. 矩阵 A 本身就是对角矩阵.

对于矩阵 B，有

$$(B-2I)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}x_2\\0\\x_3\end{pmatrix}.$$

因此，方程要求 $x_2=x_3=0$，从而

$$E_2(B)=\operatorname{span}(\boldsymbol{e}_1).$$

这时特征值 2 的几何重数只有 1，小于其代数重数 2.

特征值 3 对应的特征子空间则为

$$E_3(B)=\operatorname{span}(\boldsymbol{e}_3).$$

所以 B 最多只能找到两个线性无关的特征向量，无法组成 $\mathbb{R}^3$ 的一组基，因此 B 不可对角化. 有重特征值的矩阵可能可对角化，也可能不可对角化；关键在于是否拥有足够多的线性无关特征向量.

Yuan and Song

1.3 多元向量值函数及其性质

最初，我们考虑了（单值）多元函数，即自变量为向量，而因变量为标量的函数. 然而，我们上一节考虑的线性映射的因变量则是一个向量. 因此，本节我们将系统考虑多元向量值函数，即自变量为向量，因变量也为向量的函数，并探究其微分性质及其与多元函数的关系.

1.3.1 雅各比矩阵

多元向量值函数的定义如下：

定义 1.3.1 (多元向量值函数) 设 X 为实数集 R^n 中的一个子集，如果存在一个对应法则 $\boldsymbol{f}$，使得对于任意 $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in X$，都存在唯一 $\boldsymbol{y}=(y_1,y_2,\ldots,y_m)\in\mathbb{R}^m$ 与之对应，则称 $\boldsymbol{f}$ 为定义在 X 上（陪域为 $\mathbb{R}^m$）的 n 元 m 维向量值函数，记作：

$$\boldsymbol{f}:X\to\mathbb{R}^m;\quad \boldsymbol{x}\mapsto\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})=\begin{pmatrix} f_1(x_1,x_2,\ldots,x_n)\\ f_2(x_1,x_2,\ldots,x_n)\\ \vdots\\ f_m(x_1,x_2,\ldots,x_n)\end{pmatrix}.$$

由于向量值函数可以视为由 m 个普通多元函数拼合而成，因此每单个多元函数 $y_i=f_i(\boldsymbol{x})$ 均可分别求出所有偏导数和梯度. 事实上，我们可以将每个多元函数的梯度用行向量的形式写出，并纵向拼合成一个矩阵，该矩阵即为多元向量值函数 $\boldsymbol{f}$ 的雅各比矩阵（Jacobian matrix）.

定义 1.3.2 (雅各比矩阵) 对于定义在 $X\subseteq\mathbb{R}^n$ 上的多元向量值函数 $\boldsymbol{f}(\boldsymbol{x})\in\mathbb{R}^m$，若其所有分量的所有偏导数 $(f_i)'_j(\boldsymbol{x}),i=1,2,\ldots,m;j=1,2,\ldots,n$ 在定义域内均存在且为连续函数，则函数 $\boldsymbol{f}$ 的雅各比矩阵定义为由因变量的所有分量关于自变量的所有分量求偏导所构成的矩阵，即：

$$J_{\boldsymbol{f}}(\boldsymbol{x}):=\begin{pmatrix}(f_1)'_1(\boldsymbol{x}) & \cdots & (f_1)'_n(\boldsymbol{x})\\ \vdots & & \vdots\\ (f_m)'_1(\boldsymbol{x}) & \cdots & (f_m)'_n(\boldsymbol{x})\end{pmatrix}=\begin{pmatrix}\nabla f_1(\boldsymbol{x})^\top\\ \vdots\\ \nabla f_m(\boldsymbol{x})^\top\end{pmatrix}.$$

同样，雅各比矩阵也有不止一种表示方式，我们也可以直接借用一元函数中的导数记号，将其记为 $\boldsymbol{f}'(\boldsymbol{x})$. 此外，在机器学习相关研究中，当因变量固定为 $\boldsymbol{y}$ 时，有时会使

Yuan and Song

用偏导数记号 $\dfrac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ 来不严谨地代表雅各比矩阵，为了不易混淆，本书后续内容中如果使用这一偏导数形式的记号，都会进行说明.

以下几个例子演示了雅各比矩阵的求解过程：

例 1.3.3 我们考虑线性映射这一多元向量值函数. 设函数 $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$,

$$\boldsymbol{f}(\boldsymbol{x})=A\boldsymbol{x},\quad A\in\mathbb{R}^{m\times n},$$

记因变量为 $\boldsymbol{y}=A\boldsymbol{x}$, 则因变量的每个分量为:

$$y_i=a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n,$$

相应地每个分量的偏导数为

$$\frac{\partial y_i}{\partial x_j}=a_{ij},$$

从而该函数的雅各比矩阵即为

$$J_{\boldsymbol{f}}(\boldsymbol{x})=A.$$

例 1.3.4 (从二维输入到三维输出) 考虑向量值函数

$$\boldsymbol{f}:\mathbb{R}^2\to\mathbb{R}^3,\qquad \boldsymbol{f}(x_1,x_2)=\begin{pmatrix}x_1^2+x_2\\ x_1x_2\\ e^{x_2}\end{pmatrix}.$$

分别计算三个分量的偏导数：

$$\frac{\partial f_1}{\partial x_1}=2x_1,\qquad \frac{\partial f_1}{\partial x_2}=1,$$

$$\frac{\partial f_2}{\partial x_1}=x_2,\qquad \frac{\partial f_2}{\partial x_2}=x_1,$$

$$\frac{\partial f_3}{\partial x_1}=0,\qquad \frac{\partial f_3}{\partial x_2}=e^{x_2}.$$

将每个分量函数的偏导数排列成一行，得到

$$J_{\boldsymbol{f}}(x_1,x_2)=\begin{pmatrix}2x_1 & 1\\ x_2 & x_1\\ 0 & e^{x_2}\end{pmatrix}.$$

Yuan and Song

例如，在 $(x_1, x_2) = (1, 1)$ 处，

$$J_{\boldsymbol{f}}(1,1) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \\ 0 & e \end{pmatrix}.$$

这个函数有三个输出分量、两个输入分量，因此其雅各比矩阵是一个 3×2 矩阵. 雅各比矩阵不必是方阵，并且通常随求导点的位置变化.

例 1.3.5 Softmax 函数用于将一列实数映射为一个抽样时的概率分布，具体的函数形式为 $\boldsymbol{s} : \mathbb{R}^n \to \mathbb{R}^n$,

$$s_i(\boldsymbol{x}) = \frac{e^{x_i}}{e^{x_1} + \cdots + e^{x_n}}.$$

我们下面考虑每个 s_i 关于 x_j 的偏导数. 如果 $j = i$, 则

$$(s_i)'_i(\boldsymbol{x}) = \frac{e^{x_i}}{e^{x_1} + \cdots + e^{x_n}} - \frac{e^{x_i} e^{x_i}}{(e^{x_1} + \cdots + e^{x_n})^2};$$

如果 $j \neq i$, 则

$$(s_i)'_j(\boldsymbol{x}) = -\frac{e^{x_i} e^{x_j}}{(e^{x_1} + \cdots + e^{x_j})^2}.$$

为使得我们的记号更加简便，我们固定因变量 $\boldsymbol{y} = \boldsymbol{s}(\boldsymbol{x})$, 则 Softmax 函数的雅各比矩阵可表示为：

$$J_{\boldsymbol{s}}(\boldsymbol{x}) = \mathrm{Diag}(\boldsymbol{y}) - \boldsymbol{y}\boldsymbol{y}^\top.$$

我们可以观察到，Softmax 函数的雅各比矩阵可以使用因变量进行较为简单的表示，这在前反向传播中具有一定的优势.

1.3.2 梯度作为向量值函数、雅各比矩阵与海瑟矩阵的联系

对于一个多元函数 $f : \mathbb{R}^n \to \mathbb{R}$, 我们可以求得它的 $n \times n$ 维海瑟矩阵；而对于多元向量值函数 $g : \mathbb{R}^n \to \mathbb{R}^n$, 我们同样可以求得它的 $n \times n$ 维的雅各比矩阵. 我们自然要问，两个矩阵之间有着什么样的联系呢？

对于函数 f, 我们会去思考其中是否蕴含一个多元向量值函数，其定义域与陪域均为 $\mathbb{R}^n$, 事实上这样的向量值函数确实存在，而且就是 f 的梯度：我们知道，函数 f 的每个偏导数都为一个 n 元函数，而梯度的定义便是将这 n 个偏导数拼合成为一个向量. 因此可以很自然地将梯度 ∇f 视为一个 n 元的 n 维向量值函数，即：

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n; \quad \boldsymbol{x} \mapsto \boldsymbol{y} = \nabla f(\boldsymbol{x}) = \big(f'_1(\boldsymbol{x}), \ldots, f'_n(\boldsymbol{x})\big)^\top,$$

注意此时梯度应写为列向量. 进一步，如果我们对这一多元向量值函数求雅各比矩阵，

Yuan and Song

我们可以得到：

$$J_{\nabla f}(\boldsymbol{x}) = \begin{pmatrix} f''_{11}(\boldsymbol{x}) & \cdots & f''_{1n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ f''_{n1}(\boldsymbol{x}) & \cdots & f''_{nn}(\boldsymbol{x}) \end{pmatrix}.$$

我们惊喜地发现，这恰好是函数 f 的海瑟矩阵. 因此我们有如下重要结论：**多元函数 f 的海瑟矩阵，等价于其梯度作为多元向量值函数时的雅各比矩阵.**

例 1.3.6 (梯度的雅各比矩阵就是海瑟矩阵) 考虑标量函数

$$\phi : \mathbb{R}^2 \to \mathbb{R}, \qquad \phi(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^4.$$

它的列梯度为

$$\nabla\phi(x_1, x_2) = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + 4x_2^3 \end{pmatrix}.$$

将梯度视为向量值函数

$$\boldsymbol{q} = \nabla\phi : \mathbb{R}^2 \to \mathbb{R}^2.$$

对 $\boldsymbol{q}$ 的两个分量再次求偏导：

$$\begin{aligned} J_{\boldsymbol{q}}(x_1, x_2) &= \begin{pmatrix} \dfrac{\partial(2x_1 + x_2)}{\partial x_1} & \dfrac{\partial(2x_1 + x_2)}{\partial x_2} \\ \dfrac{\partial(x_1 + 4x_2^3)}{\partial x_1} & \dfrac{\partial(x_1 + 4x_2^3)}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 \\ 1 & 12x_2^2 \end{pmatrix}. \end{aligned}$$

这恰好是 ϕ 的海瑟矩阵：

$$H_\phi(x_1, x_2) = J_{\nabla\phi}(x_1, x_2) = \begin{pmatrix} 2 & 1 \\ 1 & 12x_2^2 \end{pmatrix}.$$

这里需要区分：

$$\phi'(\boldsymbol{x}) = \begin{pmatrix} 2x_1 + x_2 & x_1 + 4x_2^3 \end{pmatrix}$$

是标量函数的一阶导数，形状为 1×2；而 $J_{\nabla\phi}(\boldsymbol{x})$ 是梯度函数的一阶导数，形状为 2×2，其中包含的是 ϕ 的二阶偏导数. 对标量函数求一次导数得到行梯度；对列梯度作为向量值函数再求雅各比矩阵，得到海瑟矩阵.

我们下面用一个十分常用的实例来进一步展现上述结论的魅力：

例 1.3.7 考虑如下的 n 元函数：

$$f : \mathbb{R}^n \to \mathbb{R}; \quad f(\boldsymbol{x}) = \log(e^{x_1} + e^{x_2} + \cdots + e^{x_n}),$$

Yuan and Song

该函数通常称为“log-sum-exp 函数”，或“指数和对数函数”. 它的函数取值大约比 $\max\{x_1, x_2, \ldots, x_n\}$ 稍大一些，而且具有 max 函数不具有的可求导的性质，因此可以视为求最大值函数的平滑版本. 我们下面先求它的偏导数：

对任意 $i = 1, 2, \ldots, n$,

$$f_i'(\boldsymbol{x}) = \frac{e^{x_i}}{e^{x_1} + \cdots + e^{x_n}}.$$

我们惊奇地发现，这一函数的梯度 $\nabla f(\boldsymbol{x})$ 正是例1.3.5中的 Softmax 函数. 因此，我们无需进行下一步计算，便可直接断言函数 f 的海瑟矩阵即为 Softmax 函数的雅各比矩阵，即：

$$H_f(\boldsymbol{x}) = J_{\boldsymbol{s}}(\boldsymbol{x}) = \operatorname{Diag}(\nabla f(\boldsymbol{x})) - \nabla f(\boldsymbol{x}) \nabla f(\boldsymbol{x})^\top.$$

从这一例子中我们可以窥探出 Softmax 函数名称的由来，因为以 Softmax 为导数的原函数——log-sum-exp 函数的作用就是求一组数的平滑的、柔性的最大值. 但由于历史原因，其导数反而被称为 Softmax 函数. 事实上，在机器学习领域中，Softmax 函数确实相比 log-sum-exp 函数更加常用. 因为相比于最大值的具体取值，我们更关心取得最大值的那个分量是哪一个，即取出 Softmax 函数输出的概率分布中最大的那个分量，这一操作在机器学习模型决定最终输出时广为使用.

1.3.3 链式法则

我们在第1.1.4节中介绍了全微分的概念，其包含两个版本：其中一个为增量形式的可微性，它是一个比可导性更强一些的性质，主要用于理论分析；另一个为微分形式的表达式，主要用于不严谨的导数推导. 使用严谨的增量微分往往需要较为复杂的数学分析技巧来分析高阶余项的误差，而使用不严谨的微分形式通常能够大幅度简化推导过程，只要我们断言所研究的函数对象性质足够好，高阶余项的误差始终保持为小量即可. 下面我们介绍一个为链式法则推导作准备的引理，并提供使用微分形式的简洁的证明.

引理 1.3.8 (菱形求导法则) 设二元函数 $g(u, v) : \mathbb{R}^2 \to \mathbb{R}$, 而 $u(x), v(x)$ 均为关于 x 的一元函数，考虑复合一元函数 $f(x) := g(u(x), v(x))$, 其导数为：*[** 插图 **]*

$$f'(x) = g_1'\big(u(x), v(x)\big) u'(x) + g_2'\big(u(x), v(x)\big) v'(x).$$

证明. 我们记 $y = g(u(x), v(x))$, 则

$$\mathrm{d}y = \frac{\partial y}{\partial u}\mathrm{d}u + \frac{\partial y}{\partial v}\mathrm{d}v = \frac{\partial y}{\partial u}\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}x + \frac{\partial y}{\partial v}\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x = \left(\frac{\partial y}{\partial u}\frac{\mathrm{d}u}{\mathrm{d}x} + \frac{\partial y}{\partial v}\frac{\mathrm{d}v}{\mathrm{d}x}\right)\mathrm{d}x$$

根据微分形式，我们可以倒推出函数 f 的导数的具体表达形式. ■

例 1.3.9 (一个输入通过两条路径影响输出) 设

$$u(t)=t^2,\qquad v(t)=e^t,\qquad h(u,v)=u^2+uv.$$

定义复合函数

$$f(t)=h(u(t),v(t)).$$

外层函数的偏导数为

$$\frac{\partial h}{\partial u}=2u+v,\qquad \frac{\partial h}{\partial v}=u.$$

两个内层函数的导数为

$$u'(t)=2t,\qquad v'(t)=e^t.$$

因为 t 同时通过 u 和 v 影响输出，所以两条路径的贡献需要相加：

$$\begin{aligned}f'(t)&=\left.\frac{\partial h}{\partial u}\right|_{(u(t),v(t))}u'(t)+\left.\frac{\partial h}{\partial v}\right|_{(u(t),v(t))}v'(t)\\&=(2t^2+e^t)\,2t+t^2e^t\\&=4t^3+(2t+t^2)e^t.\end{aligned}$$

这也可以写成矩阵乘法：

$$f'(t)=\begin{pmatrix}2t^2+e^t & t^2\end{pmatrix}\begin{pmatrix}2t\\ e^t\end{pmatrix}.$$

其中，外层函数的导数是 1×2 行向量，内层向量值函数的雅各比矩阵是 2×1 列向量.

直接代入还可得到

$$f(t)=t^4+t^2e^t,$$

对其求导，与上面的结果一致. 菱形求导法则中的“各条路径求导后相加”，正好对应矩阵乘法中的求和.

上述引理可以自然推广至除 u,v 外更多中间变量的情形，同时结合多元函数求不同分量偏导数的并行性、向量值函数不同维度求导的并行性，我们可以得出如下的有关多元向量值函数的链式法则：

定理 1.3.10 (链式法则) 考虑两个多元向量值函数

$$\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^m,\boldsymbol{y}=\boldsymbol{g}(\boldsymbol{x});\quad \boldsymbol{h}:\mathbb{R}^m\to\mathbb{R}^l,\boldsymbol{z}=\boldsymbol{h}(\boldsymbol{y}),$$

设它们的雅各比矩阵分别为 $\boldsymbol{g}'(\boldsymbol{x}),\boldsymbol{h}'(\boldsymbol{y})$. 定义复合函数

$$\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^l,\quad \boldsymbol{z}=\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{h}\big(\boldsymbol{g}(\boldsymbol{x})\big),$$

Yuan and Song

则它的雅各比矩阵为：

$$\boldsymbol{f}'(\boldsymbol{x}) = \boldsymbol{h}'\big(\boldsymbol{g}(\boldsymbol{x})\big)\boldsymbol{g}'(\boldsymbol{x}).$$

证明. 对于向量值函数 $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x})$, 考虑 $\boldsymbol{z}$ 的第 i 维对 $\boldsymbol{x}$ 的第 j 个分量求偏导数，则此时我们有：

$$z_i = h_i\big(g_1(x_1,\dots,x_j,\dots,x_n),\quad \dots\quad ,g_m(x_1,\dots,x_j,\dots,x_n)\big),$$

由于除 x_j 外的其余自变量分量均固定，因此此时 $g_1,\dots,g_m$ 即为关于 x_j 的一元函数，我们可以使用上述的菱形求导法则（引理1.3.8），得到：

$$\frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{m}(h_i)'_k\big(g(\boldsymbol{x})\big)\cdot(g_k)'_j(\boldsymbol{x}).$$

将所有 $i = 1,\dots,l; j = 1,\dots,n$ 计算并拼合成 $l\times n$ 维矩阵即可得到最终结果. ■

例 1.3.11 (二维到三维，再从三维到二维) 定义

$$\boldsymbol{g}:\mathbb{R}^2\to\mathbb{R}^3,\qquad \boldsymbol{g}(x_1,x_2) = \begin{pmatrix} x_1+x_2\\ x_1x_2\\ x_1^2\end{pmatrix},$$

以及

$$\boldsymbol{h}:\mathbb{R}^3\to\mathbb{R}^2,\qquad \boldsymbol{h}(y_1,y_2,y_3) = \begin{pmatrix} y_1y_2\\ y_2+y_3^2\end{pmatrix}.$$

令 $\boldsymbol{f} = \boldsymbol{h}\circ\boldsymbol{g}$. 内层函数的雅各比矩阵为

$$J_{\boldsymbol{g}}(x_1,x_2) = \begin{pmatrix} 1 & 1\\ x_2 & x_1\\ 2x_1 & 0\end{pmatrix}\in\mathbb{R}^{3\times2}.$$

外层函数的雅各比矩阵为

$$J_{\boldsymbol{h}}(y_1,y_2,y_3) = \begin{pmatrix} y_2 & y_1 & 0\\ 0 & 1 & 2y_3\end{pmatrix}\in\mathbb{R}^{2\times3}.$$

应将外层雅各比矩阵在内层输出处求值：

$$J_{\boldsymbol{h}}(\boldsymbol{g}(\boldsymbol{x})) = \begin{pmatrix} x_1x_2 & x_1+x_2 & 0\\ 0 & 1 & 2x_1^2\end{pmatrix}.$$

Yuan and Song

根据链式法则，

$$J_{\boldsymbol{f}}(\boldsymbol{x}) = J_{\boldsymbol{h}}(\boldsymbol{g}(\boldsymbol{x}))J_{\boldsymbol{g}}(\boldsymbol{x}) = \begin{pmatrix} 2x_1x_2 + x_2^2 & x_1^2 + 2x_1x_2 \\ x_2 + 4x_1^3 & x_1 \end{pmatrix}.$$

例如，

$$J_{\boldsymbol{f}}(1,2) = \begin{pmatrix} 8 & 5 \\ 6 & 1 \end{pmatrix}.$$

直接展开复合函数，得到

$$\boldsymbol{f}(x_1,x_2) = \begin{pmatrix} x_1^2x_2 + x_1x_2^2 \\ x_1x_2 + x_1^4 \end{pmatrix},$$

逐分量求导也得到相同结果. 外层雅各比矩阵要在内层函数的输出处求值；外层矩阵在左、内层矩阵在右. 本例的维数关系是

$$(2\times 3)(3\times 2) = 2\times 2,$$

与复合函数从 $\mathbb{R}^2$ 到 $\mathbb{R}^2$ 的维数一致.

小贴士

上述链式法则针对向量值函数，然而也能退化至普通的多元函数情形，只需令因变量的维度为 1 即可. 需要注意的是，因变量维度为 1 的向量值函数的雅各比矩阵是其对应的多元函数的**行梯度**. 如果想要转化为列梯度，需要进行转置操作. 为了不易混淆，在涉及链式法则时，我们使用导数记号 $f'(\boldsymbol{x})$ 来表示行梯度，其与雅各比矩阵是相合的；而用算子记号 $\nabla f(\boldsymbol{x})$ 来表示列梯度.

在一元函数中，链式法则求导可以写为紧凑且优雅的形式

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x},$$

或是微分形式

$$\mathrm{d}z = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x.$$

在多元函数中，我们同样有相应的微分形式：

$$\mathrm{d}\boldsymbol{z} = \boldsymbol{h}'(\boldsymbol{y})\boldsymbol{g}'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

然而，请不要写成导数形式

$$(\times)\quad \frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}\boldsymbol{x}} = \boldsymbol{h}'(\boldsymbol{y})\boldsymbol{g}'(\boldsymbol{x}),$$

也不要将雅各比矩阵写成 $(\times)\,\dfrac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}}$. 这是因为，一元函数中微分具有形式不变性，可以直观理解为导数与微分直接相除的结果是一致的. 而在多元函数中，我们不存在向量间相除的定义. 上述的错误记号仅是形式表示，不具有实际意义.

我们下面用几个使用链式法则的实例来结束本节内容：

例 1.3.12 设矩阵 $A\in\mathbb{R}^{m\times n}$, 向量 $\boldsymbol{b}\in\mathbb{R}^m$, 多元函数 $f:\mathbb{R}^n\to\mathbb{R}$,

$$f(\boldsymbol{x})=\frac{1}{2}(A\boldsymbol{x}-\boldsymbol{b})^\top(A\boldsymbol{x}-\boldsymbol{b}),$$

试求其（列）梯度.

解：考虑外层函数 $h(\boldsymbol{y})=\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y}$ 与内层函数 $\boldsymbol{y}=\boldsymbol{g}(\boldsymbol{x})=A\boldsymbol{x}-\boldsymbol{b}$, 分别参考例1.1.4与例1.3.3, 我们求得 h 的行梯度为：$h'(\boldsymbol{y})=\boldsymbol{y}^\top$; $\boldsymbol{g}$ 的雅各比矩阵为 $\boldsymbol{g}'(\boldsymbol{x})=A$, 因此根据链式法则，我们可以求得函数 f 的行梯度为：

$$f'(\boldsymbol{x})=h'\big(\boldsymbol{g}(\boldsymbol{x})\big)\boldsymbol{g}'(\boldsymbol{x})=(A\boldsymbol{x}-\boldsymbol{b})^\top A.$$

转置为列梯度可得：

$$\nabla f(\boldsymbol{x})=A^\top(A\boldsymbol{x}-\boldsymbol{b}).$$

例 1.3.13 (非线性平方损失的列梯度) 定义残差向量

$$\boldsymbol{r}:\mathbb{R}^2\to\mathbb{R}^2,\qquad \boldsymbol{r}(\boldsymbol{x})=\begin{pmatrix}x_1^2+x_2-1\\ 2x_1-x_2^2\end{pmatrix}.$$

考虑平方损失

$$L(\boldsymbol{x})=\frac{1}{2}\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{r}(\boldsymbol{x}).$$

将它分为外层函数 $h(\boldsymbol{y})=\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y}$ 和内层函数 $\boldsymbol{y}=\boldsymbol{r}(\boldsymbol{x})$. 外层函数的行梯度为

$$h'(\boldsymbol{y})=\boldsymbol{y}^\top.$$

内层函数的雅各比矩阵为

$$J_{\boldsymbol{r}}(\boldsymbol{x})=\begin{pmatrix}2x_1 & 1\\ 2 & -2x_2\end{pmatrix}.$$

因此，链式法则首先给出行梯度：

$$L'(\boldsymbol{x})=\boldsymbol{r}(\boldsymbol{x})^\top J_{\boldsymbol{r}}(\boldsymbol{x}).$$

转置后得到列梯度：

$$\nabla L(\boldsymbol{x})=J_{\boldsymbol{r}}(\boldsymbol{x})^\top\boldsymbol{r}(\boldsymbol{x}).$$

Yuan and Song

记两个残差分量为 r_1, r_2，则

$$\nabla L(\boldsymbol{x}) = \begin{pmatrix} 2x_1 & 2 \\ 1 & -2x_2 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 2x_1 r_1 + 2r_2 \\ r_1 - 2x_2 r_2 \end{pmatrix}.$$

代入残差表达式：

$$\nabla L(\boldsymbol{x}) = \begin{pmatrix} 2x_1(x_1^2 + x_2 - 1) + 2(2x_1 - x_2^2) \\ (x_1^2 + x_2 - 1) - 2x_2(2x_1 - x_2^2) \end{pmatrix}.$$

例如，在 $\boldsymbol{x} = (1,1)^\top$ 处，

$$\boldsymbol{r}(\boldsymbol{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad J_{\boldsymbol{r}}(\boldsymbol{x}) = \begin{pmatrix} 2 & 1 \\ 2 & -2 \end{pmatrix},$$

所以

$$\nabla L(1,1) = \begin{pmatrix} 2 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ -1 \end{pmatrix}.$$

对标量复合函数求列梯度时，内层雅各比矩阵需要转置. 线性最小二乘公式

$$\nabla L(\boldsymbol{x}) = A^\top (A\boldsymbol{x} - \boldsymbol{b})$$

正是这里取 $\boldsymbol{r}(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$ 的特殊情形.

例 1.3.14 设向量 $\boldsymbol{a} \in \mathbb{R}^n$, 常数 $c \in \mathbb{R}$, 多元函数 $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\boldsymbol{x}) = \ln\left(1 + e^{-c\boldsymbol{a}^\top \boldsymbol{x}}\right),$$

试求其（列）梯度.

解：考虑外层函数 $h(y) = \ln(1 + e^{-y})$ 与内层函数 $y = g(\boldsymbol{x}) = c\boldsymbol{a}^\top \boldsymbol{x}$. 外层函数 $h(y)$ 为一个一元函数，易求得其导数为：

$$h'(y) = \frac{-e^{-y}}{1 + e^{-y}} = \frac{-1}{1 + e^{y}}.$$

内层函数 $g(\boldsymbol{x})$ 为多元函数，通过分别计算偏导数易得到其行梯度为

$$g'(\boldsymbol{x}) = c\boldsymbol{a}^\top.$$

因此根据链式法则，我们可以得到函数 f 的行梯度为：

$$f'(\boldsymbol{x}) = h'\big(g(\boldsymbol{x})\big) g'(\boldsymbol{x}) = \frac{-c\boldsymbol{a}^\top}{1 + e^{y}},$$

转置为列梯度，并代入 y 可得：

$$\nabla f(\boldsymbol{x}) = \frac{-c}{1 + e^{c\boldsymbol{a}^\top \boldsymbol{x}}} \boldsymbol{a}.$$

Yuan and Song

1.4 实线性空间的内积与正定性

我们本节中仅考虑实数域的向量空间，即 $\mathbb{R}^n$, 这是因为实数域具有我们本节中将要介绍的正定性这一重要性质. 本节中我们考虑这样一个问题：先前有关线性空间的讨论大多围绕单个向量，而我们想要进一步探究任意两个向量之间的关系. 将上述想法抽象后，我们的目标是定义一个函数 $p(\boldsymbol{x},\boldsymbol{y})$, 它接受两个 $\mathbb{R}^n$ 空间中的向量 $\boldsymbol{x},\boldsymbol{y}$ 作为自变量，并输出一个 $\mathbb{R}$ 中的标量，而进一步，为了使上述函数的应用范围尽可能广泛，我们希望该函数具有下述的三个性质：

1. **对称性**：对于任意 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, 有 $p(\boldsymbol{x},\boldsymbol{y})=p(\boldsymbol{y},\boldsymbol{x})$.
2. **线性性**：对于任意 $\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2\in\mathbb{R}^n$, 任意 $c_1,c_2\in\mathbb{R}$, 有
$$p(\boldsymbol{x},c_1\boldsymbol{y}_1+c_2\boldsymbol{y}_2)=c_1p(\boldsymbol{x},\boldsymbol{y}_1)+c_2p(\boldsymbol{x},\boldsymbol{y}_2).$$
3. **正定性**：对于任意 $\boldsymbol{x}\in\mathbb{R}^n$, 有 $p(\boldsymbol{x},\boldsymbol{x})\geq 0$, 且 $p(\boldsymbol{x},\boldsymbol{x})=0\Leftrightarrow\boldsymbol{x}=\boldsymbol{0}$.

第一条与第二条性质结合可以保证该函数对于左右两个变量以及任意多个线性组合均保持线性，我们一般称这样的函数为**双线性函数**. 第三条性质即正定性是实数域 $\mathbb{R}$ 中的特有性质，它可以确保当线性空间退化为一维时，上述函数能够自然退化为取值恒为非负数的一元函数 $f(x)=x^2$.

本节的后续小节将逐步揭开这一双线性函数的神秘面纱.

1.4.1 矩阵的合同及谱定理

得益于双线性函数的性质，对于任意给定 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, 我们可以使用线性空间中的一个基 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 来对它们进行表示，并代入双线性函数 $p(\cdot,\cdot)$ 进行计算. 设 $\boldsymbol{x},\boldsymbol{y}$ 在上述基下的坐标表示分别为 $\boldsymbol{u},\boldsymbol{v}$, 即我们有 $\boldsymbol{x}=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)\boldsymbol{u}$, $\boldsymbol{y}=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)\boldsymbol{v}$, 我们有：

$$p(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\sum_{j=1}^{n}u_iv_jp(\boldsymbol{e}_i,\boldsymbol{e}_j)=\boldsymbol{u}^{\top}\begin{pmatrix}p(\boldsymbol{e}_1,\boldsymbol{e}_1) & \cdots & p(\boldsymbol{e}_1,\boldsymbol{e}_n)\\ \vdots & & \vdots\\ p(\boldsymbol{e}_n,\boldsymbol{e}_1) & \cdots & p(\boldsymbol{e}_n,\boldsymbol{e}_n)\end{pmatrix}\boldsymbol{v}:=\boldsymbol{u}^{\top}M\boldsymbol{v}.$$

因此，和线性映射类似，我们只需将矩阵 M 的性质研究清楚，便可得知有关双线性函数 $p(\cdot,\cdot)$ 的全部信息. 我们称该矩阵为双线性函数 $p(\cdot,\cdot)$ 在基 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 下的度量矩阵，我们注意到该矩阵是一个实对称矩阵. 根据上述的结论，如果函数 $p(\cdot,\cdot)$ 满足对称性和线性性，那么选定一个基之后，它一定可以表示为 $\boldsymbol{u}^{\top}M\boldsymbol{v}$ 的形式，我们通常称 $\boldsymbol{u}^{\top}M\boldsymbol{v}$ 这一形式为一个**二次型**（quadratic form）. 然而，这一形式尚且不能保证第三条性质，即

正定性，譬如我们取 M 为零矩阵，则很容易找到非零向量 $\boldsymbol{x}$ 使得 $p(\boldsymbol{x},\boldsymbol{x})=0$.

我们下面依然借鉴线性映射的研究思路，考虑当基进行变换后，相应的 M 矩阵之间有怎样的关系. 对于同一个双线性函数 $p(\cdot,\cdot)$, 假如我们已有一个基 $\boldsymbol{e}_1,\dots,\boldsymbol{e}_n$ 及对应的度量矩阵 M, 下面我们考虑另一个基 $(\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_n)$, 它与原先基的关系为

$$(\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_n)=(\boldsymbol{e}_1,\dots,\boldsymbol{e}_n)P,$$

其中 P 为可逆矩阵. 设 $\boldsymbol{x},\boldsymbol{y}$ 在新的基下的表示为 $\boldsymbol{u}',\boldsymbol{v}'$, 则

$$\boldsymbol{x}=(\boldsymbol{e}_1,\dots,\boldsymbol{e}_n)\boldsymbol{u}=(\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_n)\boldsymbol{u}'=(\boldsymbol{e},\dots,\boldsymbol{e}_n)P\boldsymbol{u}',$$

从而 $\boldsymbol{u}=P\boldsymbol{u}'$, 同理 $\boldsymbol{v}=P\boldsymbol{v}'$. 从而

$$p(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{u}^{\top}M\boldsymbol{v}=(\boldsymbol{u}')^{\top}P^{\top}MP\boldsymbol{v}',$$

由此我们可以得出，$P^{\top}MP$ 即为新的基 $\boldsymbol{e}_1,\dots,\boldsymbol{e}_n$ 下的度量矩阵.

事实上，限定在实对称矩阵范围内，形式 $P^{\top}MP$ 代表了矩阵之间的另一种重要关系——合同关系（congruence），其定义如下：

定义 1.4.1 (合同矩阵) 给定实对称矩阵 $M\in\mathbb{R}^{n\times n}$, 对于另一形状相同的实对称矩阵 $M'\in\mathbb{R}^{n\times n}$, 若存在可逆矩阵 $P\in\mathbb{R}^{n\times n}$ 使得

$$M'=P^{\top}MP,$$

则称矩阵 M' 与矩阵 M 合同.

同样根据矩阵合同的定义，我们不难验证其满足自反性、对称性与传递性.

例 1.4.2 (合同换基保持双线性函数的取值) 在 $\mathbb{R}^2$ 的标准基下，考虑双线性函数

$$p(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^{\top}\boldsymbol{y},\qquad M=I_2.$$

将基换为 $\boldsymbol{e}'_1=2\boldsymbol{e}_1$, $\boldsymbol{e}'_2=3\boldsymbol{e}_2$，过渡矩阵及新的度量矩阵分别为

$$P=\begin{pmatrix}2&0\\0&3\end{pmatrix},\qquad M'=P^{\top}MP=\begin{pmatrix}4&0\\0&9\end{pmatrix}.$$

取 $\boldsymbol{x}=(2,3)^{\top}$, $\boldsymbol{y}=(4,-3)^{\top}$，它们在新基下的坐标为

$$\boldsymbol{u}'=P^{-1}\boldsymbol{x}=(1,1)^{\top},\qquad \boldsymbol{v}'=P^{-1}\boldsymbol{y}=(2,-1)^{\top}.$$

在原坐标下，$p(\boldsymbol{x},\boldsymbol{y})=2\cdot4+3\cdot(-3)=-1$；在新坐标下，

$$(\boldsymbol{u}')^{\top}M'\boldsymbol{v}'=4\cdot1\cdot2+9\cdot1\cdot(-1)=-1.$$

因此，换基时坐标和度量矩阵必须一起改变，双线性函数的取值才能保持不变. 此外，

Tian and Song

M 的特征值为 $1,1$，而 M' 的特征值为 $4,9$，所以二者合同但不相似. 一般的合同变换并不保持特征值.

对于矩阵合同之间的研究，我们没有像相似关系中的特征值、特征向量等工具. 但是，我们可以通过某一类特殊的矩阵——正交矩阵，将合同关系转化为我们更为熟悉的相似关系. 由于正交矩阵 Q 满足 $Q^{-1}=Q^{\top}$, 我们可以将合同关系 $Q^{\top}MQ$ 转化为相似关系 $Q^{-1}MQ$, 从而使用相似矩阵的工具研究合同的相关性质.

我们知道，对于一般的方形矩阵，其不一定能够相似于对角矩阵. 然而，对于实对称矩阵，我们可以证明其一定能够相似于对角矩阵，且相似关系中使用的可逆矩阵一定可以化为正交矩阵，我们通常将上述性质简称为**实对称矩阵一定可以正交相似于对角矩阵**. 这是一个非常强的性质，在线性代数中我们将其总结为一个重要的定理，即谱定理（spectral theorem），或称主轴定理（principal axis theorem），意为该定理将实对称矩阵相似为对角矩阵上的一串光谱值（特征值），或者将实对称矩阵理解为在某个主坐标轴（特征向量构成的坐标轴）上的缩放变换. 我们下面正式叙述谱定理的定义：

定理 1.4.3 (谱定理) 对于一个 $n\times n$ 的实对称矩阵 M，一定存在一个正交矩阵 Q 与一个实对角矩阵 Λ, 使得

$$M=Q\Lambda Q^{\top}=(\boldsymbol{q}_1,\dots,\boldsymbol{q}_n)\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}\begin{pmatrix}\boldsymbol{q}_1^{\top}\\ \vdots \\ \boldsymbol{q}_n^{\top}\end{pmatrix}=\sum_{i=1}^{n}\lambda_i\boldsymbol{q}_i\boldsymbol{q}_i^{\top},$$

其中 $\lambda_1,\dots,\lambda_n$ 为矩阵 M 的 n 个特征值（可能有重复），$\boldsymbol{q}_1,\dots,\boldsymbol{q}_n$ 分别为特征值 $\lambda_1,\dots,\lambda_n$ 对应的特征向量，且它们两两线性无关.

小贴士

实对称矩阵可以正交相似于对角矩阵，但并不意味着其只能正交相似于对角矩阵. 从特征向量的角度理解，我们可以对特征向量数乘，得到的依然为特征向量；如果某个特征值有多个线性无关的特征向量，则对它们线性组合得到的依然是特征向量. 我们用新的特征向量完全可以构建出非正交的用于相似关系的矩阵.

例 1.4.4 (用正交换元消去二次型的交叉项) 考虑实对称矩阵及相应的二次型

$$M=\begin{pmatrix}5 & 2\\ 2 & 2\end{pmatrix},\qquad q(\boldsymbol{x})=\boldsymbol{x}^{\top}M\boldsymbol{x}=5x_1^2+4x_1x_2+2x_2^2.$$

其特征多项式为

$$|\lambda I - M| = (\lambda - 5)(\lambda - 2) - 4 = (\lambda - 6)(\lambda - 1).$$

对应于 $6, 1$ 的单位特征向量可分别取为

$$\boldsymbol{q}_1 = \frac{1}{\sqrt{5}}(2, 1)^\top, \qquad \boldsymbol{q}_2 = \frac{1}{\sqrt{5}}(-1, 2)^\top.$$

它们满足 $\boldsymbol{q}_i^\top \boldsymbol{q}_i = 1$ 及 $\boldsymbol{q}_1^\top \boldsymbol{q}_2 = 0$. 因此

$$Q = (\boldsymbol{q}_1, \boldsymbol{q}_2) = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}, \qquad Q^\top M Q = \begin{pmatrix} 6 & 0 \\ 0 & 1 \end{pmatrix}.$$

令 $\boldsymbol{x} = Q\boldsymbol{z}$，则

$$q(\boldsymbol{x}) = \boldsymbol{z}^\top Q^\top M Q \boldsymbol{z} = 6z_1^2 + z_2^2.$$

例如，$\boldsymbol{x} = (1, 0)^\top$ 对应 $\boldsymbol{z} = (2, -1)^\top/\sqrt{5}$，代入后得到 $6 \cdot 4/5 + 1/5 = 5$，与原式一致. 谱定理在这里给出了一个具体的正交换元，使二次型的交叉项消失.

1.4.2* 谱定理的证明

在本小节中，我们将详细证明谱定理（定理1.4.3），谱定理的证明过程不要求掌握，但读者需要学会使用谱定理.

证明. **第一步：证明实对称矩阵的特征值全为实数.** 我们引入如下的两个操作：

- 矩阵或向量共轭：$\bar{A}$ 或 $\bar{\boldsymbol{v}}$, 即对矩阵或向量中每个元素取共轭复数.
- 矩阵的共轭转置：A^H, 即对矩阵 A 先求共轭再求转置（或先转置再共轭，结果相同）.

设 λ 为矩阵 M 的一个特征值，$\boldsymbol{x}$ 是它的一个特征向量，则 $M\boldsymbol{x} = \lambda\boldsymbol{x}$. 对等式两边取共轭转置，则

$$\boldsymbol{x}^H M^H = (M\boldsymbol{x})^H = (\lambda\boldsymbol{x})^H = \bar{\lambda}\boldsymbol{x}^H.$$

因为 M 为实对称矩阵，所以 $M^H = \bar{M}^\top = \bar{M} = M$, 从而有 $\boldsymbol{x}^H M = \bar{\lambda}\boldsymbol{x}^H$.

现在我们对 $M\boldsymbol{x} = \lambda\boldsymbol{x}$ 左乘 $\boldsymbol{x}^H$, 对 $\boldsymbol{x}^H M = \bar{\lambda}\boldsymbol{x}^H$ 右乘 $\boldsymbol{x}$ 得到：

$$\boldsymbol{x}^H M\boldsymbol{x} = \lambda\boldsymbol{x}^H\boldsymbol{x}; \quad \boldsymbol{x}^H M\boldsymbol{x} = \bar{\lambda}\boldsymbol{x}^H\boldsymbol{x}.$$

由于特征向量为非零向量，$\boldsymbol{x}^H\boldsymbol{x} = \sum_{i=1}^n \bar{x}_i x_i > 0$（一个复数与其共轭的乘积为其模长，具有非负性）. 从而我们可以得出 $\lambda = \bar{\lambda}$, 即 λ 为实数，从而实对称矩阵 M 的任何特征值均为实数.

Yuan and Song

第二步：证明实对称矩阵分属不同特征值的特征向量必正交. 所谓两向量 $\boldsymbol{x},\boldsymbol{y}$ 的正交性即为 $\boldsymbol{x}^\top\boldsymbol{y}=0$.（我们在此处仅使用形式定义，正交性含义将会在本节最后提及）

设 $M\boldsymbol{x}=\lambda\boldsymbol{x}, M\boldsymbol{y}=\mu\boldsymbol{y}$, 且 $\lambda\neq\mu$. 对于 $M\boldsymbol{x}=\lambda\boldsymbol{x}$, 取转置得 $\boldsymbol{x}^\top M=\lambda\boldsymbol{x}^\top$, 右乘 $\boldsymbol{y}$ 得到

$$\boldsymbol{x}^\top M\boldsymbol{y}=\lambda\boldsymbol{x}^\top\boldsymbol{y},$$

类似地，我们可以得到 $\boldsymbol{x}^\top M\boldsymbol{y}=\mu\boldsymbol{x}^\top\boldsymbol{y}$, 从而可以得到

$$(\mu-\lambda)\boldsymbol{x}^\top\boldsymbol{y}=0.$$

由于 $\lambda\neq\mu$, 则必有 $\boldsymbol{x}^\top\boldsymbol{y}=0$.

第三步：使用数学归纳法，考虑归纳假设. 对于一维的情形，$M=(m)$ 已是对角矩阵，且取 $Q=(1)$ 即位正交矩阵，满足定理结论. 现假设定理对 $n-1$ 维成立，考虑 n 维情形下的矩阵 M 是否同样满足定理.

第四步：归纳推理. 根据第一步的结论，矩阵 M 一定有一个实特征值 λ_1, 取它的一个单位向量 $\boldsymbol{v}_1$, 即 $\boldsymbol{v}_1^\top\boldsymbol{v}_1=1$, 以 $\boldsymbol{v}_1$ 为基础，我们可以扩充出一个 $\mathbb{R}^n$ 中的单位正交基 $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n\}$（可以任意选取一个基并进行施密特正交化）. 则我们有 $V_1:=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n)$ 为一个正交矩阵.

现考虑 $\boldsymbol{v}_1$ 与其他向量关于 M 的双线性函数 $\boldsymbol{v}_1^\top M\boldsymbol{v}_i$. 当 $i=1$ 时，$\boldsymbol{v}_1^\top M\boldsymbol{v}_1=\lambda_1\boldsymbol{v}_1^\top\boldsymbol{v}_1=\lambda_1$; 当 $i\neq1$ 时，由于 V_1 中向量两两正交，$\boldsymbol{v}_1^\top M\boldsymbol{v}_i=\lambda\boldsymbol{v}_1\boldsymbol{v}_i=0$. 从而

$$V_1^\top MV_1=\begin{pmatrix}\boldsymbol{v}_1^\top\\ \vdots\\ \boldsymbol{v}_n^\top\end{pmatrix}M\begin{pmatrix}\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\end{pmatrix}=\begin{pmatrix}\boldsymbol{v}_1^\top M\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_1^\top M\boldsymbol{v}_n\\ \vdots & & \vdots\\ \boldsymbol{v}_n^\top M\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n^\top M\boldsymbol{v}_n\end{pmatrix}=\begin{pmatrix}\lambda_1 & \mathbf{0}^\top\\ \mathbf{0} & \tilde{M}\end{pmatrix},$$

其中 $\mathbf{0}$ 为 $n-1$ 维零向量，$\tilde{M}$ 为 $(n-1)\times(n-1)$ 维矩阵. 由于双线性函数的对称性，我们不难得出 $V_1^\top MV_1$ 为实对称矩阵，从而 $\tilde{M}$ 也为实对称矩阵.

第五步：使用归纳假设. 根据归纳假设，对于矩阵 $\tilde{M}$, 存在正交矩阵 P_{n-1} 和对角矩阵 D_{n-1} 使得

$$\tilde{M}=P_{n-1}D_{n-1}P_{n-1}^\top.$$

Yuan and Cong

我们另取 $V_2 = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & P_{n-1} \end{pmatrix}$, 易验证 V_2 也为正交矩阵. 从而

$$\begin{aligned} M &= V_1 \begin{pmatrix} \lambda_1 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{M} \end{pmatrix} V_1^\top = V_1 \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & P_{n-1} \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{0}^\top \\ \mathbf{0} & D_{n-1} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} P_{n-1}^\top & \end{pmatrix} V_1^\top \\ &= (V_1 V_2) \begin{pmatrix} \lambda_1 & \mathbf{0}^\top \\ \mathbf{0} & D_{n-1} \end{pmatrix} (V_1 V_2)^\top. \end{aligned}$$

从而有正交矩阵 $Q = V_1 V_2$ 及实对角矩阵 $\Lambda = \begin{pmatrix} \lambda_1 & \mathbf{0}^\top \\ \mathbf{0} & D_{n-1} \end{pmatrix}$ 满足 $M = Q\Lambda Q^\top$. 归纳步成立. 从而定理对于任意 n 阶实对称矩阵均成立. ■

1.4.3 正定矩阵

以谱定理为基础，我们可以将任意一个实对称矩阵正交相似（即正交合同）为对角矩阵. 然而，在合同的概念下，对角矩阵依然包含一部分的冗余信息. 假如我们现在有一个对角矩阵 $\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_n)$, 我们可以另取一个可逆的对角矩阵 $P = \text{Diag}(p_1, \ldots, p_n), p_i \neq 0, \forall i = 1, \ldots, n$, 得到一个新的合同后的矩阵：

$$P^\top \Lambda P = \text{Diag}(p_1^2 \lambda_1, \ldots, p_n^2 \lambda_n).$$

得益于 $p_1^2 > 0$ 取值的任意性, 它可以将任意正数变为 $+1$, 任意负数变为 -1, 0 保持 0. 通过选取合适的矩阵 P, 我们可以将对角矩阵进一步合同变换为对角元只为

$$\{+1, -1, 0\}$$

其中之一的对角矩阵. 我们称只包含以上三种元素的对角矩阵为所有与之合同的实对称矩阵的**规范形**.

规范形中 $+1, -1, 0$ 元素的个数可以反映出实对称矩阵在合同意义下的所有信息. 进一步，如果对一个实对称矩阵 M 取负得到 $-M$，则其规范形中的 $+1, -1$ 元素也将互换. 因此综合来看，我们感兴趣的实对称矩阵可以分为如下三类：

- **正定矩阵**（positive definite matrix）：规范形中所有元素均为 $+1$.
- **半正定矩阵**（positive semi-definite matrix）：规范形中元素为 $+1$ 或 0, 不存在 -1.
- **不定矩阵**（indefinite matrix）：规范形中元素同时存在 $+1$ 和 -1.

在上述三种类型的实对称矩阵中，我们最感兴趣的是第一种，即**正定矩阵**. 正定矩阵具有许多优良的性质，我们在此简要列举正定矩阵 $M \in \mathbb{R}^{n \times n}$ 的若干基本的及与机器学

Yuan and Song

习密切相关的性质：

1. 所有特征值严格为正，即 $\lambda_1, \ldots, \lambda_n > 0$. 这可由上述关于规范形的讨论直接得出.
2. 存在可逆矩阵 $C \in \mathbb{R}^{n\times n}$ 使得
$$M = CC^\top.$$
这是因为 M 的合同规范形为对角元全为 1 的对角矩阵，即单位阵 I, 从而我们有可逆矩阵 C 及合同变换 $I = C^\top MC$.
3. 行列式 $|M| > 0$. 这是因为其可以正交对角化为对角元全为正数的对角矩阵，而相似矩阵具有相同的行列式值.
4. 正定性：对于任意实数向量 $\boldsymbol{x} \in \mathbb{R}^n$, 有：
$$\boldsymbol{x}^\top M\boldsymbol{x} \geq 0,$$
且不等式取等号当且仅当 $\boldsymbol{x} = \boldsymbol{0}$.
5. 可开方性：由于 $M = Q\Lambda Q^\top$, 且 $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ 中所有元素均为正，因此我们可以定义 $M^{1/2} := Q\Lambda^{1/2}Q^\top$, 其中 $\Lambda := \mathrm{Diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$, 我们不难验证，$M^{1/2}M^{1/2} = M$.

我们简要证明一下第四条性质，即正定性. 这也是正定矩阵最重要的一条性质.

证明. 由第二条性质可得，存在可逆矩阵 C 使得 $M = CC^\top$, 从而
$$\boldsymbol{x}^\top M\boldsymbol{x} = \boldsymbol{x}^\top CC^\top\boldsymbol{x} = (C^\top\boldsymbol{x})^\top(C^\top\boldsymbol{x}),$$
记 $\boldsymbol{y} = C^\top\boldsymbol{x}$, 则
$$\boldsymbol{y}^\top\boldsymbol{y} = \sum_{i=1}^{n} y_i^2 \geq 0.$$
且等号取等时当且仅当 $\boldsymbol{y} = \boldsymbol{0}$. 由于 C 为可逆矩阵，$\boldsymbol{x} = (C^\top)^{-1}\boldsymbol{y} = \boldsymbol{0}$. ■

例 1.4.5 (含参数矩阵的正定性与边界情形) 考虑实对称矩阵
$$M_t = \begin{pmatrix} 2 & t \\ t & 3 \end{pmatrix}, \qquad t \in \mathbb{R}.$$
对任意 $\boldsymbol{x} = (x_1, x_2)^\top$，配方得到
$$\boldsymbol{x}^\top M_t\boldsymbol{x} = 2x_1^2 + 2tx_1x_2 + 3x_2^2 = 2\left(x_1 + \frac{t}{2}x_2\right)^2 + \left(3 - \frac{t^2}{2}\right)x_2^2.$$
当 $|t| < \sqrt{6}$ 时，两个平方项的系数均为正，且两项同时为零只可能发生在 $\boldsymbol{x} = \boldsymbol{0}$. 因此 M_t 正定.

当 $|t| = \sqrt{6}$ 时，二次型仍然非负，但非零向量 $(-t/2, 1)^\top$ 使其取零，所以 M_t 半正

Yuan and Song

定而非正定. 当 $|t|>\sqrt{6}$ 时，取 $(1,0)^\top$ 使二次型取正值 2，取 $(-t/2,1)^\top$ 使其取负值 $3-t^2/2$，所以 M_t 不定.

这里使用二次型判断正定性的依据仍是谱定理：在单位特征向量上，二次型的值等于相应特征值. 这个例子也说明，两个对角元都为正并不足以保证矩阵正定.

例 1.4.6 (正定矩阵的平方根不是逐元素开方) 考虑

$$M=\begin{pmatrix}5&4\\4&5\end{pmatrix},\qquad Q=\frac{1}{\sqrt{2}}\begin{pmatrix}1&1\\1&-1\end{pmatrix}.$$

直接计算得 $Q^\top MQ=\mathrm{Diag}(9,1)$，所以 M 正定. 根据正定矩阵可开方的性质，

$$M^{1/2}=Q\,\mathrm{Diag}(3,1)Q^\top=\begin{pmatrix}2&1\\1&2\end{pmatrix}.$$

验证可得

$$\begin{pmatrix}2&1\\1&2\end{pmatrix}\begin{pmatrix}2&1\\1&2\end{pmatrix}=\begin{pmatrix}5&4\\4&5\end{pmatrix}=M.$$

相反，若把 M 的各个元素分别开方，得到 $B=\begin{pmatrix}\sqrt{5}&2\\2&\sqrt{5}\end{pmatrix}$，则

$$B^2=\begin{pmatrix}9&4\sqrt{5}\\4\sqrt{5}&9\end{pmatrix}\neq M.$$

矩阵平方根要求满足矩阵乘法下的 $M^{1/2}M^{1/2}=M$；谱分解中开方的是特征值.

1.4.4 内积与度量

在上一小节中，我们介绍了正定矩阵，并提及了其最重要的一条性质——正定性. 回顾我们本节最初的目标，我们要寻找一个满足正定性的双线性函数. 结合正定矩阵的正定性性质以及矩阵合同的含义，我们得出如下结论：**欲使一个双线性函数满足正定性，只需使其在任何一个基下的度量矩阵为正定矩阵即可.**

因此，从理论上来讲，任给定一个正定矩阵 M, 函数 $p(\boldsymbol{x},\boldsymbol{y}):=\boldsymbol{x}^\top M\boldsymbol{y}$ 均为一个满足正定性的双线性函数，我们称其为实数向量空间的一个广义内积. 进一步，由于通过坐标变换可以将度量矩阵做相应的合同变换，我们一定可以找到一个合适的基，使得该双线性函数对应的度量矩阵为单位阵，那么此时坐标表示后的向量 $\boldsymbol{u},\boldsymbol{v}$ 在该双线性函数下的定义即为 $p'(\boldsymbol{u},\boldsymbol{v})=\boldsymbol{u}^\top\boldsymbol{v}$. 不失一般性，我们可以取上述最简单的 $M=I$ 的形式作为一种标准的形式，将其定义为标准内积，简称内积（inner product）.

Yuan and Song

定义 1.4.7 (内积) 对于 n 维实数向量空间，定义内积为其上的一个双线性函数

$$f(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^{\top}\boldsymbol{y},\quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n.$$

内积也可记为 $\langle\boldsymbol{x},\boldsymbol{y}\rangle$.

其实，我们在正式引入内积的概念之前，已经在大量使用 $\boldsymbol{x}^{\top}\boldsymbol{y}$ 的形式了. 而本节的推导进一步告诉我们，所谓内积即为实数向量空间中一个满足正定性的双线性函数. 其中双线性性使得内积操作可以与其他多种线性操作兼容，而正定性则为在实数向量空间中引入**度量**提供了基础.

所谓度量，即衡量线性空间中一个向量的“长度”，它是一个以向量为自变量，非负数为因变量的函数，得益于内积的正定性，我们可以引入模（长）（module）的概念作为实数向量空间的度量：

定义 1.4.8 (模) 给定 $\boldsymbol{x}\in\mathbb{R}^n$, 定义其模为

$$\|\boldsymbol{x}\|:=\sqrt{\langle\boldsymbol{x},\boldsymbol{x}\rangle}.$$

有了模长的定义，两向量 $\boldsymbol{x},\boldsymbol{y}$ 间的距离可以直接表示为二者之差的模 $\|\boldsymbol{x}-\boldsymbol{y}\|$.

内积与模之间有一个重要关系，即柯西-施瓦茨不等式（Cauchy-Schwarz Inequality）：

定理 1.4.9 (柯西-施瓦茨不等式) 给定 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, 我们有不等式：

$$|\langle\boldsymbol{x},\boldsymbol{y}\rangle|\le\|\boldsymbol{x}\|\|\boldsymbol{y}\|,$$

不等式取等当且仅当 $\boldsymbol{x},\boldsymbol{y}$ 共线，即存在 $c\in\mathbb{R}$ 使得 $\boldsymbol{x}=c\boldsymbol{y}$ 或 $\boldsymbol{y}=c\boldsymbol{x}$.

柯西-施瓦茨不等式有许多种证明方式，[** 附录 **] 中给出了其中一种证明.

在柯西-施瓦茨不等式的基础上，我们有 $-1\le\dfrac{\langle\boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\le 1$, 从而我们可以定义两向量的夹角余弦为

$$\cos\theta:=\frac{\langle\boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|},$$

其中 $\theta\in[0,\pi]$ 定义为两向量的夹角.

两向量的夹角可以有多种情况，而其中一种情况我们最感兴趣，即夹角为 $\pi/2$ 的情况. 此时从几何角度来看，两向量是相互垂直的；而从代数角度来看，$\cos\theta=0\Leftrightarrow\langle\boldsymbol{x},\boldsymbol{y}\rangle=0$, 我们用正交性（orthogonality）来描述这一性质：

定义 1.4.10 (正交性) 称向量 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ 相互正交，若其满足 $\langle\boldsymbol{x},\boldsymbol{y}\rangle=0$.

例 1.4.11 (利用正交性寻找直线上最近的点) 给定 $\boldsymbol{x}=(2,0,1)^{\top}$，考虑由 $\boldsymbol{u}=(1,1,0)^{\top}$ 张成的直线，其上的点均可写为 $t\boldsymbol{u}$. 我们希望找到与 $\boldsymbol{x}$ 距离最近的一个点. 先要求误差

Yuan and Song

向量与直线方向正交：

$$\langle \boldsymbol{x} - t\boldsymbol{u}, \boldsymbol{u} \rangle = \langle \boldsymbol{x}, \boldsymbol{u} \rangle - t\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 2 - 2t = 0.$$

因此 $t = 1$，对应的点为 $\boldsymbol{p} = (1, 1, 0)^\top$，误差为 $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{p} = (1, -1, 1)^\top$. 为验证它确实最近，直接计算任意 t 下的距离平方：

$$\|\boldsymbol{x} - t\boldsymbol{u}\|^2 = (2 - t)^2 + t^2 + 1 = 2(t - 1)^2 + 3 \geq 3.$$

等号恰在 $t = 1$ 时成立，最短距离为 $\sqrt{3}$. 这说明，正交性不仅描述夹角，也可以用来确定投影点并解释最短距离.

小贴士

模（长度）、距离、夹角、正交性（垂直）这些概念原本在几何学中均为非常直观的概念. 然而，有了内积的定义，上述的概念均可通过严格的推导得出. 我们一般称配备了标准内积的实数向量空间为欧几里得空间（Euclidean space），简称欧氏空间，它是直观的空间几何的数学抽象与严格化.

我们最后来考虑一下正交矩阵的性质. 在此前，我们仅通过形式 $Q^\top = Q^{-1}$ 来判定一个矩阵为正交矩阵，却未阐释其背后的原理. 事实上，如果我们将正交矩阵 Q 分解为 n 个列向量 $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$, 则

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{q}_1^\top \\ \vdots \\ \boldsymbol{q}_n^\top \end{pmatrix} (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = \begin{pmatrix} \boldsymbol{q}_1^\top \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_1^\top \boldsymbol{q}_n \\ \vdots & & \vdots \\ \boldsymbol{q}_n^\top \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n^\top \boldsymbol{q}_n \end{pmatrix} = I_n.$$

从而有

$$\boldsymbol{q}_i^\top \boldsymbol{q}_j = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

因此，如果将正交矩阵分解为 n 个列向量，则这 n 个列向量构成了 $\mathbb{R}^n$ 中的一个单位正交基. 将正交矩阵分解为行向量可以得到类似的结果. 单位正交基的几何意义即为所有基向量两两垂直，且模长均为 1.

例 1.4.12 (改变内积会改变长度与正交关系) 取 $M = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. 由于

$$\boldsymbol{z}^\top M \boldsymbol{z} = (z_1 + z_2)^2 + z_1^2 + z_2^2,$$

它在非零向量上严格为正，所以 M 可以定义广义内积. 为区别于标准内积，本例记

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_M := \boldsymbol{x}^\top M \boldsymbol{y}, \qquad \|\boldsymbol{x}\|_M := \sqrt{\boldsymbol{x}^\top M \boldsymbol{x}}.$$

取 $\boldsymbol{x} = (1,0)^\top$, $\boldsymbol{y} = (1,-2)^\top$，有

$$M\boldsymbol{y} = (0,-3)^\top, \qquad \langle \boldsymbol{x}, \boldsymbol{y} \rangle_M = 0.$$

所以它们在这个广义内积下正交，但标准内积为 $\boldsymbol{x}^\top \boldsymbol{y} = 1$，并不为零. 相应的长度与距离为

$$\|\boldsymbol{x}\|_M = \sqrt{2}, \qquad \|\boldsymbol{y}\|_M = \sqrt{6}, \qquad \|\boldsymbol{x} - \boldsymbol{y}\|_M = 2\sqrt{2}.$$

向量的坐标没有改变，但长度、距离以及正交关系依赖于我们选定的内积.

1.5 矩阵函数

此前，我们考虑的多元函数自变量是以向量的形式表示的，我们本节考虑将自变量推广为矩阵的形式，即矩阵函数：

$$f : \mathbb{R}^{n\times m}; \quad y = f(X), X \in \mathbb{R}^{n\times n}.$$

由于矩阵 X 本质上也是一个有序数组，我们可以将所有元素按照一个特定的顺序（通常为保持每列元素连续，所有列首尾相接）展平为一个向量，我们记为 $\mathrm{vec}(X) \in \mathbb{R}^{mn}$. 从而函数 $f(X)$ 可以转化为关于向量 $\mathrm{vec}(X)$ 的普通多元函数 $\tilde{f}\big(\mathrm{vec}(X)\big)$. 因此，从理论上讲，矩阵函数完全可以通过展平操作转化为关于向量的普通多元函数；但是从实践上讲，展平操作有时会丢失矩阵的某些紧凑性质，使求导过程变得困难，或是使导函数的形式变得非常不直观. 所以，我们仍然有必要独立地研究矩阵函数的性质.

1.5.1 矩阵函数及其导数

为了使记号更加简便，我们定义 E_{ij} 为第 i 行第 j 列元素为 1, 其余元素均为 0 的矩阵. 我们定义矩阵函数的偏导数如下：

定义 1.5.1 (矩阵函数偏导数) 对于定义在 $\mathcal{X} \subseteq \mathbb{R}^{m\times n}$ 上的矩阵函数 $f(X)$, 给定一点 $X^{(0)} \in \mathcal{X}$, 若以下极限存在且有限

$$\ell = \lim_{t\to 0} \frac{f(X^{(0)} + tE_{ij}) - f(X^{(0)})}{t},$$

则称 ℓ 为多元函数 f 在 $X^{(0)}$ 点处关于第 i 行第 j 列分量的(单点)偏导数,记为 $f'_{i,j}(X^{(0)})$. 考虑所有 $X \in \mathcal{X}$, 一系列偏导数构成了一个关于矩阵 X 的矩阵函数，我们称其为偏导

函数，简称偏导数，记为 $f'_{i,j}(X)$.

矩阵函数的偏导数同样有 f' 型记号，$\partial\cdot/\partial\cdot$ 型记号等多种形式，这里不再赘述.

类比关于向量的多元函数梯度的概念，我们理应可以定义矩阵函数梯度的概念. 然而我们遇到了一些困难. 多元函数的梯度为一个同维度的向量，其行梯度、列梯度形式分别有不同的用途；而对于矩阵函数的梯度，我们可能不再能定义出一对行、列梯度的形式. 但我们依旧希望矩阵函数的梯度保留尽可能多的性质. 因此，折衷的方案为：**让矩阵函数的梯度与自变量保持相同的形状**，即 $m\times n$ 维矩阵. 这么做的优势是可以继承列梯度的与自变量直接相加减的性质，从而使得梯度下降等算法依然可行；劣势是其不再直接与链式法则相容，后续行文中我们将介绍含有矩阵函数时，链式法则应当如何处理. 我们下面严格叙述矩阵函数的梯度定义：

定义 1.5.2 (矩阵函数梯度) 对于定义在 $\mathcal{X}\subseteq\mathbb{R}^{m\times n}$ 上的矩阵函数 $f(X)$, 若其所有偏导数 $f'_{i,j}(X), i=1,2,\dots,m; j=1,2,\dots,n$ 在定义域内均存在且为连续函数，则函数 f 的梯度定义为所有分量的偏导数排列成的 $m\times n$ 维矩阵，即：

$$\nabla f(X):=\begin{pmatrix} f'_{1,1}(X) & \cdots & f'_{1,m}(X)\\ \vdots & & \vdots\\ f'_{m,1}(X) & \cdots & f'_{m,n}(X)\end{pmatrix}.$$

同样 grad$f(X)$,$f'(X)$ 也可以用来表示梯度.

本质上讲，将矩阵函数的梯度排列成与自变量相同的形状并不会带来额外的结构与信息，求偏导数并拼合成矩阵依然是最基础的方法. 然而，存在多种可以快速求得矩阵函数梯度的方法，下文的若干实例循序渐进地提供了一些求导的思路.

例 1.5.3 (把分量偏导数放回矩阵中的原位置) 对 $X=(x_{i,j})\in\mathbb{R}^{2\times 3}$，定义

$$f(X)=x_{1,1}x_{2,3}+x_{1,2}^2x_{2,1}+\sin x_{1,3}+e^{x_{2,2}}.$$

分别对六个元素求偏导并按原来的行、列位置排列，得到

$$\nabla f(X)=\begin{pmatrix} x_{2,3} & 2x_{1,2}x_{2,1} & \cos x_{1,3}\\ x_{1,2}^2 & e^{x_{2,2}} & x_{1,1}\end{pmatrix}.$$

例如，在

$$X^{(0)}=\begin{pmatrix} 1 & 2 & 0\\ -1 & 0 & 3\end{pmatrix}$$

处，有 $f(X^{(0)})=0$ 以及

$$\nabla f(X^{(0)})=\begin{pmatrix}3 & -4 & 1\\ 4 & 1 & 1\end{pmatrix}.$$

还可以直接用偏导数的定义检查第 $(1,2)$ 个元素：

$$f(X^{(0)}+tE_{12})-f(X^{(0)})=-(2+t)^2+4=-4t-t^2.$$

除以 t 并令 $t\to 0$，得到 $f'_{1,2}(X^{(0)})=-4$. 梯度与 X 同为 2×3 矩阵，每个位置表示对应元素变化时的一阶影响.

例 1.5.4 设 $A\in\mathbb{R}^{m\times n}$ 为一个常数矩阵，矩阵函数 $f:\mathbb{R}^{m\times n}\to\mathbb{R}$,

$$f(X)=\mathrm{tr}(A^\top X).$$

求解该函数的梯度可以直接将其展开为标量形式，我们不难推导出：

$$f(X)=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{i,j}x_{i,j},$$

聚焦 $x_{i,j}$, 固定其余分量后的偏导数即为

$$\frac{\partial f}{\partial x_{i,j}}=a_{i,j},$$

是一个关于 X 的常数函数. 从而我们直接将其拼合成对应形状的矩阵即可得到：

$$\nabla f(X)=A.$$

例 1.5.5 设 $A\in\mathbb{R}^{m\times m}$, 为一个常数方阵（不一定对称），矩阵函数 $f:\mathbb{R}^{m\times n}\to\mathbb{R}$,

$$f(X)=\mathrm{tr}(X^\top AX).$$

对于该函数，我们依然可以通过展开成标量形式、求偏导数、合并为梯度矩阵的方法去计算. 然而，我们注意到，如果我们将 X 分解为 $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, 每个 $\boldsymbol{x}_j$ 为一个 m 维向量，则矩阵 $X^\top AX$ 的第 i 个对角元可表示为：

$$\boldsymbol{x}_i^\top A\boldsymbol{x}_i.$$

从而函数 f 可以表示为分别关于 $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ 的 n 个独立的函数的和的形式，即

$$f(X)=\boldsymbol{x}_1^\top A\boldsymbol{x}_1+\cdots+\boldsymbol{x}_n^\top A\boldsymbol{x}_n:=\sum_{i=1}^{n}f_i(\boldsymbol{x}_i),$$

我们通常称满足这种形式的函数为一个“可分函数”. 如果我们此时考虑的是关于 $\mathrm{vec}(X)=(\boldsymbol{x}_1^\top,\dots,\boldsymbol{x}_n^\top)^\top$ 的普通多元函数，那么我们不难得出，该函数的梯度就是将关于每块的函数的梯度直接拼合. 矩阵函数的梯度由于只是将偏导数进行了重排，自然也

Yuan and Song

满足上述的性质.

参照例*1.1.13*的推导，我们可以求得

$$\nabla f_i(\boldsymbol{x}_i) = (A + A^\top)\boldsymbol{x}_i.$$

从而我们将其按照矩阵函数的梯度形状将每块进行横向拼合，即得到：

$$\nabla f(X) = (A + A^\top)X.$$

我们将上例的洞察总结为如下定理：

定理 1.5.6 (可分函数的梯度) 设函数 $f : \mathbb{R}^{n_1+\cdots+n_k} \to \mathbb{R}$, 若其可以表示为

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = f_1(\boldsymbol{x}_1) + \cdots + f_k(\boldsymbol{x}_k),$$

其中 $f_i(\boldsymbol{x}_i) : \mathbb{R}^{n_i} \to \mathbb{R}, i = 1, \ldots, k$ 为 k 个相互独立的函数，则其梯度为

$$\nabla f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = (\nabla f_1(\boldsymbol{x}_1)^\top, \ldots, \nabla f_k(\boldsymbol{x}_k)^\top)^\top.$$

例 1.5.7* 考虑行列式 $\det(X)$, 它可以视为关于方阵 $X \in \mathbb{R}^{n\times n}$ 的一个矩阵函数，且其本质上是关于矩阵 X 所有元素的一个 n 次多项式，因此是可导的，可以计算梯度.

直接计算关于单个元素的偏导数并不容易，所幸我们有行列式的拉普拉斯展开定理：

$$\det(X) = \sum_{k=1}^{n} x_{i,k} C_{i,k}, \tag{1.2}$$

其中 $C_{i,k}$ 是元素 $x_{i,k}$ 的代数余子式，即 $C_{i,k} = (-1)^{i+k} M_{i,k}$, $M_{i,k}$ 为 $x_{i,k}$ 的余子式，即矩阵 X 去掉第 i 行第 k 列后的子矩阵的行列式.

考虑 $x_{i,k}$, 我们观察到，所有 $C_{i,k}, k = 1, \ldots, n$ 中不含有第 i 行的任何元素，自然也不含有 $x_{i,k}$. 而(1.2)同时也具有可分的结构，因此我们求关于 $x_{i,k}$ 的偏导数时，只需考虑 $x_{i,k}C_{i,k}$ 这一项即可. 很显然，

$$\frac{\partial \det(X)}{\partial x_{i,k}} = C_{i,k}.$$

拼合所有的偏导数，我们可以得到：

$$\nabla \det(X) = \begin{pmatrix} C_{1,1} & \cdots & C_{1,n} \\ \vdots & & \vdots \\ C_{n,1} & \cdots & C_{n,n} \end{pmatrix}.$$

在线性代数中存在与上述矩阵十分相似的形式. 事实上，上述形式即矩阵 X 的伴随矩阵

Yuan and Song

的转置，我们记为 $\mathrm{adj}(X)^\top$. 从而我们有结论：

$$\nabla \det(X) = \mathrm{adj}(X)^\top.$$

进一步，如果 X 为可逆矩阵，我们有线性代数中另一重要结果：

$$X\,\mathrm{adj}(X) = \mathrm{adj}(X)X = \det(X)I,$$

从而 $\mathrm{adj}(X) = \det(X)X^{-1}$. 因此

$$\nabla \det(X) = \det(X)(X^{-1})^\top.$$

小贴士

多元函数的二阶导数可以表示为海瑟矩阵. 而对于矩阵函数，其二阶导数只能表示为四阶张量，或者是将其展平为向量的多元函数再求海瑟矩阵. 鉴于矩阵函数的二阶导数表示十分不方便，我们在本书中不会详细讨论.

1.5.2 矩阵内积、矩阵函数的微分

在向量空间中，我们有标准内积：

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

那么对于矩阵来讲，我们是否也可以定义一个内积呢？事实上，矩阵 X,Y 的内积就等价于将两个矩阵展平为向量 $\mathrm{vec}(X)$,$\mathrm{vec}(X)$ 后的内积，即：

$$\langle X, Y \rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{i,j}y_{i,j}, \quad X, Y \in \mathbb{R}^{m\times n}.$$

事实上，我们可以证明矩阵内积可以写为如下更紧凑的格式：

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y).$$

例 1.5.8 (非方阵的内积与迹表示) 考虑两个 2×3 矩阵

$$X = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 3 & 1 \end{pmatrix}, \qquad Y = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 0 & 3 \end{pmatrix}.$$

按照逐元素相乘后求和的定义，

$$\langle X, Y \rangle = 1\cdot 2 + (-1)\cdot 1 + 2\cdot 0 + 0\cdot(-1) + 3\cdot 0 + 1\cdot 3 = 4.$$

另一方面，

$$X^\top Y = \begin{pmatrix} 2 & 1 & 0 \\ -5 & -1 & 9 \\ 3 & 2 & 3 \end{pmatrix}, \qquad \mathrm{tr}(X^\top Y) = 2 - 1 + 3 = 4.$$

若按列展平，则

$$\mathrm{vec}(X) = (1, 0, -1, 3, 2, 1)^\top, \qquad \mathrm{vec}(Y) = (2, -1, 1, 0, 0, 3)^\top,$$

这两个向量的内积同样为 4. 因此矩阵内积是一个标量；它并不是矩阵乘法 XY，本例中的 XY 甚至因维度不匹配而没有定义.

对于向量的多元函数，我们有全微分

$$\Delta z = f_1'(\boldsymbol{x})\Delta x_1 + \cdots + f_n'(\boldsymbol{x})\Delta x_n + o(\rho) = \nabla f(\boldsymbol{x})^\top \Delta \boldsymbol{x} + o(\rho),$$

或

$$\mathrm{d}z = \nabla f(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x},$$

这里的 $\nabla f(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}$ 其实正是内积，我们可以自然地将多元函数的微分推广到矩阵函数中，即

$$\Delta z = \langle \nabla f(X), \Delta X \rangle + o(\rho),$$

此处 ρ 的定义应当类推为 $\rho = \sqrt{\sum_{i=1}^m \sum_{j=1}^n x_{i,j}^2}$, 或可写成微分形式

$$\mathrm{d}z = \langle \nabla f(X), \mathrm{d}X \rangle.$$

无论关于向量还是矩阵的微分，其本质上均是一个多元函数的极限，即关于 $\Delta \boldsymbol{x}$ 或是 ΔX 的极限. 这个极限涉及到 n 个乃至 $m \times n$ 个自变量，并不好研究. 因此与之相对的，可微性有另一个稍微弱一些的版本，即将多元函数极限转换为任意方向上的一元函数极限. 我们称这个弱一些的版本为 Gâteaux 可微性，而称原先的标准版本为 Fréchet 可微性. Gâteaux 可微性的定义如下（我们以矩阵形式为例，向量形式可以类推得到）：

定义 1.5.9 (Gâteaux 可微性) 设函数 $z = f(X) : \mathbb{R}^{m\times n} \to \mathbb{R}$ 在点 X_0 的一个邻域中有定义，任意给定方向矩阵 $V \in \mathbb{R}^{m\times n}$, 记 $\Delta z(t) := f(X_0 + tV) - f(X_0)$, 其中 $t \in \mathbb{R}$ 为自由变元. 如果 Δz 满足

$$\Delta z(t) = \langle G, tV \rangle + o(t),$$

其中 $G \in \mathbb{R}^{m\times n}$ 为不依赖于 V 及 t 的常数，则我们称函数 $f(X)$ 在点 X_0 处 *Gâteaux* 可微.

注意：Fréchet 可微性可以直接推导出可导性，且微分中的线性主部系数就是所有的

Yuan and Song

偏导数. Gâteaux 可微性虽然弱于 Fréchet 可微性，但仍然可以推出可导性且线性主部与偏导数完全相同，证明思路与 Fréchet 可微性完全一致. 我们用如下示意图表达三者的关系，其中右侧的条件成立可以推导出左侧的条件成立.

$$\boxed{\text{可导性} \Leftarrow \text{Gâteaux 可微性} \Leftarrow (\text{Fréchet}) \text{可微性}}$$

小贴士

事实上，机器学习中绝大部分的多元函数或是矩阵函数都是性质良好的函数. 可导但不是 Gâteaux 可微，或是 Gâteaux 可微但不是 Fréchet 可微的函数仅存在于少数一些数学分析的复杂例子中.

例 1.5.10 (由矩阵方向上的增量求梯度) 对方阵 $X \in \mathbb{R}^{n\times n}$，定义 $f(X) = \operatorname{tr}(X^3)$. 固定一个方向矩阵 V，展开时保留矩阵乘积的顺序：

$$(X+tV)^3 = X^3 + t(X^2V + XVX + VX^2) + O(t^2).$$

这里对固定的 X, V，省略项中 t 的次数至少为 2. 利用迹的循环性质，三项 X^2V, XVX, VX^2 的迹相等，于是

$$\begin{aligned} f(X+tV) - f(X) &= 3t\operatorname{tr}(X^2V) + o(t) \\ &= \langle 3(X^2)^\top, tV\rangle + o(t). \end{aligned}$$

与可微性的定义比较，得到

$$\nabla f(X) = 3(X^2)^\top.$$

例如，取 $X_0 = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$，则

$$X_0^2 = \begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix}, \qquad \nabla f(X_0) = \begin{pmatrix} 3 & 0 \\ 9 & 12 \end{pmatrix}.$$

在方向 $V = E_{21}$ 上，一阶变化率为 $\langle \nabla f(X_0), E_{21}\rangle = 9$. 直接展开也得到 $f(X_0 + tE_{21}) = 9 + 9t$. 本例中的转置来自内积 $\langle G, V\rangle = \operatorname{tr}(G^\top V)$，不能省略.

可微性的实用之处在于，如果我们通过某种方式事先“猜测”出了某个函数梯度的可能形式，同时又证明了其满足可微的定义式，那么我们就可以直接断言上述猜测的形式就是该函数的梯度. 下述两个例子提供了利用可微性求得梯度的典型应用.

例 1.5.11* (正定矩阵行列式的 Gâteaux 导数) 考虑定义在 n 阶正定矩阵集合 $\mathcal{S}_{++}^n$ 上的

Yuan and Song

函数

$$f : \mathcal{S}_{++}^n \to \mathbb{R}, \qquad f(X) = \det(X).$$

利用 *Gâteaux* 可微性的定义，求 f 在正定矩阵 X 处的梯度.

解：由于自变量属于实对称矩阵空间，我们任取实对称方向矩阵 V. 由正定矩阵的可开方性，$X^{1/2}$ 存在且可逆，记 $X^{-1/2} := (X^{1/2})^{-1}$，则

$$X + tV = X^{1/2}(I_n + tX^{-1/2}VX^{-1/2})X^{1/2}.$$

令 $A = X^{-1/2}VX^{-1/2}$. 由于 $X^{-1/2}$ 与 V 均为实对称矩阵，有 $A^\top = A$，因此可利用谱定理作正交分解

$$A = Q\Lambda Q^\top, \qquad \Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n),$$

其中 Q 为正交矩阵，$\lambda_1, \ldots, \lambda_n$ 均为实数. 当 $|t|$ 充分小时，所有 $1 + t\lambda_i > 0$，所以 $I_n + tA$ 及与之合同的 $X + tV$ 均正定，函数值 $f(X + tV)$ 有定义.

根据行列式的乘法性质以及 $Q^\top Q = I_n$，有

$$\begin{aligned}\det(X + tV) &= \det(X^{1/2})^2 \det(I_n + tA) \\ &= \det(X) \det\big(Q(I_n + t\Lambda)Q^\top\big) \\ &= \det(X) \prod_{i=1}^n (1 + t\lambda_i).\end{aligned}$$

展开乘积，常数项为 1，一次项为 $t\sum_i \lambda_i$，其余项至少含有 t^2，因此

$$\prod_{i=1}^n (1 + t\lambda_i) = 1 + t\sum_{i=1}^n \lambda_i + o(t).$$

另一方面，利用迹在相似变换下的不变性及其循环性质，

$$\begin{aligned}\sum_{i=1}^n \lambda_i &= \mathrm{tr}(\Lambda) = \mathrm{tr}(A) \\ &= \mathrm{tr}(X^{-1/2}VX^{-1/2}) = \mathrm{tr}(X^{-1}V).\end{aligned}$$

所以函数增量为

$$\begin{aligned}\Delta z(t) &:= f(X + tV) - f(X) \\ &= t \det(X) \,\mathrm{tr}(X^{-1}V) + o(t).\end{aligned}$$

除以 t 并令 $t \to 0$，得到沿方向 V 的方向导数：

$$\lim_{t \to 0} \frac{f(X + tV) - f(X)}{t} = \det(X) \,\mathrm{tr}(X^{-1}V).$$

最后，X^{-1} 仍为实对称矩阵，根据本节的矩阵内积 $\langle G, V \rangle = \mathrm{tr}(G^\top V)$，上述增量可

Yuan and Song

写为

$$\Delta z(t) = \langle \det(X)X^{-1}, tV \rangle + o(t).$$

其中 $G = \det(X)X^{-1}$ 为实对称矩阵，且仅依赖于 X，不依赖于方向 V 或 t. 因此，按照实对称矩阵空间中同一内积下的 *Gâteaux* 可微性定义，得到

$$\nabla f(X) = \det(X)X^{-1}.$$

相应的微分形式为

$$\mathrm{d}\big(\det(X)\big) = \det(X)\,\mathrm{tr}(X^{-1}\mathrm{d}X).$$

这与例*1.5.7*的结果一致，因为正定情形下 $(X^{-1})^\top = X^{-1}$.

例 1.5.12* 考虑定义在 n 阶正定矩阵集合 $\mathcal{S}_{++}^n$ 上的矩阵函数 f:

$$f(X) = \ln\big(\det(X)\big).$$

我们下面使用 *Gâteaux* 可微性来求该函数的梯度. 给定正定矩阵 X, 由于正定矩阵的线性组合仍然为实对称矩阵，因此我们应该任意假设方向矩阵 V 为实对称矩阵，同时考虑 $t \in \mathbb{R}$, 有

$$\begin{aligned} f(X+tV) - f(X) &= \ln\big(\det(X+tV)\big) - \ln\big(\det(X)\big) \\ &= \ln\Big(\det\big(X^{1/2}(I + tX^{-1/2}VX^{-1/2})X^{1/2}\big)\Big) - \ln\big(\det(X)\big) \\ &= \ln\big(\det(X)\det(I + tX^{-1/2}VX^{-1/2})\big) - \ln\big(\det(X)\big) \\ &= \ln\big(\det(I + tX^{-1/2}VX^{-1/2})\big), \end{aligned}$$

其中 $X^{1/2}$ 的分解得益于正定矩阵的可开方性，$X^{-1/2} := (X^{1/2})^{-1}$. 我们下面考虑将 $X^{-1/2}VX^{-1/2}$ 正交对角化，即 $X^{-1/2}VX^{-1/2} = Q\Lambda Q^\top$, 其中 $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, 则

$$\begin{aligned} &\ln\big(\det(I + tX^{-1/2}VX^{-1/2})\big) \\ &= \ln\big(\det(QIQ^\top + tQ\Lambda Q^\top)\big) = \ln\big(\det(I + t\Lambda)\big) \\ &= \ln\prod_{i=1}^n (1+t\lambda_i) = \sum_{i=1}^n \ln(1+t\lambda_i) = \sum_{i=1}^n t\lambda_i + o(t) \\ &= t\cdot\mathrm{tr}(X^{-1/2}VX^{-1}) + o(t) = \langle (X^{-1})^\top, tV\rangle + o(t), \end{aligned}$$

其中第三行第三个等号使用了 ln 函数的泰勒展开式，第四行第一个等号使用了相似矩阵的保持迹的性质 $\mathrm{tr}(QAQ^{-1}) = \mathrm{tr}(Q^{-1}QA) = \mathrm{tr}(A)$. 自此，我们将函数 f 成功写为了 *Gâteaux* 导数的形式，同时由于 $X^\top = X$, 我们有

$$\nabla f(X) = X^{-1}.$$

Yuan and Song

小贴士

使用可微性来推导梯度的方法是一种“事后诸葛亮”的方法，猜测和推导梯度的可能形式往往需要很强的洞察力. 在下一小节中，我们将介绍一些更加具有可操作性的求导策略.

1.5.3 复合求导法则

我们在引入矩阵函数的梯度时，曾指出其不再直接支持向量值函数的链式法则. 我们自然可以选择将矩阵 X 展平为向量 $\mathrm{vec}(X)$ 再使用链式法则，但是这往往会显著增加求导的难度，因此我们需要使用一些额外的工具来帮助我们快速求解复合下的矩阵函数的梯度.

我们在先前已指出，微分记号 $\mathrm{d}\boldsymbol{x}, \mathrm{d}X$ 等是一种不严谨的用于快速推导梯度的记号. 对于函数 $f : \mathbb{R}^{m\times n} \to \mathbb{R}, X \mapsto z$, 如果我们能通过若干微分运算推导得出

$$\mathrm{d}z = \langle G, \mathrm{d}X\rangle,$$

则我们就可以断言 $\nabla f(X) = G$. 这实质上正是整理 Fréchet 或 Gâteaux 微分表达式的过程，只是我们默认余项一定是无穷小量，无需担心余项可能出现的问题，这在绝大多数问题中都是满足的. 此外，微分运算下也存在多种形式优雅的运算法则，以使得求导更加便利，我们将会在本小节介绍最常用的几条微分法则.

需要注意的是，微分记号 $\mathrm{d}X$ 可能有不同的含义，如果 X 是底层的自变量，那么 $\mathrm{d}X$ 就是一个自由变元，无需进一步运算；如果 X 是一个中间变量，那么我们需要将其所有分量分别考虑为一个多元函数的微分，再分别展开直至得到底层的自变量.

下述第一条定理与链式法则的关系最为密切，当外层为一个一元函数，内层为一个矩阵函数时，微分运算与链式法则完全一致.

定理 1.5.13 设自变量 $X \in \mathbb{R}^{m\times n}$, 中间变量 $y = f(X) \in \mathbb{R}$, 因变量 $z = g(y) \in \mathbb{R}$, 则：

$$\mathrm{d}z = g'(y)\mathrm{d}y = \Big\langle g'\big(f(X)\big)\nabla f(X), \mathrm{d}X\Big\rangle.$$

证明. 考虑偏导数

$$\frac{\partial z}{\partial x_{i,j}} = g'\big(f(X)\big)f'_{i,j}(X),$$

直接拼合成矩阵即可得到最终结果. ■

以下例子使用上述定理来验证例1.5.12中 ln-det 函数导数的正确性.

例 1.5.14 (续例1.5.12) 再次考虑定义在 n 阶正定矩阵集合上的函数 $h(X) := \ln\big(\det(X)\big)$, 其中外层函数 $z = g(y) = \ln y$ 为一元函数, 其导数易求得为 $g'(y) = y^{-1}$; 内层 $y = f(X) = \det(X)$ 为行列式对应的矩阵函数, 使用例*1.5.7*的结果, 我们得到:

$$\nabla f(X) = \det(X)(X^{-1})^\top = \det(X)X^{-1},$$

其中第二个等式是因为 X 为正定矩阵. 从而我们直接使用定理*1.5.13*得:

$$\nabla h(X) = g'(f(X))\nabla f(X) = \frac{1}{\det(X)}\det(X)X^{-1} = X^{-1}.$$

这与我们在例*1.5.12*中的计算结果是完全吻合的.

以下两个性质与内积运算及矩阵的迹运算有关, 它们的证明也是比较直接的.

性质 1.5.15 设自变量或中间变量为 $X \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{m\times n}$ 为常数矩阵, 则

$$\mathrm{d}\big(\langle A, X\rangle\big) = \langle A, \mathrm{d}X\rangle$$

证明. 直接根据矩阵函数的微分形式来展开函数 $f(X) = \langle A, X\rangle$ 即可. ■

性质 1.5.16 设自变量或中间变量为 $X \in \mathbb{R}^{n\times n}$, 则

$$\mathrm{d}\big(\mathrm{tr}(X)\big) = \mathrm{tr}\big(\mathrm{d}X\big).$$

证明. 由于 $\mathrm{tr}(X) = \mathrm{tr}(I_n X) = \langle I_n, X\rangle$, 因此直接使用性质1.5.15有:

$$\mathrm{d}\big(\mathrm{tr}(X)\big) = \mathrm{d}\big(\langle I_n, X\rangle\big) = \langle I_n, \mathrm{d}X\rangle = \mathrm{tr}(I_n \mathrm{d}X) = \mathrm{tr}(\mathrm{d}X).$$

■

以下三个性质则是进一步处理矩阵作为中间变量时的推导方式. 它们如果展平为向量再使用链式法则进行证明的话将会非常复杂, 而由于它们都是线性的运算, 直接拆分成分量进行推导反而更加容易.

性质 1.5.17 设 $X \in \mathbb{R}^{n\times n}$ 为自变量或中间变量, 则

$$\mathrm{d}(X^\top) = (\mathrm{d}X)^\top.$$

证明. 转置仅影响分量的排列顺序, 不会影响分量自身的取值, 上述等式自然成立. ■

Yuan and Song

性质 1.5.18 设 $X, Y \in \mathbb{R}^{m\times n}$ 为自变量或中间变量，则

$$\mathrm{d}(X+Y) = \mathrm{d}X + \mathrm{d}Y.$$

证明. 记 $Z = X + Y$, 则 $z_{i,j} = x_{i,j} + y_{i,j}$, 此时我们直接有

$$\mathrm{d}z_{i,j} = \mathrm{d}x_{i,j} + \mathrm{d}y_{i,j},$$

直接拼合成矩阵即得证. ■

性质 1.5.19 设 $X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{n\times p}$ 为自变量或中间变量，则

$$\mathrm{d}(XY) = (\mathrm{d}X)Y + X(\mathrm{d}Y).$$

证明. 记 $Z = XY$, 则 $z_{i,j} = \sum_{k=1}^{n} x_{i,k}y_{k,j}$. 使用一元函数的乘积微分法则 $\mathrm{d}(xy) = x\mathrm{d}y + y\mathrm{d}x$, 我们得到：

$$\mathrm{d}z_{i,j} = \sum_{k=1}^{n}(x_{i,k}\mathrm{d}y_{k,j} + y_{k,j}\mathrm{d}x_{i,k}) = \sum_{k=1}^{n} x_{i,k}\mathrm{d}y_{k,j} + \sum_{k=1}^{n} y_{k,j}\mathrm{d}x_{i,k},$$

再拼合成矩阵我们即可得到最后结果. ■

小贴士

1. 对于性质1.5.19，注意矩阵乘积不满足交换律，矩阵乘积的顺序不可变. 进一步，如果矩阵 A, B 为常数矩阵，则 $\mathrm{d}A = 0, \mathrm{d}B = 0$, 从而我们有：

$$\mathrm{d}(AX) = A\mathrm{d}X; \quad \mathrm{d}(XB) = (\mathrm{d}X)B.$$

2. 以上所有性质均可适用于普通的多元函数，只需将其视为列数为 1 的列矩阵即可，此时求出的梯度为列梯度.

使用上述的微分运算法则，我们来再次分别验证例1.5.4, 例1.5.5求得梯度的正确性.

例 1.5.20 (续例1.5.4) 考虑函数 $f(X) = \mathrm{tr}(A^\top X)$, 其中 A 为常数矩阵.

使用微分形式，

$$\mathrm{d}\big(\mathrm{tr}(A^\top X)\big) = \langle A, \mathrm{d}X\rangle,$$

从而我们立即得到 $\nabla f(X) = A$.

例 1.5.21 (续例1.5.5) 考虑函数 $f(X) = \mathrm{tr}(X^\top AX)$, 其中 A 为常数方阵（不一定对称）.

使用微分形式，

$$\begin{aligned}\mathrm{d}\big(\mathrm{tr}(X^\top AX)\big) &= \mathrm{tr}\big(\mathrm{d}(X^\top AX)\big) = \mathrm{tr}\big(X^\top \mathrm{d}(AX) + (\mathrm{d}X^\top)(AX)\big)\\ &= \mathrm{tr}(X^\top A\mathrm{d}X) + \mathrm{tr}\big((\mathrm{d}X)^\top(AX)\big) = \langle A^\top X, \mathrm{d}X\rangle + \langle AX, \mathrm{d}X\rangle\\ &= \big\langle (A^\top + A)X, \mathrm{d}X\big\rangle.\end{aligned}$$

从而我们可以得到 $\nabla f(X) = (A + A^\top)X$.

例 1.5.22 (矩阵两侧都含常数因子的平方损失) 设 $A \in \mathbb{R}^{p\times m}$, $B \in \mathbb{R}^{n\times q}$, $C \in \mathbb{R}^{p\times q}$ 为常数矩阵，自变量为 $X \in \mathbb{R}^{m\times n}$. 记

$$R = AXB - C, \qquad f(X) = \frac{1}{2}\langle R, R\rangle = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{q} r_{i,j}^2.$$

对每个残差分量求微分，可得 $\mathrm{d}f = \langle R, \mathrm{d}R\rangle$. 由于 A, B, C 固定，乘积微分法则给出 $\mathrm{d}R = A(\mathrm{d}X)B$. 因此

$$\begin{aligned}\mathrm{d}f &= \mathrm{tr}\big(R^\top A(\mathrm{d}X)B\big)\\ &= \mathrm{tr}\big(BR^\top A\,\mathrm{d}X\big)\\ &= \langle A^\top RB^\top, \mathrm{d}X\rangle.\end{aligned}$$

从而

$$\nabla f(X) = A^\top(AXB - C)B^\top \in \mathbb{R}^{m\times n}.$$

例如，取

$$A = \begin{pmatrix}1 & 2\end{pmatrix}, \quad B = \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}, \quad C = \begin{pmatrix}0 & 1\end{pmatrix}, \quad X = I_2,$$

则 $R = \begin{pmatrix}3 & 1\end{pmatrix}$，并且

$$\nabla f(I_2) = \begin{pmatrix}1\\ 2\end{pmatrix}\begin{pmatrix}3 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} = \begin{pmatrix}3 & 4\\ 6 & 8\end{pmatrix}.$$

左右两个常数因子分别作用于矩阵的行、列；整理微分时保留乘法顺序，就能得到与自变量同形状的梯度.

我们最后来叙述一个在深度学习前反向传播中广泛使用的结论，它是微分形式的性质1.5.19的直接推论.

推论 1.5.23 设自变量或中间变量 $X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{n\times p}$, 中间变量 $Z = XY$, 因变量 $l = f(Z)$, 其中 $\nabla f(\cdot)$ 已知，则我们有分块导数：

$$\frac{\partial l}{\partial X} = \nabla f(Z)Y^\top; \quad \frac{\partial l}{\partial Y} = X^\top \nabla f(Z).$$

Yun and Song

证明. 当求关于 X 的偏导数时，Y 视为常数. 则由微分形式：

$$\begin{aligned}\mathrm{d}l &= \langle \nabla f(Z), \mathrm{d}Z\rangle = \langle \nabla f(Z), (\mathrm{d}X)Y\rangle = \operatorname{tr}\big(\nabla f(Z)^\top (\mathrm{d}X)Y\big)\\ &= \operatorname{tr}\big(Y\nabla f(Z)^\top \mathrm{d}X\big) = \langle \nabla f(Z)Y^\top, \mathrm{d}X\rangle.\end{aligned}$$

从而有 $\dfrac{\partial l}{\partial X} = \nabla f(Z)Y^\top$. 关于 Y 的分块导数可以类似地得到. ■

例 1.5.24 (矩阵乘积两端的分块导数) 设 $X, Y \in \mathbb{R}^{2\times 2}$，中间变量 $Z = XY$，并定义标量函数

$$l = f(Z) = \frac{1}{3}\sum_{i=1}^{2}\sum_{j=1}^{2} z_{i,j}^3.$$

逐元素求导得到

$$G := \nabla f(Z) = \begin{pmatrix} z_{1,1}^2 & z_{1,2}^2 \\ z_{2,1}^2 & z_{2,2}^2 \end{pmatrix}.$$

根据推论*1.5.23*，求 X 的分块导数时固定 Y，求 Y 的分块导数时固定 X，于是

$$\frac{\partial l}{\partial X} = GY^\top, \qquad \frac{\partial l}{\partial Y} = X^\top G.$$

具体取

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad Y = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix},$$

先计算 $Z = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$，再得到 $G = \begin{pmatrix} 1 & 4 \\ 1 & 1 \end{pmatrix}$. 最终

$$\frac{\partial l}{\partial X} = \begin{pmatrix} 9 & -4 \\ 3 & -1 \end{pmatrix}, \qquad \frac{\partial l}{\partial Y} = \begin{pmatrix} 2 & 5 \\ 1 & 1 \end{pmatrix}.$$

例如，增加 $x_{1,2}$ 会使 $z_{1,2}$ 按变化率 $y_{2,2} = -1$ 变化，因此这一项对 $\partial l/\partial x_{1,2}$ 的贡献为 $4\cdot(-1) = -4$；另一项因 $y_{2,1} = 0$ 没有贡献. 这与上面的矩阵计算一致.

1.6　范数与奇异值分解

在本节中，我们将考虑两个问题. 其一、我们先前在实数向量空间中引入了度量，即模长. 我们可能会进一步考虑如下问题：**是否存在其他形式的关于向量 x 的非负函数，也能充当实数向量空间中的一种度量**？其二、对于方形矩阵，我们可以通过特征值来直

Yuanqing

观地表示其作为线性变换时对向量的缩放作用，而**对于非方形矩阵，其对应的线性映射的放大或缩小作用要使用什么样的度量来刻画呢**？通过本节的学习，我们可以认识到上述两个问题存在一定的联系，我们也将对实数向量空间的度量有更深刻的理解.

1.6.1 范数

所谓度量即线性空间上的一个具有特殊性质的非负函数，而模长 $\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ 即是实数向量空间上的一个最典型的度量. 我们将模长满足的性质抽象为以下的三点：

1. **正定性**：$\|\boldsymbol{x}\| \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^n$, 且 $\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$.
2. **齐次性**：对于任意 $c \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^n$, 有 $\|c\boldsymbol{x}\| = |c|\|\boldsymbol{x}\|$.
3. **三角不等式**：对于任意 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, 有 $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

其中第三条性质可由柯西-施瓦茨不等式得到.

满足上述三条性质的函数自然不只有模长这一个函数，在实数向量空间中可以找到许多这样的函数，我们称这些函数为 $\mathbb{R}^n$ 上的范数（norm）：

定义 1.6.1 (范数) *实数向量空间 $\mathbb{R}^n$ 上的函数 $m : \mathbb{R}^n \to \mathbb{R}$ 若满足如下三个条件：*

1. *正定性：$m(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^n$, 且 $m(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$.*
2. *齐次性：对于任意 $c \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^n$, 有 $m(c\boldsymbol{x}) = |c|m(\boldsymbol{x})$.*
3. *三角不等式：对于任意 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, 有 $m(\boldsymbol{x} + \boldsymbol{y}) \leq m(\boldsymbol{x}) + m(\boldsymbol{y})$.*

则称函数 $m(\cdot)$ 为实数向量空间 $\mathbb{R}^n$ 的一个范数.

以下我们尝试寻找除模长外其他的实数向量空间中的范数. 我们观察到，如果记 $\boldsymbol{x} = (x_1, \ldots, x_n)$, 则模长可以表示为：

$$\|\boldsymbol{x}\| = \left(\sum_{i=1}^{n} x_n^2 \right)^{1/2},$$

一个自然的想法是，上述式子中的 2 与 $1/2$ 能否换成其他数？这当然有可能，只要我们能满足上述三条性质即可. 我们初步拟定一个可能的范数形式应当为 $m(\boldsymbol{x}) = (\sum_{i=1}^{n} x_i^p)^q$. 欲满足正定性，我们可能需要在每个分量上添加绝对值，这样该函数就恒为非负数了，且函数取零当且仅当每个分量取零. 欲满足齐次性，我们进一步需要 $q = p^{-1}$. 因此，上述形式进一步化为了 $m(\boldsymbol{x}) = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$. 唯一比较棘手的是三角不等式，事实上，当 $p \geq 1$ 时，三角不等式是成立的，相关的推导比较繁琐，我们将其放在了[** 附录 **]中. 综上所述，我们找到了一系列实数向量空间上的范数，其包含一个参数 p, 从而我们将其称为 p-范数（或 L_p 范数/ℓ_p 范数，其中“L”为 Lebesgue 的简写），记作 $\|\cdot\|_p$：

Yuan and Song

定义 1.6.2 (p-范数/L_p 范数/ℓ_p 范数) 实数向量空间 $\mathbb{R}^n$ 中的函数

$$\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad p \in [1, +\infty]$$

满足范数的定义，称为 p-范数.

在上述定义中，p 可以取正无穷，此时范数应当取极限的定义，而我们有如下极限成立：

$$\lim_{p\to+\infty} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} = \max_{1\le i\le n} |x_i|.$$

证明. 记 $M = \max_{1\le i\le n} |x_i|$, 则有：

$$M = (M^p)^{1/p} \le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \le (nM^p)^{1/p} = \sqrt[p]{n}M.$$

由于极限

$$\lim_{p\to+\infty} M = M = \lim_{p\to+\infty} \sqrt[p]{n}M,$$

因此中间序列极限也为 M, 即得证. ■

从众多 p-范数中，以下三个范数最具有代表性，使用频率也最高：

- 1-范数：$\|\boldsymbol{x}\|_1 = |x_1| + \cdots + |x_n|, \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.
- 2-范数：$\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}, \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.
- ∞-范数：$\|\boldsymbol{x}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}, \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

小贴士

实数向量空间的模长等于实数向量空间的 2-范数，此外还存在实数向量空间的 1-范数. 然而，不存在实数向量空间的“1-模”的说法.

我们知道，两个向量的（欧氏）距离就是两个向量之差的模（或 2-范数）. 类比地，两个向量之差的 1-范数或 ∞-范数也可以分别定义一种距离. 我们称 1-范数对应的距离称为“曼哈顿距离”或“街区距离”，因其形象地描绘了 1-范数下沿着坐标轴“街区”行走时走过的距离；称 ∞-范数对应的距离为切比雪夫（Chebyshev）距离.

例 1.6.3 (不同范数对两个误差向量的比较) 假设两个近似结果的误差向量分别为

$$\boldsymbol{a} = (3, 0)^\top, \qquad \boldsymbol{b} = (2, 2)^\top.$$

Yuna Song

按照三种常用范数的定义，分别得到

$$\|\boldsymbol{a}\|_1 = 3, \qquad \|\boldsymbol{b}\|_1 = 4,$$
$$\|\boldsymbol{a}\|_2 = 3, \qquad \|\boldsymbol{b}\|_2 = 2\sqrt{2},$$
$$\|\boldsymbol{a}\|_\infty = 3, \qquad \|\boldsymbol{b}\|_\infty = 2.$$

因此，用 1-范数衡量误差时，$\boldsymbol{a}$ 较小；用 2-范数或 ∞-范数衡量时，$\boldsymbol{b}$ 较小. 这里 1-范数累加各分量的绝对误差，而 ∞-范数只取最坏分量的绝对误差，所以不同的衡量方式可能给出不同的大小排序.

图1.2分别绘制了二维空间中 $\|\boldsymbol{x}\|_1 = 1, \|\boldsymbol{x}\|_2 = 1, \|\boldsymbol{x}\|_\infty = 1$ 三个范数值等于 1 时的等势面. 三个范数的等势面分别为菱形、球形以及正方形.



图 1.2: 1-范数，2-范数，∞-范数的值 1 等势面示意图

例 1.6.4 (长度的平方为什么不是范数) 在 $\mathbb{R}^2$ 上定义 $m(\boldsymbol{x}) = x_1^2 + x_2^2$. 它恒为非负，且只在零向量处取零，但还需要检查范数的另外两条性质. 取 $\boldsymbol{u} = (1, 0)^\top$，则

$$m(2\boldsymbol{u}) = 4, \qquad |2|m(\boldsymbol{u}) = 2,$$

不满足齐次性. 再取 $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{u}$，有

$$m(\boldsymbol{x} + \boldsymbol{y}) = 4 > 2 = m(\boldsymbol{x}) + m(\boldsymbol{y}),$$

也不满足三角不等式. 因此，平方长度虽然可以用于描述误差，却不能直接称为范数；取平方根后的 $\sqrt{m(\boldsymbol{x})} = \|\boldsymbol{x}\|_2$ 才是本节介绍的 2-范数.

实数向量空间中的范数自然也不只局限于 p-范数. 事实上我们可以证明，两个范数的和构成的新的函数也是范数，除此之外仍然有其他的构造范数的方式. 我们不再一一列举更多的范数. 而且我们如下的范数等价性定理. 它告诉我们，不同的范数之间只存在常数的差异，因此无需考虑过多的范数结构. 范数等价性定理的证明需要用到拓扑的紧致性等高级技巧，因此此处省略.

定理 1.6.5 (范数等价性定理) 设 $m_1(\cdot), m_2(\cdot)$ 为 $\mathbb{R}^n$ 上的两个范数（n 为有限值），则存在两个常数 $c_1, c_2 > 0$，使得 $\forall \boldsymbol{x} \in \mathbb{R}^n$, 都有：

$$m_1(\boldsymbol{x}) \leq c_1 m_2(\boldsymbol{x}), \quad m_2(\boldsymbol{x}) \leq c_2 m_2(\boldsymbol{x}).$$

1.6.2 矩阵范数

我们下面来考虑矩阵的范数. 仍然沿袭我们先前的思路，矩阵 X 可以通过展平操作 $\mathrm{vec}(X)$ 变为一个向量，我们自然可以类比向量的 1-范数、2-范数、∞-范数推导出三种可能的矩阵范数的形式. 它们分别可表示为

$$\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{i,j}|, \quad \left(\sum_{i=1}^{m}\sum_{k=1}^{n}x_{i,j}^2\right)^{1/2}, \quad \max_{1\leq i\leq m, 1\leq j\leq n}|x_{i,j}|.$$

然而，我们知道，矩阵相比其展平后的向量，还有额外的结构，矩阵乘法运算便是一个最基本的体现. 回到我们在本节最开始提出的第二个问题，如果我们将矩阵视为一个线性映射或线性变换，我们如何去度量矩阵带来的放大倍数作用呢？如果某个矩阵范数能够度量这一放大倍数的话，我们从直观上来讲需要满足如下的条件：两次线性映射的复合带来的放大的倍数，不会比两次线性映射分别放大的倍数的乘积要大，因为两次放大最大倍数的方向可能不同.

我们下面考虑用代数形式刻画上面这条性质. 我们定义一系列矩阵线性空间 $\mathbb{R}^{m\times n}$, $m = 1, 2, \ldots; n = 1, 2, \ldots$ 上的一系列函数 $\mu_{m,n} : \mathbb{R}^{m\times n} \to \mathbb{R}, m = 1, 2, \ldots; n = 1, 2, \ldots$. 则上述这条性质可以表示为：

$$\mu_{m,q}(XY) \leq \mu_{m,n}(X)\mu_{n,q}(Y), \quad \forall m, n, q = 1, 2, \ldots; \forall X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{n\times q}.$$

我们称上述性质为**相容性**，该性质保证了矩阵范数可以作为矩阵线性映射放大倍数的良好度量. 相比普通的向量范数，矩阵范数需要同步地定义一系列范数，且需要额外满足相容性. 我们将矩阵范数的定义陈述如下：

定义 1.6.6 (矩阵范数) 一系列实数矩阵线性空间 $\mathbb{R}^{m\times n}, m = 1, 2, \ldots; n = 1, 2, \ldots$ 上的一系列函数 $\mu_{m,n} : \mathbb{R}^{m\times n} \to \mathbb{R}, m = 1, 2, \ldots; n = 1, 2, \ldots$ 若满足如下的四个条件：

1. 正定性：$\mu_{m,n}(X) \geq 0, \forall m, n = 1, 2, \ldots, \forall X \in \mathbb{R}^{m\times n}$, 且 $\mu_{m,n}(X) = 0 \Leftrightarrow X = \mathbf{0}$.
2. 齐次性：$\forall c \in \mathbb{R}, \forall m, n = 1, 2, \ldots, \forall X \in \mathbb{R}^{m\times n}$, 有 $\mu_{m,n}(cX) = |c|\mu_{m,n}(X)$.
3. 三角不等式：$\forall m, n = 1, 2, \ldots, \forall X, Y \in \mathbb{R}^{m\times n}$, 有

$$\mu_{m,n}(X + Y) \leq \mu_{m,n}(X) + \mu_{m,n}(Y).$$

4. 相容性：$\forall m,n,q=1,2,\ldots,\forall X\in\mathbb{R}^{m\times n},Y\in\mathbb{R}^{n\times q}$, 有：

$$\mu_{m,q}(XY)\leq\mu_{m,n}(X)\mu_{n,q}(Y).$$

则称每个函数 $\mu_{m,n}(\cdot)$ 为对应的实数矩阵线性空间 $\mathbb{R}^{m\times n}$ 上的一个矩阵范数.

例 1.6.7 (最大元素绝对值与相容性) 对矩阵 X 定义 $m(X)=\max_{i,j}|x_{i,j}|$. 将矩阵展平后，它就是向量的 ∞-范数，因此满足范数的前三条性质. 不过，本书定义的矩阵范数还要求相容性. 取

$$X=Y=\begin{pmatrix}1&1\\1&1\end{pmatrix},\qquad XY=\begin{pmatrix}2&2\\2&2\end{pmatrix}.$$

此时 $m(X)=m(Y)=1$，而 $m(XY)=2$，所以

$$m(XY)>m(X)m(Y).$$

矩阵乘法中的求和可能累积多个元素的贡献；只取单个元素的最大绝对值不能控制这种累积，因此它不满足本书矩阵范数定义中的相容性.

从向量模长或 2-范数推导得来的形式 $\left(\sum_{i=1}^{m}\sum_{k=1}^{n}x_{i,j}^2\right)^{1/2}$, 我们可以证明其满足相容性，证明过程见下文. 从 1-范数推导得来的形式 $\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{i,j}|$ 虽然也可证明满足相容性，但该范数应用并不广泛. 从 ∞-范数推导得来的形式 $\max_{1\leq i\leq m,1\leq j\leq n}|x_{i,j}|$ 则不满足相容性，因此不能构成一个矩阵范数. 我们着重介绍从向量模长或 2-范数推导得来的矩阵范数，我们一般称其为矩阵的模或矩阵的弗罗贝尼乌斯（Frobenius）范数：

定义 1.6.8 (矩阵的模/弗罗贝尼乌斯范数) 实数矩阵线性空间 $\mathbb{R}^{m\times n}$ 中的函数

$$\|X\|_F:=\sqrt{\langle X,X\rangle}=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_{i,j}^2\right)^{1/2}$$

满足矩阵范数的定义，称为矩阵的模或弗罗贝尼乌斯范数（*Frobenius norm*）.

下面证明弗罗贝尼乌斯范数满足相容性：

证明. 任意给定矩阵 $X\in\mathbb{R}^{m\times n},Y\in\mathbb{R}^{n\times q}$, 我们分别将其按行、列分解为

$$X=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_m)^\top,Y=(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_q),$$

则矩阵 $Z=XY$ 中的第 i,j 个元素为 $z_{i,j}=\boldsymbol{a}_i^\top\boldsymbol{b}_j$, 则

$$\begin{aligned}\|Z\|_F^2&=\sum_{i=1}^{m}\sum_{j=1}^{q}z_{i,j}^2=\sum_{i=1}^{m}\sum_{j=1}^{q}(\boldsymbol{a}_i^\top\boldsymbol{b}_j)^2\leq\sum_{i=1}^{m}\sum_{j=1}^{q}(\boldsymbol{a}_i^\top\boldsymbol{a}_i)(\boldsymbol{b}_j^\top\boldsymbol{b}_j)\\&=\sum_{i=1}^{m}(\boldsymbol{a}_i^\top\boldsymbol{a}_i)\sum_{j=1}^{q}(\boldsymbol{b}_j^\top\boldsymbol{b}_j)=\|X\|_F^2\|Y\|_F^2,\end{aligned}$$

Yuan-Hang Song

其中的小于号使用到了柯西-施瓦茨不等式. ■

除弗罗贝尼乌斯范数外，我们还可以考虑下述的一系列范数，我们称其为矩阵的算子范数（operator norm）：

定义 1.6.9 (p-算子范数) 实数矩阵线性空间 $\mathbb{R}^{m\times n}$ 中的函数

$$\|X\|_p := \max_{\boldsymbol{u}\in\mathbb{R}^n, \|\boldsymbol{u}\|_p=1} \|X\boldsymbol{u}\|_p,$$

（其中 $\|\boldsymbol{u}\|_p, \|X\boldsymbol{u}\|_p$ 分别为 $\mathbb{R}^n, \mathbb{R}^m$ 上的 p-范数）满足矩阵范数的定义，称为矩阵的 p-算子范数.

小贴士

所有算子范数均满足 $\|I\|_p = 1$, 其中 I 为单位阵. 而矩阵的模或者弗罗贝尼乌斯范数不满足上述性质，因此不是一个算子范数，但仍然是一个矩阵范数. 简单而言，弗罗贝尼乌斯范数刻画的是矩阵自身的度量性质，而算子范数则刻画的是矩阵作为线性映射时的度量性质.

我们下面简要证明 p-算子范数如何满足矩阵范数的四条性质.

证明. 首先，由于

$$\|X\|_p \geq \|X\boldsymbol{u}\|_p, \forall \boldsymbol{u}\in\mathbb{R}^n,$$

我们可以直接推导出正定性. 其次，由于

$$\|cX\|_p = \max_{\|\boldsymbol{u}\|_p=1} \|cX\boldsymbol{u}\|_p = |c| \max_{\|\boldsymbol{u}\|_p=1} \|X\boldsymbol{u}\|_p = |c|\|X\|_p,$$

齐次性得证. 进一步，由于

$$\begin{aligned}\|X+Y\|_p &= \max_{\|\boldsymbol{u}\|_p=1} \|(X+Y)\boldsymbol{u}\|_p \leq \max_{\|\boldsymbol{u}\|_p=1} (\|X\boldsymbol{u}\|_p + \|Y\boldsymbol{u}\|_p)\\ &\leq \max_{\|\boldsymbol{u}\|_p=1} \|X\boldsymbol{u}\|_p + \max_{\|\boldsymbol{v}\|_p=1} \|Y\boldsymbol{v}\|_p = \|X\|_p + \|Y\|_p,\end{aligned}$$

三角不等式得证. 最后，我们分两种情况讨论相容性. 当 $XY = 0$ 时，由正定性得相容性自然成立；当 $XY \neq 0$ 时，由正定性得 $\|XY\|_p \neq 0$, 记 $\boldsymbol{u}_0 = \arg\max_{\|\boldsymbol{u}\|_p=1} \|XY\boldsymbol{u}\|_p$, 则 $Y\boldsymbol{u}_0 \neq 0$, 从而 $Y\boldsymbol{u}_0/\|Y\boldsymbol{u}_0\|_p$ 为一个 p-范数为 1 的向量，我们有

$$\|XY\|_p = \|XY\boldsymbol{u}_0\|_p = \|Y\boldsymbol{u}_0\|_p \left\|X\frac{Y\boldsymbol{u}_0}{\|Y\boldsymbol{u}_0\|_p}\right\|_p \leq \|Y\boldsymbol{u}_0\|_p \max_{\|\boldsymbol{v}\|_p=1} \|X\boldsymbol{v}\|_p \leq \|Y\|_p\|X\|_p.$$

■

根据算子范数的定义，我们很容易得出其与对应的向量范数有如下关系：

性质 1.6.10 矩阵的 p-算子范数与其对应的向量 p-范数满足

$$\|X\boldsymbol{y}\|_p \leq \|X\|_p \|\boldsymbol{y}\|_p.$$

同样 p-算子范数中有三种范数使用最为广泛，即1-算子范数、2-算子范数和 ∞-算子范数. 我们先考虑 1-算子范数和 ∞-算子范数. 对于这两个范数，我们有如下性质：

性质 1.6.11 矩阵的 1-算子范数和 ∞-算子范数满足如下性质：

$$\|X\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |x_{i,j}|, \quad \|X\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |x_{i,j}|,$$

即 1-算子范数可以看作每列元素绝对值求和的最大值，∞-算子范数则可以看作每行元素绝对值求和的最大值. 因此，二者分别可以称作“列和范数”与“行和范数”.

证明. 对于 1-算子范数，我们将矩阵按列拆分为 $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, 则

$$\begin{aligned}\|X\|_1 &= \max_{\|\boldsymbol{u}\|_1=1} \|X\boldsymbol{u}\|_1 \leq \max_{\|\boldsymbol{u}\|_1=1} (\|u_1\boldsymbol{x}_1\|_1 + \cdots + \|u_n\boldsymbol{x}_n\|_1) \\ &= \max_{\|\boldsymbol{u}\|_1=1} (|u_1|\|\boldsymbol{x}_1\|_1 + \cdots + |u_n|\|\boldsymbol{x}_n\|_1) = \max_{1\leq j\leq n} \|\boldsymbol{x}_j\|_1 = \|X\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |x_{i,j}|,\end{aligned}$$

不等式取等的条件为 $\boldsymbol{u}$ 的第 j 个分量为 1, 其余分量为 0, 其中 j 为绝对值求和最大的那一列的索引.

对于 ∞-算子范数，我们将矩阵按行拆分为 $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^\top$, 则

$$\|X\|_\infty = \max_{\|\boldsymbol{u}\|_\infty=1} \|X\boldsymbol{u}\|_\infty = \max_{\|\boldsymbol{u}\|_\infty=1} \max_{1\leq i\leq m} |\boldsymbol{x}_i^\top \boldsymbol{u}| \leq \max_{1\leq i\leq m} \sum_{j=1}^{n} |x_{i,j}|,$$

不等式取等的条件为 $\boldsymbol{u} = \operatorname{sign}(\boldsymbol{x}_i)$, 其中 $\operatorname{sign}(\cdot)$ 为符号函数，分别将正数、负数、零映射为 $+1, -1, 0$, i 为绝对值求和最大的那一行的索引. ■

例 1.6.12 (列和范数、行和范数及达到上界的向量) 考虑

$$X = \begin{pmatrix} 1 & -2 & 0 \\ 3 & 1 & -4 \end{pmatrix} \in \mathbb{R}^{2\times 3}.$$

各列元素绝对值之和为 $4, 3, 4$，各行元素绝对值之和为 $3, 8$，因此

$$\|X\|_1 = 4, \qquad \|X\|_\infty = 8.$$

按弗罗贝尼乌斯范数的定义，还有

$$\|X\|_F = \sqrt{1^2 + (-2)^2 + 0^2 + 3^2 + 1^2 + (-4)^2} = \sqrt{31}.$$

为验证 1-算子范数确实能达到 4，取 $\boldsymbol{u} = (1, 0, 0)^\top$，则

$$\|\boldsymbol{u}\|_1 = 1, \qquad X\boldsymbol{u} = (1, 3)^\top, \qquad \|X\boldsymbol{u}\|_1 = 4.$$

Yuan and Song

为验证 ∞-算子范数能达到 8，按第二行各元素的符号取 $\boldsymbol{v}=(1,1,-1)^\top$，则

$$\|\boldsymbol{v}\|_\infty=1,\qquad X\boldsymbol{v}=(-1,8)^\top,\qquad \|X\boldsymbol{v}\|_\infty=8.$$

这个例子把列和、行和的计算公式与算子范数的最大放大倍数定义联系了起来.

我们最后考虑一下矩阵的 2-算子范数，它的定义式为 $\|X\|_2=\max_{\|\boldsymbol{u}\|=1}\|X\boldsymbol{u}\|$, 其中 $\|\cdot\|$ 为向量模长或 2-范数. 这个范数具有什么性质，又将如何去计算它呢？我们将会在下一小节讨论.

1.6.3 奇异值分解

在这一小节中，我们继续考虑矩阵的 2-算子范数. 在 2-算子范数的计算中，出现了 $\|X\boldsymbol{u}\|$ 的形式，我们进一步推导得到：

$$\|X\boldsymbol{u}\|=\sqrt{\langle X\boldsymbol{u},X\boldsymbol{u}\rangle}=\sqrt{\boldsymbol{u}^\top X^\top X\boldsymbol{u}}.$$

我们观察到上述计算结果中出现了 $X^\top X$ 的形式，这一形式值得我们深入去研究. 以下几条关于 $X^\top X$ 的性质帮助我们对上述形式建立层层深入的理解.

性质 1.6.13 (半正定性) 设 $X\in\mathbb{R}^{m\times n}$ 为任意矩阵，则矩阵 $X^\top X$ 与 $XX^\top$ 均为半正定矩阵.

证明. 欲证 $X^\top X$ 为半正定矩阵，由谱定理1.4.3，只需证其所有特征值均为非负数即可. 假设其存在一个负特征值 λ_-, 设其对应的一个特征向量为 $\boldsymbol{q}\in\mathbb{R}^n$, 则有 $X^\top X\boldsymbol{q}=\lambda_-\boldsymbol{q}$. 等式左右同乘 $\boldsymbol{q}^\top$ 得：

$$\boldsymbol{q}^\top X^\top X\boldsymbol{q}=(X\boldsymbol{q})^\top(X\boldsymbol{q})=\|X\boldsymbol{q}\|^2\geq 0,$$

而

$$\lambda_-\boldsymbol{q}^\top\boldsymbol{q}=\lambda_-\|\boldsymbol{q}\|^2<0,$$

矛盾！因此 $X^\top X$ 的所有特征值必然均为非负数，从而 $X^\top X$ 为半正定矩阵. 将 X 替换为 $X^\top$ 进行相同步骤，可得到 $XX^\top$ 也为半正定矩阵. ■

性质 1.6.14 (秩相同) 设 $X\in\mathbb{R}^{m\times n}$ 为任意矩阵，则矩阵 X, $X^\top X$ 与 $XX^\top$ 拥有相同的秩.

证明. 由于 X 与 $X^\top X$ 的列数相同，欲证明二者有相同的秩，只需证明二者构成的齐次线性方程组具有相同的解即可. 正向来看，若 $X\boldsymbol{y}=0$, 则必有 $X^\top X\boldsymbol{y}=0$. 反向来看，

Yuan and Song

若 $X^\top X\boldsymbol{y}=0$, 则等式两侧左乘 $\boldsymbol{y}^\top$ 得：

$$0=\boldsymbol{y}^\top X^\top X\boldsymbol{y}=\|X\boldsymbol{y}\|^2,$$

由模长的正定性得 $X\boldsymbol{y}=0$. 从而二者的解空间相同，因此秩也相同.

同样地，我们可以推导得出 $X^\top$ 与 $XX^\top$ 的秩相同. 又由于 $X, X^\top$ 的秩自然相同，我们可以得出 $X, X^\top X$ 与 $XX^\top$ 三者拥有相同的秩. ■

性质 1.6.15 (特征值相同) 设 $X\in\mathbb{R}^{m\times n}$ 为任意矩阵，则矩阵 $X^\top X$ 与 $XX^\top$ 有相同的特征值，且每个非零特征值的重数也相同.

证明. **第一步：证明二者有相同的特征值**. 假设矩阵 $X^\top X$ 有一个特征值 λ 及其对应的特征向量 $\boldsymbol{p}$ 满足 $X^\top X\boldsymbol{p}=\lambda\boldsymbol{p}$, 两边左乘 X 得：

$$XX^\top(X\boldsymbol{p})=\lambda(X\boldsymbol{p}).$$

从而矩阵 $XX^\top$ 也有一个特征值为 λ, 且有一个对应的特征向量为 $X\boldsymbol{p}$.

第二步：证明非零特征值的重数相同. 由于 $X^\top X, XX^\top$ 均为实对称矩阵，根据谱定理1.4.3, 二者均可对角化，从而它们的每个特征值的几何重数与代数重数是一样的. 我们证明它们的几何重数相同即可.

下面考虑 $X^\top X, XX^\top$ 共有的非零特征向量 λ, 由于属于同一个特征值的所有特征向量构成了一个特征子空间，我们只需找到一个二者之间的**双射**，即可证明两个空间的维度相同，即几何重数相同. 事实上，上文的证明已经为我们提供了一个可能的映射：

$$f:\mathcal{V}\to\mathcal{U};\quad f(\boldsymbol{v})=X\boldsymbol{v},$$

其中 $\mathcal{V},\mathcal{U}$ 分别为 $X^\top X, XX^\top$ 的属于特征值 λ 的特征子空间.

单射：对于 $\mathcal{V}$ 中的两个向量 $\boldsymbol{v}_1,\boldsymbol{v}_2$, 若 $f(\boldsymbol{v}_1)=f(\boldsymbol{v}_2)$, 则 $X(\boldsymbol{v}_1-\boldsymbol{v}_2)=\boldsymbol{0}$. 两边左乘 $X^\top$ 得：

$$0=X^\top\cdot\boldsymbol{0}=X^\top X(\boldsymbol{v}_1-\boldsymbol{v}_2)=\lambda(\boldsymbol{v}_1-\boldsymbol{v}_2).$$

由于 $\lambda\neq0$, 因此必有 $\boldsymbol{v}_1=\boldsymbol{v}_2$, 单射得证.

满射：对于任意 $\boldsymbol{u}\in\mathcal{U}$, 有 $XX^\top\boldsymbol{u}=\lambda\boldsymbol{u}$, 由于 $\lambda\neq0$, 我们有

$$\boldsymbol{u}=X(\lambda^{-1}X^\top\boldsymbol{u}).$$

令 $\boldsymbol{v}=\lambda^{-1}X^\top\boldsymbol{u}$, 即有 $\boldsymbol{u}=X\boldsymbol{v}$, 我们下面验证 $\boldsymbol{v}\in\mathcal{V}$:

$$X^\top X\boldsymbol{v}=X^\top X(\lambda^{-1}X^\top\boldsymbol{u})=\lambda^{-1}X^\top(XX^\top\boldsymbol{u})=\lambda^{-1}X^\top\lambda\boldsymbol{u}=\lambda\boldsymbol{v}.$$

从而 $\boldsymbol{v}\in\mathcal{V}$, 满射得证.

Yuan and Song

因此，映射 f 是一个双射，且上文已经为我们找到了它的逆映射 $f^{-1}(\boldsymbol{u}) = \lambda^{-1}X^\top \boldsymbol{u}$. 综上，$X^\top X, XX^\top$ 拥有相同的特征值，且每个非零特征值的重数也相同. ■

根据以上性质，我们有如下认识：矩阵 $X^\top X, XX^\top$ 可能大小并不相等，但它们的非零特征值（包括重数）是一一对应的，唯一区别在于形状较大的那个矩阵可能存在额外的更多零特征值. 而以上所有非零特征值（包括重数）的总个数，就等于矩阵 X 的秩. 进一步，又由于半正定性，上述所有的特征值能够取算术平方根. 这些特征值的算术平方根能够反映矩阵 X 的若干性质，我们称这些值为矩阵 X 的奇异值（singular values）：

定义 1.6.16 (奇异值) *设矩阵 $X \in \mathbb{R}^{m\times n}$, 由它构成的两个对称阵 $X^\top X, XX^\top$ 的所有特征值构成的集合为 $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$（不计重数），则 $\sigma_1 := \sqrt{\lambda_1}, \sigma_2 := \sqrt{\lambda_2}, \ldots, \sigma_k := \sqrt{\lambda_k}$ 称为矩阵 X 的奇异值.*

我们知道对于可对角化的矩阵 A，它可以分解为与特征值构成的对角矩阵相似的形式 $A = PDP^{-1}$, 而对于一般的矩阵 X 的奇异值，是否也存在类似的形式呢？事实上我们有如下的奇异值分解定理：

定理 1.6.17 (奇异值分解) *设矩阵 $X \in \mathbb{R}^{m\times n}$, 则存在正交矩阵 $U \in \mathbb{R}^{m\times m}$ 与正交矩阵 $V \in \mathbb{R}^{n\times n}$，使得*

$$X = U\Sigma V^\top,$$

其中 $\Sigma \in \mathbb{R}^{m\times n}$ 中的对角元为 X 的所有奇异值，非对角元为 0, U 中的每一列向量为 X 的一个左奇异向量，V 中的每一列向量为 X 的一个右奇异向量.

证明. 设矩阵 X 的所有奇异值不计重数按降序排列为 $\sigma_1, \ldots, \sigma_k, 0$. 我们先考虑 σ_1, 设 σ_1 的重数为 r_1. 由于 σ_1^2 也是 $X^\top X$ 的 r_1 重特征值，且根据谱定理1.4.3, 存在 r_1 个相互正交且模长为 1 的单位向量 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r_1} \in \mathbb{R}^n$ 张成 σ_1^2 的特征子空间. 进一步，根据性质1.6.15的证明，我们可以使用双射 $f(\boldsymbol{v}) = X\boldsymbol{v}$ 得到 $XX^\top$ 的属于 σ_1^2 的特征向量 $\boldsymbol{u}_1' = X\boldsymbol{v}_1, \ldots, \boldsymbol{u}_{r_1}' = X\boldsymbol{v}_{r_1}$, 它们是相互正交的，即：

$$(\boldsymbol{u}_i')^\top \boldsymbol{u}_j' = \boldsymbol{v}_i^\top X^\top X\boldsymbol{v}_j = \sigma_1^2 \boldsymbol{v}_i^\top \boldsymbol{v}_j = 0, \quad \forall i \neq j.$$

但每个向量并不是单位向量. 由于

$$(\boldsymbol{u}_i')^\top \boldsymbol{u}_i' = \boldsymbol{v}_i^\top X^\top X\boldsymbol{v}_i = \sigma_1^2 \boldsymbol{v}_i^\top \boldsymbol{v}_i = \sigma_1^2,$$

因此取 $\boldsymbol{u}_i = \sigma_1^{-1}\boldsymbol{u}_i'$ 后，$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_1}$ 就是彼此正交且模长为 1 的单位向量了.

记 $U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_1}) \in \mathbb{R}^{m\times r_1}, V_1 = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r_1}) \in \mathbb{R}^{n\times r_1}$, 则我们有 $U_1 =$

$\sigma_1^{-1}XV_1$. 我们下面考虑矩阵 $X_{-1} := X - \sigma_1 U_1 V_1^\top$, 我们有

$$\begin{aligned} X_{-1}^\top X_{-1} &= X^\top X - \sigma_1 X^\top U_1 V_1^\top - \sigma_1 V_1 U_1^\top X + \sigma_1^2 V_1 U_1^\top U_1 V_1^\top \\ &= X^\top X - X^\top X V_1 V_1^\top - V_1 V_1^\top X^\top X + \sigma_1^2 V_1 V_1^\top \\ &= X^\top X - \sigma_1^2 V_1 V_1^\top. \end{aligned}$$

根据 $X^\top X$ 的特征值分解的结果，$X_{-1}^\top X_{-1}$ 实现的效果就是**将 $X^\top X$ 的特征值 σ_1^2 变为了 0，特征向量不变，而剩余的特征值和特征向量均不变**.

因此，我们将矩阵 X 分解为了 $\sigma_1 U_1 V_1^\top + X_{-1}$, 而矩阵 X_{-1} 的奇异值为 $\sigma_2, \ldots, \sigma_k, 0$（原先的 σ_1 变为了 0）. 上述过程可以循环进行，直至将所有正特征值均分解出来，得到

$$X = \sigma_1 U_1 V_1^\top + \cdots + \sigma_k U_k V_k^\top + X_{-k},$$

而余项 X_{-k} 的所有奇异值均为 0，即 $X_{-k}^\top X_{-k}$ 相似于零矩阵，秩为 0，因此 X_{-k} 秩也为 0，从而为零矩阵. 因此我们有

$$X = \sigma_1 U_1 V_1^\top + \cdots + \sigma_k U_k V_k^\top.$$

这一形式一般称为矩阵 X 的**瘦奇异值分解**. 进一步，由于 X 的秩 $r \leq \min\{m, n\}$, 我们可以分别将 $(U_1, \ldots, U_k), (V_1, \ldots, V_k)$ 通过施密特正交化进一步扩充为正交矩阵 $U = (U_1, \ldots, U_k, U_-), V = (V_1, \ldots, V_k, V_-)$, 得到 X 的最终奇异值分解为：

$$X = (U_1, \ldots, U_k, U_-) \begin{pmatrix} \sigma_1 I_{r_1} & & & \\ & \ddots & & \\ & & \sigma_k I_{r_k} & \\ & & & 0_{m-r, n-r} \end{pmatrix} \begin{pmatrix} V_1^\top \\ \vdots \\ V_k^\top \\ V_-^\top \end{pmatrix} := U \Sigma V^\top.$$

■

小贴士

由于 $U^\top X X^\top U = \Sigma\Sigma^\top, V^\top X^\top X V = \Sigma^\top \Sigma$, 因此 U, V 中的每一列向量分别为矩阵 $XX^\top, X^\top X$ 的特征向量.

例 1.6.18 (一个非方阵的完整奇异值分解) *考虑*

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad X^\top X = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

Yuan and Song

后一个矩阵的特征值为 $4, 2$，对应的单位特征向量为

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{2}}(1,1)^\top, \qquad \boldsymbol{v}_2 = \frac{1}{\sqrt{2}}(1,-1)^\top.$$

因此 X 的非零奇异值为 $\sigma_1 = 2, \sigma_2 = \sqrt{2}$. 按 $\boldsymbol{u}_i = X\boldsymbol{v}_i/\sigma_i$ 构造左奇异向量，得到

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{2}}(1,1,0)^\top, \qquad \boldsymbol{u}_2 = (0,0,1)^\top.$$

再补入 $\boldsymbol{u}_3 = (1,-1,0)^\top/\sqrt{2}$，三者构成 $\mathbb{R}^3$ 的单位正交基. 于是可取

$$U = \begin{pmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}, \qquad V = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}, \qquad X = U\Sigma V^\top.$$

也可以直接验证较紧凑的形式

$$X = 2\boldsymbol{u}_1\boldsymbol{v}_1^\top + \sqrt{2}\boldsymbol{u}_2\boldsymbol{v}_2^\top.$$

对任意输入 $\boldsymbol{x} = a\boldsymbol{v}_1 + b\boldsymbol{v}_2$，输出为

$$X\boldsymbol{x} = 2a\boldsymbol{u}_1 + \sqrt{2}b\boldsymbol{u}_2.$$

输入空间与输出空间各使用一组单位正交基，奇异值描述对应方向间的缩放. Σ 的第三行为零，表示输出不会含有 $\boldsymbol{u}_3$ 方向的分量.

自此，拥有了奇异值分解这一强大的工具，我们可以完美回答本小节最开始提出的问题了——2-算子范数的值具体如何计算. 由 $X = U\Sigma V^\top$, 对于任意 $\boldsymbol{u}$ 满足 $\|\boldsymbol{u}\| = 1$, 我们有：

$$\boldsymbol{u}X^\top X\boldsymbol{u} = \boldsymbol{u}^\top V\Sigma^\top U^\top U\Sigma V^\top\boldsymbol{u} = \boldsymbol{u}^\top V\Sigma^\top\Sigma V^\top\boldsymbol{u},$$

令 $\boldsymbol{w} = V^\top\boldsymbol{u}$, 则 $\boldsymbol{w}^\top\boldsymbol{w} = \boldsymbol{u}^\top VV^\top\boldsymbol{u} = \boldsymbol{u}^\top\boldsymbol{u} = I$, 则

$$\boldsymbol{u}^\top V\Sigma^\top\Sigma V^\top\boldsymbol{u} = \boldsymbol{w}^\top\Sigma^\top\Sigma\boldsymbol{w} = \sum_{i=1}^{n}\sigma_i^2 w_i^2 \le \sigma_*^2\sum_{i=1}^{n} w_i^2 = \sigma_*^2,$$

其中 σ_i 为矩阵 Σ 的所有对角元，即计重数的奇异值，σ_* 为最大的奇异值. 上述不等式取等的条件为 σ_* 对应的分量 w_i 取 1, 其余分量取 0. 因此 $\|X\boldsymbol{u}\| = \sqrt{\boldsymbol{u}^\top X^\top X\boldsymbol{u}} \le \sigma_*$, 从而我们得出结论：**矩阵的 2-算子范数等于矩阵的最大奇异值**. 鉴于矩阵 2-算子范数与实对称矩阵的谱定理有很大的关系，2-算子范数也被称为谱范数（spectral norm）.

最后，我们还可以使用奇异值分解得到弗罗贝尼乌斯范数与奇异值之间的关系如

Yuan and Song

下：

$$\|X\|_F = \sqrt{\mathrm{tr}(X^\top X)} = \sqrt{\mathrm{tr}(\Sigma^\top \Sigma)} = \sqrt{\sum_{i=1}^n \sigma_i^2},$$

即**矩阵的弗罗贝尼乌斯范数等于矩阵所有奇异值的平方和的算术平方根**, 从而我们有关系 $\|X\|_2 \le \|X\|_F$, $\forall X \in \mathbb{R}^{m\times n}$.

例 1.6.19 (特征值全为零仍然可能放大向量) 考虑方阵

$$X = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix}.$$

其特征多项式为 $|\lambda I - X| = \lambda^2$，所以两个特征值均为零. 但是

$$X^\top X = \begin{pmatrix} 0 & 0 \\ 0 & 9 \end{pmatrix},$$

因而奇异值为 $3, 0$，根据本小节刚得到的公式，

$$\|X\|_2 = 3, \qquad \|X\|_F = \sqrt{3^2 + 0^2} = 3.$$

这一放大倍数可以直接达到：$\boldsymbol{e}_2 = (0,1)^\top$ 的长度为 1，而

$$X\boldsymbol{e}_2 = (3,0)^\top = 3\boldsymbol{e}_1, \qquad \|X\boldsymbol{e}_2\|_2 = 3.$$

更一般地，对 $\boldsymbol{x} = (x_1, x_2)^\top$，有 $\|X\boldsymbol{x}\|_2 = 3|x_2| \le 3\|\boldsymbol{x}\|_2$. 特征值描述保持自身方向的向量如何变化；对于一般矩阵，最大长度放大倍数由最大奇异值给出，不能直接用特征值的最大绝对值代替.

1.6.4* 对偶范数

对于向量的模长或是 2-范数，我们有柯西-施瓦茨不等式 $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \le \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$, 而对于其他的范数，是否也有类似的不等式呢？这就引出了对偶范数（dual norm）的概念：

定义 1.6.20 (对偶范数) 对于实数向量空间 $\mathbb{R}^n$ 上的一个范数 $\|\cdot\|_\cdot$, 定义其对偶范数为：

$$\|\boldsymbol{y}\|_* := \max_{\|\boldsymbol{x}\|_\cdot = 1} \langle \boldsymbol{x}, \boldsymbol{y}\rangle.$$

我们依然验证一下其是否满足范数的三条性质. 对于正定性，只要 $\boldsymbol{y} \neq 0$, 则

$$\max_{\|\boldsymbol{x}\|_\cdot = 1} \langle \boldsymbol{x}, \boldsymbol{y}\rangle \ge \|\boldsymbol{y}\|_\cdot^{-1} \langle \boldsymbol{y}, \boldsymbol{y}\rangle = \|\boldsymbol{y}\|_\cdot^{-1} \|\boldsymbol{y}\|_2^2 > 0.$$

对于三角不等式，我们有

$$\|\boldsymbol{y}_1\|_* + \|\boldsymbol{y}_2\|_* = \max_{\|\boldsymbol{x}_1\|_\cdot = 1} \langle \boldsymbol{x}_1, \boldsymbol{y}_1\rangle + \max_{\|\boldsymbol{x}_2\|_\cdot = 1} \langle \boldsymbol{x}_2, \boldsymbol{y}_2\rangle \ge \max_{\|\boldsymbol{x}\|_\cdot = 1} \langle \boldsymbol{x}, \boldsymbol{y}_1 + \boldsymbol{y}_2\rangle = \|\boldsymbol{y}_1 + \boldsymbol{y}_2\|_*.$$

Yuanzhi Song

对于齐次性，其几何直观是较为显然的，但是严格证明却稍显复杂，我们单独开辟如下的篇幅来证明齐次性：

证明. 我们分三种情况来讨论. 当 $c=0$ 时，齐次性自然成立.

当 $c>0$ 时，

$$\|c\boldsymbol{y}\|_*=\max_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},c\boldsymbol{y}\rangle=c\max_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},\boldsymbol{y}\rangle=c\|\boldsymbol{y}\|_*.$$

当 $c<0$ 时，

$$\|c\boldsymbol{y}\|_*=\max_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},c\boldsymbol{y}\rangle=c\min_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},\boldsymbol{y}\rangle.$$

我们注意到，对于任意范数 $\|\cdot\|_\cdot$，集合 $U:=\{\boldsymbol{x}|\|\boldsymbol{x}\|_\cdot=1\}$ 是一个中心对称集合，即有 $\boldsymbol{x}\in U\Leftrightarrow-\boldsymbol{x}\in U$. 因此其次线性函数 $f(\boldsymbol{x})=\langle\boldsymbol{x},\boldsymbol{y}\rangle$ 在 U 上的值域也是关于 0 对称的，因此我们有 $\min_{\boldsymbol{x}\in U}\langle\boldsymbol{x},\boldsymbol{y}\rangle=-\max_{\boldsymbol{x}\in U}\langle\boldsymbol{x},\boldsymbol{y}\rangle$，从而原式为

$$c\min_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},\boldsymbol{y}\rangle=-c\max_{\|\boldsymbol{x}\|_\cdot=1}\langle\boldsymbol{x},\boldsymbol{y}\rangle=|c|\|\boldsymbol{y}\|_*.$$

综上所述，齐次性得证. ■

根据对偶范数的定义，我们很容易证明原范数 $\|\cdot\|_\cdot$ 与其对偶范数 $\|\cdot\|_*$ 满足一个**广义的柯西-施瓦茨不等式**：

$$|\langle\boldsymbol{x},\boldsymbol{y}\rangle|\leq\|\boldsymbol{x}\|_\cdot\|\boldsymbol{y}\|_*,\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n.$$

证明. 当 $\boldsymbol{x}=\boldsymbol{0}$ 时，不等式自然成立. 当 $\boldsymbol{x}\neq0$ 时，有

$$\|\boldsymbol{x}\|_\cdot\|\boldsymbol{y}\|_*=\|\boldsymbol{x}\|_\cdot\max_{\|\boldsymbol{z}\|_\cdot=1}\langle\boldsymbol{z},\boldsymbol{y}\rangle\geq\|\boldsymbol{x}\|_\cdot\left\langle\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_\cdot},\boldsymbol{y}\right\rangle=\langle\boldsymbol{x},\boldsymbol{y}\rangle.$$

而对于不等式另一边，我们用反证法更方便说明问题. 假设 $\langle\boldsymbol{x},\boldsymbol{y}\rangle<-\|\boldsymbol{x}\|_\cdot\|\boldsymbol{y}\|_*$，则我们立即有

$$\langle-\boldsymbol{x},\boldsymbol{y}\rangle>\|\boldsymbol{x}\|_\cdot\|\boldsymbol{y}\|_*=\|-\boldsymbol{x}\|_\cdot\|\boldsymbol{y}\|_*.$$

与上述的证明矛盾，因此不等式整体得证. ■

直接利用普通的柯西-施瓦茨不等式，我们可以得出结论：向量的 2-范数的对偶范数就是其自身，即**2-范数是自对偶的**. 这也进一步表明了 2-范数在欧氏空间中的特殊地位，所以我们才将其视为欧氏空间中的默认的模长.

我们下面考虑其他范数的对偶范数.

例 1.6.21 考虑向量的 1-范数，则它的对偶范数可表示为

$$\|\boldsymbol{y}\|_*=\max_{\|\boldsymbol{x}\|_1=1}=\langle\boldsymbol{x},\boldsymbol{y}\rangle=\max_{\|\boldsymbol{x}\|_1=1}(x_1y_1+\cdots+x_ny_n).$$

Yuan and Song

对于任何一个可能的 $\boldsymbol{x}=(x_1,\ldots,x_n)$ 的取值，将其中若干个分量取负号，共可得到 2^n 种取值，而相应的与 $(y_1,\ldots,y_n)$ 每个分量的符号相同的那个取值将会使得上式在 2^n 个取值中取得最大值. 进一步，我们让 $\boldsymbol{y}$ 中绝对值最大的那个分量 y_i 对应的 x_i 取 $+1$ 或 -1（取决于 y_i 的符号），此时整个式子能够达到最大值 $|y_i|$.

因此，我们有结论，对于 1-范数，其对偶范数就是 ∞-范数. 此外，我们可以类似地推导出，∞-范数的对偶范数就是 1-范数，我们将其留作习题.

小贴士

范数 $\|\cdot\|$. 对偶范数 $\|\cdot\|_*$ 的对偶范数 $\|\cdot\|_{**}$ 就是其自身，即 $\|\boldsymbol{x}\|_{**}=\|\boldsymbol{x}\|.,\ \forall \boldsymbol{x}\in\mathbb{R}^n$.

例 1.6.22 (用对偶范数计算输入扰动的最大影响) 设 $f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}$，其中 $\boldsymbol{w}=(2,-3,1)^\top$. 若输入增加扰动 $\boldsymbol{\delta}$，则输出变化恰为

$$f(\boldsymbol{x}+\boldsymbol{\delta})-f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{\delta}.$$

给定 $\varepsilon>0$. 当约束为 $\|\boldsymbol{\delta}\|_\infty\le\varepsilon$ 时，利用 ∞-范数与 1-范数的对偶关系，有

$$|\boldsymbol{w}^\top\boldsymbol{\delta}|\le\|\boldsymbol{w}\|_1\|\boldsymbol{\delta}\|_\infty\le(2+3+1)\varepsilon=6\varepsilon.$$

取 $\boldsymbol{\delta}=\varepsilon(1,-1,1)^\top$ 便达到上界. 若改为 $\|\boldsymbol{\delta}\|_1\le\varepsilon$，则

$$|\boldsymbol{w}^\top\boldsymbol{\delta}|\le\|\boldsymbol{w}\|_\infty\varepsilon=3\varepsilon,$$

在 $\boldsymbol{\delta}=(0,-\varepsilon,0)^\top$ 处取等. 若改为 $\|\boldsymbol{\delta}\|_2\le\varepsilon$，则最大绝对变化为 $\sqrt{14}\varepsilon$，在 $\boldsymbol{\delta}=\varepsilon\boldsymbol{w}/\sqrt{14}$ 处达到. 三种约束分别给出 $6\varepsilon,3\varepsilon,\sqrt{14}\varepsilon$. 对偶范数不仅给出了变化的上界，还帮助我们找到达到上界的扰动方向.

我们下面来考虑矩阵范数的对偶范数，我们知道，矩阵范数相比于向量范数需要额外满足相容性，即 $\|XY\|.\le\|X\|.\|Y\|.$. 但是，仅使用对偶范数的定义，我们不能直接得到关于对偶范数的相容性，即 $\|XY\|_*\le\|X\|_*\|Y\|_*$ 不一定总能成立. 也就是说，一个良定义的矩阵范数的对偶范数可能只能作为一个展平为向量后的向量范数. 不过对于某些具体的实例，我们依然能够得到其良定义的对偶矩阵范数.

首先，由于弗罗贝尼乌斯范数就是向量模长在矩阵上的自然推广，因此弗罗贝尼乌斯范数就是矩阵上的自对偶的范数，自然一定满足相容性. 此外，还有如下的一些例子，它们也同样是良定义的对偶矩阵范数.

Yuan and Song

例 1.6.23 考虑矩阵的 1-算子范数 $\|X\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^m |x_{i,j}|$, 我们有

$$\|Y\|_* = \max_{\|X\|_1=1} \langle X, Y\rangle.$$

我们 X, Y 均按列拆分为 $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), Y = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, 则上述对偶范数的表达式变为：

$$\|Y\|_* = \max_{\|\boldsymbol{x}_1\|_1,\dots,\|\boldsymbol{x}_n\|_1\leqslant 1} \sum_{j=1}^n \boldsymbol{x}_j^\top \boldsymbol{y}_j = \sum_{j=1}^n \max_{\|\boldsymbol{x}_j\|_1\leqslant 1} \boldsymbol{x}_j^\top \boldsymbol{y}_j.$$

仿照例*1.6.21*中相同的步骤，我们可以得到上式为

$$\|Y\|_* = \sum_{j=1}^n \max_{1\leqslant i\leqslant m} |y_{i,j}|.$$

注意它并不等于 $\|Y\|_\infty$, 而是一个可能比其更大的值. 可以证明这个范数满足相容性，我们将其证明留给读者.

同理可以得出，∞-算子范数的对偶范数是

$$\|Y\|_* = \sum_{i=1}^m \max_{1\leqslant j\leqslant n} |y_{i,j}|,$$

它是一个取值可能比 1-算子范数稍大一些的范数，它同样也满足相容性. 而且利用单位矩阵不难验证，这两个对偶范数都不是算子范数.

例 1.6.24 考虑矩阵的 2-算子范数 $\|X\|_2 = \max_{i=1,\dots,\min\{m,n\}} \sigma_i(X)$, 其中 $\sigma_i(X)$ 表示矩阵 $X\in\mathbb{R}^{m\times n}$ 的所有奇异值. 我们有：

$$\|Y\|_* = \max_{\|X\|_2=1} \langle X, Y\rangle.$$

我们对矩阵 Y 做奇异值分解，得到 $Y = U\Sigma V^\top$. 根据矩阵内积的定义，我们有：

$$\langle X, Y\rangle = \langle X, U\Sigma V^\top\rangle = \mathrm{tr}(X^\top U\Sigma V^\top) = \mathrm{tr}(V^\top X^\top U\Sigma) = \langle U^\top XV, \Sigma\rangle.$$

如果 $\|X\|_2 = 1$, 即矩阵 X 的最大奇异值为 1，那么矩阵 $U^\top XV$ 的最大奇异值也是 1. 这是因为，我们可以将 X 进一步奇异值分解为 $U^\top XV = U^\top U'\Sigma'(V')^\top V$, 其中 Σ' 的对角元包含 X 的所有奇异值，而由于 $U^\top U'$ 与 $(V')^\top V$ 仍然也是正交矩阵，因此上式自然是矩阵 $U^\top XV$ 的奇异值分解，故其奇异值与 X 的奇异值完全相同.

下面我们继续断言矩阵 $U^\top XV$ 的每个元素绝对值不大于 1. 这是因为，假如该矩阵有一个绝对值大于 *1* 的元素 a, 不失一般性，我们假设第一行第一列的元素为 a. 这时我们选一个输入向量为 $\boldsymbol{e}_1 = (1, 0, \dots, 0)^\top$, 满足 $\|\boldsymbol{e}_1\|_2 = 1$, 我们有：

$$\begin{pmatrix} a & \cdots \\ \vdots & \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \end{pmatrix} = \begin{pmatrix} a \\ \vdots \end{pmatrix}.$$

此时输出的向量 $(a, \dots)^\top$ 的 2-范数一定至少为 $|a| > 1$, 这与 $\|X\|_2 = \max_{\|\boldsymbol{u}\|_2=1}\|X\boldsymbol{u}\|_2$

$=1$ 矛盾.

进一步，由于 Σ 只有对角元素非零，其余元素均为 0, 原先的内积即为两个矩阵的对角元对应相乘后相加. 由于 $U^\top XV$ 每个元素绝对值不大于 1, 自然对角元均取 1 时能够使内积最大，而令非对角元取 0, 我们此时可以得到一个符合 $\|X\|_2=1$ 的 X 取值，记 $s=\min\{m,n\}$, 则 $X=U_{[:,1:s]}I_sV_{[:,1:s]}^\top$. 其中 $[:,1:s]$ 表示取所有行，1 至 s 列得到的子矩阵.

当 X 取上述值时，最大值为 Σ 的所有对角元值的和，即 Y 的所有奇异值之和. 我们称矩阵 2-算子范数的对偶范数为核范数（*nuclear norm*），它的值等于矩阵所有奇异值之和，即：

$$\|X\|_*:=\sum_{i=1}^{\min\{m,n\}}\sigma_i(X).$$

小贴士

在向量范数中，2-范数为自对偶范数，其取值为所有分量的平方和的算术平方根，1-范数与 ∞-范数互为对偶范数，它们取值分别为所有元素的绝对值之和以及绝对值最大的元素. 而在矩阵范数中，Frobenius 范数为自对偶范数，其取值等于所有奇异值的平方和的算术平方根，核范数与谱范数互为对偶范数，他们的取值分别为所有奇异值之和与最大奇异值. 范数理论对于向量和矩阵两个对象的结论具有极高的相似性.

例 1.6.25 (三种矩阵范数与对偶上界的取等) 对矩阵

$$Y=\begin{pmatrix}3&0&0\\0&1&0\end{pmatrix},$$

有 $YY^\top=\mathrm{Diag}(9,1)$，所以两个非零奇异值为 $3,1$. 于是

$$\|Y\|_2=3,\qquad \|Y\|_F=\sqrt{10},\qquad \|Y\|_*=4.$$

三者分别计算奇异值的最大值、平方和的平方根、以及总和.

从对偶范数的角度看，取

$$X_0=\begin{pmatrix}1&0&0\\0&1&0\end{pmatrix},$$

则 $\|X_0\|_2=1$ 且 $\langle X_0,Y\rangle=3+1=4=\|Y\|_*$，恰好达到谱范数单位球上内积的最大值.

若约束改为 $\|X\|_F \le 1$，由弗罗贝尼乌斯范数的自对偶性，上界变为 $\sqrt{10}$；此时可取

$$X_1 = \frac{Y}{\sqrt{10}}, \qquad \|X_1\|_F = 1, \qquad \langle X_1, Y\rangle = \frac{10}{\sqrt{10}} = \sqrt{10}.$$

注意 $\|X_0\|_F = \sqrt{2} > 1$，所以前一个取等矩阵并不满足后一个约束. 不同的范数单位球会给出不同的最大内积.

1.7　小结

的

习题 1

1. 这是习题.

Yuan and So

第二章　概率与统计

2.1　事件及其概率

生活中充满了随机事件，对于随机事件发生的可能性大小进行度量就产生了概率. 从纯数学角度来讲，概率可以通过公理化体系来定义；而从实际意义角度，概率通常认为是在大量且可重复的独立实验中，某事件发生的频率的极限. 在本节我们将要学习事件的定义、概率及其基本的性质.

2.1.1 样本空间与事件

想要研究一个随机对象（譬如骰子），首先我们需要进行试验，得到随机对象的所有可能的确定的结果（即将骰子掷出去）. 每一次化随机为某个确定的结果的试验过程就称为一次**采样**（sample）. 该随机对象的每一个可能的结果就称为一个**样本点**（sample point），所有样本点构成了该随机对象的**样本空间**（sample space），通常记为 Ω. 样本空间中的元素可以是有限的，如掷硬币的 $\{\text{正}, \text{反}\}$，掷骰子的 $\{1,2,3,4,5,6\}$, 也可以是无限的，如 $(0,1]$(单位 m) 的长度值等.

有了样本空间，我们可以定义事件（events），事件即定义为**样本空间中的一个子集**，通常用大写字母 $A, B\ldots$ 来表示. 如“掷骰子数值 ≤ 3”，“掷硬币为正面”均可视为事件.

由于事件本质为上为样本空间的集合，我们可以有事件的包含关系：$A \subset B$ 即表示所有属于 A 的样本点也属于 B, 或者可以理解为事件 A 发生，则事件 B 一定发生. 事件 A 也可以取补集，即样本空间中所有不属于 A 的样本点构成的集合，我们称这一事件为原事件 A 的**对立事件**（complementary events）或**逆事件**，记作 $\bar{A}$. 事件之间可以进行取并集运算 $A \cup B$, 表示两个事件至少有一个发生；也可以进行取交集运算，$A \cap B$, 表示两个事件同时发生，$A \cap B$ 通常简写为 AB.

Yuan and Song

特别地，当 $A\cap B=\varnothing$ 时，我们称事件 A,B 为**互斥事件**或**互不相容事件**（exclusive events）. 事件 A 与其对立事件 $\bar{A}$ 一定是互斥事件. 如果事件 A,B 互斥，此时我们称 $A\cup B$ 为两个事件的和，可以改记为 $A+B$.

事件的运算满足如下的定理，它们本质上都是通过集合理论推导得出的结论：

1. 交换律：$A\cup B=B\cup A,\ AB=BA.$
2. 结合律：$(A\cup B)\cup C=A\cup(B\cup C),\ (AB)C=A(BC).$
3. 分配律：$(A\cup B)C=(AC)\cup(BC),\ (AB)\cup C=(A\cup C)(B\cup C).$
4. 德摩根（De Morgan）律：$\overline{A\cup B}=\bar{A}\bar{B},\ \overline{AB}=\bar{A}\cup\bar{B}.$

例 2.1.1 (把日常语言写成事件) 掷一次骰子，样本空间为 $\Omega=\{1,2,3,4,5,6\}$. 设 $A=$“点数为偶数”，$B=$“点数不超过 *3*”，则

$$A=\{2,4,6\},\quad B=\{1,2,3\},\quad AB=\{2\},\quad A\cup B=\{1,2,3,4,6\}.$$

“两个条件都不满足”对应 $\overline{A\cup B}=\{5\}$；“两个条件没有同时满足”对应 $\overline{AB}=\{1,3,4,5,6\}$. 前者要求既不是偶数也不小于等于 *3*，后者只要求至少有一个条件不满足，二者是不同的事件.

另外，$C=$“点数大于 *4*”与 B 互斥，但 $B\cup C=\{1,2,3,5,6\}\neq\Omega$，所以互斥事件未必互为对立事件.

2.1.2 概率模型

到目前为止，我们仅对一个随机对象给出了所有的样本点，并且划分了事件，但并未对其分配概率. 以下两种概率模型分别从离散与连续的角度给出了对于事件概率的基本合理的定义.

定义 2.1.2 (古典概型) 假设样本空间 Ω 的样本点个数为有限个，且 n 个样本点中每个样本点 $\omega_1,\ldots,\omega_n$ 被采样到的机会都是均等的，那么对于一个事件 $A\subset\Omega$, 假如其包含的样本点个数为 m 个，则定义事件 A 的概率为：

$$P(A)=m/n.$$

我们用一个实例演示古典概型如何使用：

例 2.1.3 考虑先后独立地掷两枚硬币，假设硬币的质地均匀，掷出正反面的机会是均等的，则此时我们可以将样本空间写作：

$$\Omega=\{(正,正),(正,反),(反,正),(反,反)\}.$$

Yan and Song

事件 $A=$“第一次掷出正面”，共包含 *2* 个样本点，因此 $P(A)=1/2$.

事件 $B=$“第二次掷出正面”，共包含 *2* 个样本点，因此 $P(B)=1/2$.

事件 $C=$“两次均掷出正面”，共包含 *1* 个样本点，因此 $P(C)=1/4$.

事件 $D=$“至少一次掷出正面”，共包含 *3* 个样本点，因此 $P(D)=3/4$.

例 2.1.4 (等可能的是样本点，而不一定是点数之和) 独立地掷两枚均匀骰子，求点数之和为 *7* 的概率. 用有序数对 (i,j) 记录两枚骰子的点数，样本空间共有 $6\times 6=36$ 个等可能的样本点. 点数之和为 *7* 的样本点为

$$(1,6),(2,5),(3,4),(4,3),(5,2),(6,1),$$

所以所求概率为 $6/36=1/6$.

虽然点数之和可以取 $2,3,\ldots,12$ 这 *11* 个数，但它们并不等可能. 例如和为 *2* 只有 $(1,1)$ 这一种情形，其概率为 $1/36$. 因此不能把“和为 *7*”的概率写成 $1/11$；使用古典概型时，必须先确认所选的基本结果等可能.

通过上述掷硬币的例子，我们可以引出另一种事件间的相互关系. 从直观上来讲，第一次掷硬币的结果与第二次掷硬币的结果理应没有关系，它们二者之间应当是相互独立的. 从计算结果来看，我们有 $P(C)=P(AB)=P(A)P(B)$. 我们下面引入独立事件（independent events）的定义：

定义 2.1.5 (独立事件) 设 A,B 为同一个样本空间中的两个事件，若其满足

$$P(AB)=P(A)P(B),$$

则称事件 A,B 相互独立.

小贴士

两个事件相互独立的定义较为直接，而多个事件之间相互独立的定义则较为复杂，需要保证从所有事件中任取若干个事件，它们的交集的概率均为这些取定的事件的概率的乘积.

例 2.1.6 (互斥与独立是两回事) 掷一次均匀骰子，设 $A=$“点数为偶数”，$B=$“点数不超过 *3*”，则

$$P(A)=\frac{1}{2},\quad P(B)=\frac{1}{2},\quad P(AB)=\frac{1}{6}\neq\frac{1}{4}=P(A)P(B).$$

因此 A,B 既不互斥，也不独立. 若改取 $C=$“点数不超过 *4*”，则 $AC=\{2,4\}$，有

$$P(AC)=\frac{2}{6}=\frac{1}{2}\times\frac{4}{6}=P(A)P(C),$$

Yuan and Song

所以 A, C 相互独立，即使它们来自同一次掷骰子试验.

再取 $D =$“点数为奇数”，则 A, D 互斥，却不独立，因为 $P(AD) = 0$ 而 $P(A)P(D) = 1/4$. 一般地，两个概率均为正的互斥事件不可能独立.

例 2.1.7 (两两独立还不够) 独立地掷两枚均匀硬币，设 $A =$“第一枚为正面”，$B =$“第二枚为正面”，$C =$“两枚朝向相同”. 这三个事件的概率均为 $1/2$，任意两个同时发生的概率均为 $1/4$，所以它们两两独立.

但 ABC 只包含“两枚均为正面”这一个结果，故

$$P(ABC) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C).$$

因此三个事件并不相互独立. 判断多个事件的相互独立性，只检查每一对事件是不够的.

回到最初掷两枚硬币的例子，沿用其中的事件 A, B, D，我们也观察到，

$$\boxed{P(D) = P(A \cup B) = P(A) + P(B) - P(AB),}$$

这一等式可以从集合的视角较为容易地推导出，这也是概率计算中普遍成立的一个等式. 特别地，当 A, B 为互斥事件时，我们有

$$\boxed{P(A \cup B) = P(A + B) = P(A) + P(B),}$$

后者一般称作**加法公式**.

例 2.1.8 (至少一个发生与对立事件) 某设备有两套独立工作的备用模块，每套模块在一次任务中正常工作的概率均为 0.9. 设 A, B 分别表示两套模块正常工作，则“至少一套正常工作”的概率为

$$P(A \cup B) = 0.9 + 0.9 - 0.9 \times 0.9 = 0.99.$$

也可以先求“两套都失效”的概率，得到

$$P(A \cup B) = 1 - P(\bar{A}\bar{B}) = 1 - 0.1^2 = 0.99.$$

前一种方法扣除了重复计算的“两套都正常工作”，后一种方法利用了对立事件. 若有 n 套这样的独立模块，至少一套正常工作的概率就是 $1 - 0.1^n$.

最后，从古典概型的计算中，我们可以总结出概率作为样本空间上所有事件的函数，所满足的三条性质：

1. **非负性**：对于任意事件 A, $P(A) \geq 0$.
2. **规范性**：样本空间上全集 Ω 的概率 $P(\Omega) = 1$.
3. **可列可加性**：若事件 $A_1, A_2, \ldots$ 互斥，则有 $P(A_1 + A_2 + \cdots) = P(A_1) + P(A_2) + \cdots$.

Yuan and Song

实际上，上述三条性质即为概率的公理化定义中的三条基本公理.

由于古典概型只能处理样本空间容量有限的情形，对于无限但是可以度量的情形，我们相对应地引入几何概型：

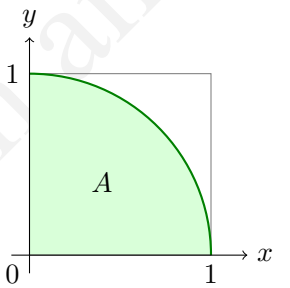
定义 2.1.9 (几何概型) 假设样本空间 Ω 是可测的，其测度（长度、面积、体积等）是 M，满足 $0 < M < \infty$，且采样在 Ω 上相对于这一测度均匀分布，即等测度的可测区域被采样到的概率相同. 那么对于一个事件 A, 假如其对应 Ω 的可测子集测度为 m, 则定义事件 A 的概率为：

$$P(A) = m/M.$$

我们同样用实例来演示几何概型的使用方法：

例 2.1.10 在区间 $[0,1]$ 内独立、均匀地取两个实数 x, y, 考虑事件 $A =$ “$x^2 + y^2 \leq 1$” 的概率.

我们将 x, y 的取值画在一个平面直角坐标系中. 其中样本空间 $\Omega = \{(x,y)|0 \leq x, y \leq 1\}$ 为一个正方形，其面积为 1. 而事件 $A = \{(x,y)|0 \leq x, y \leq 1, x^2 + y^2 \leq 1\}$ 为四分之一圆，其面积为 $\pi/4$. 因此事件 A 的概率即为 $P(A) = \pi/4$.



例 2.1.11 (约会问题中的面积法) 两人约定在一小时内到达某处，各自的到达时刻在这一小时内独立、均匀分布. 先到的人最多等待 *15* 分钟，求两人能够见面的概率.

以分钟为单位记录两人的到达时刻 x, y,样本空间是 $[0,60]^2$,见面条件为 $|x-y| \leq 15$. 不能见面的区域是正方形中满足 $y > x + 15$ 或 $x > y + 15$ 的两个直角三角形，每个三角形的直角边长均为 *45*. 因此

$$P(\text{见面}) = 1 - \frac{2 \times \frac{1}{2} \times 45^2}{60^2} = \frac{7}{16}.$$

这里并不是“等待时间占一小时的 $1/4$，所以见面概率就是 $1/4$”；需要度量的是两个到达时刻共同构成的平面区域.

Yuan and Song

小贴士

对于几何概型，有关“样本空间究竟关于哪种测量方式是均匀的”在某些问题中会有不唯一的回答，也会进一步导致概率计算的差异. 经典反例可参看贝特朗（Bertrand）奇论. 这一问题也进一步推动了概率的定义朝着公理化方向发展，即仅规定概率满足的最基本性质，而将具体概率值的计算从概率定义中剥离出来.

除古典概型和几何概型之外，在实际生活中，我们有时也会通过长期观察事件发生的频率来估计事件的概率，随后使用事件之间的关系来推导更加复杂的事件的概率. 下一小节我们将讨论，在已经给定某些事件的概率后，如何求更复杂的事件的概率，其中最基本的概念即为**条件概率**.

2.1.3 条件概率、全概率公式与贝叶斯公式

对于掷骰子问题，我们想要探究：在已知掷出的点数小于等于 3 的情况下，掷出的点数为 1 的概率进一步是多少？直观告诉我们，掷出点数小于等于 3 共有 3 种情形，而掷出点数为 1 是其中的一种，因此上述概率应当为 $\frac{1}{3}$. 我们记事件“掷出点数小于等于 3”为事件 A,“掷出的点数为 1”为事件 B, 我们一般记“已知掷出的点数小于等于 3 的情况下，掷出的点数为 1”的概率为 $P(B|A)$, 进一步我们发现，

$$P(B|A)=\frac{1}{3}=\frac{\frac{1}{6}}{\frac{3}{6}}=\frac{P(AB)}{P(A)},$$

这实际上是一个普遍成立的等式，也是条件概率（conditional probability）的严格定义方式.

定义 2.1.12 (条件概率) 设 A, B 为同一个样本空间中的两个事件，且 $P(A)>0$, 记

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

为已知事件 A 发生的条件下事件 B 发生的概率，即条件概率.

有时候，条件概率可能是更加直观或是更容易求出的值，而我们通常可以利用条件概率反过来求得交事件的概率，即

$$\boxed{P(AB)=P(A)P(B|A),}$$

该公式一般被称作**乘法公式**. 乘法公式可以通过递推的方式推广到多个事件的情形中：

$$\begin{aligned} P(A_1A_2\cdots A_n) &= P(A_1A_2\cdots A_{n-1})P(A_n|A_1A_2\cdots A_{n-1}) \\ &= \cdots \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1}). \end{aligned}$$

例 2.1.13 (不放回抽样与乘法公式) 袋中有 *3* 个红球和 *2* 个蓝球，每次等可能地摸出一个球，连续摸两次且不放回. 设 $A=$“第一次为红球”，$B=$“第二次为红球”. 已知 A 发生后，袋中只剩 *2* 个红球和 *2* 个蓝球，故

$$P(B|A)=\frac{2}{4}=\frac{1}{2},\qquad P(AB)=P(A)P(B|A)=\frac{3}{5}\times\frac{2}{4}=\frac{3}{10}.$$

不能将 $P(AB)$ 写成 $(3/5)^2$，因为第一次摸球改变了第二次面对的球的组成. 若继续摸第三个球，三次均为红球的概率为

$$\frac{3}{5}\times\frac{2}{4}\times\frac{1}{3}=\frac{1}{10}.$$

每一步的条件概率都应根据前面已经发生的结果重新计算.

我们下面考虑如下情形：现在我们刚刚完成高考，我们想估计我们上北京大学的希望有多大. 直接估计这一概率并不容易，但我们可以查阅资料大概了解到，位于不同分数段的同学报考北京大学被录取的比率是多少，从而估计相应的概率；同时我们又可以根据自己高中时的经验，大致给出自己能够考到各个分数段的概率，那么就可以最终给出上北京大学的概率大致为多少了.

我们下面将上述过程抽象为一般的概率问题. 首先，我们需要像分数段一样的一组事件 $A_1,\ldots,A_n$, 它们满足两两互斥，且并集为全集 Ω, 我们称这样的事件为**完备事件组**，或称它们构成了样本空间的一个**分割**. 以下假设各 $P(A_i)>0$，以便使用相应的条件概率. 随后，任给一个事件 B, 利用事件的分配律，我们有：

$$B=B\cap\Omega=B\cap(A_1\cup A_2\cup\cdots\cup A_n)=(BA_1)\cup(BA_2)\cup\cdots\cup(BA_n),$$

且我们注意到 $(BA_1),\ldots,(BA_n)$ 之间也是两两互斥的，这是因为

$$(BA_i)\cap(BA_j)=B\cap(A_i\cap A_j)=B\cap\varnothing=\varnothing,\quad 1\le i,j\le n,\ i\ne j.$$

从而使用加法公式，我们有

$$\boxed{P(B)=\sum_{i=1}^{n}P(A_iB)=\sum_{i=1}^{n}\big(P(A_i)P(B|A_i)\big),}$$

这一公式通常被称为**全概率公式**，它体现了概率论中分而治之的思想. 我们下面举一个具体的实例.

yuan wilson

例 2.1.14 箱子中有 a 个黄球与 b 个蓝球，其中 a,b 为正整数，小明和小红两人先后等可能地摸球. 小明先摸出一个球后不放回，求小红摸出黄球的概率.

在不知道小明摸球结果时，由抽取位置的对称性，我们猜测小红摸出黄球的概率仍为 $a/(a+b)$. 但若已知小明摸出的颜色，小红摸出黄球的条件概率会随之改变，所以两人的摸球结果并不独立. 下面我们尝试使用全概率公式来计算未获知小明摸球结果时的概率.

显然“小明摸出黄球”与“小明摸出蓝球”为一个完备事件组，我们分别记为 A_1, A_2，同时我们也可以直接求得小明摸出一个球后，小红摸出黄球的条件概率，我们记“小红摸出黄球”为事件 B，其概率的计算如下：

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) = \frac{a-1}{a+b-1}\frac{a}{a+b} + \frac{a}{a+b-1}\frac{b}{a+b} \\ &= \frac{a(a+b-1)}{(a+b)(a+b-1)} = \frac{a}{a+b} \end{aligned},$$

也就是说，小红在小明后摸出黄球的概率，与第一次小明摸出黄球的概率相同，均为 $a/(a+b)$，这就是抽签结果与抽签顺序无关的原理.

例 2.1.15 (全概率公式是按来源加权) 某仓库的零件有 60% 来自甲工厂、40% 来自乙工厂，两厂零件的次品率分别为 1% 和 3%. 从仓库中等可能地取一个零件，求它为次品的概率.

设 A_1, A_2 分别表示“来自甲厂”和“来自乙厂”，B 表示“为次品”. A_1, A_2 构成完备事件组，因此

$$P(B) = 0.6 \times 0.01 + 0.4 \times 0.03 = 0.018.$$

所以总体次品率为 1.8%，而不是两厂次品率的简单平均值 2%. 每种来源的条件概率必须按该来源所占比例加权.

最后，我们再来介绍一下**贝叶斯公式**（Bayesian Equation），它完全使用已有的条件概率与全概率公式的工具，但却体现着截然不同的思想.

定义 2.1.16 (贝叶斯公式) 给定事件 $A_1, \ldots, A_n$ 为一个完备事件组，满足 $P(A_i) > 0$，及另一个事件 B，满足 $P(B) > 0$，已知 $P(B|A_i), \forall i = 1, \ldots, n$，我们有

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

从公式本身来看，贝叶斯公式只是将条件概率与全概率公式组合了起来，但其背后却有着深刻的意义. 在许多应用中，A_i 表示某种可能的来源或原因，B 表示观测到的结果，$P(B|A_i)$ 描述已知来源时结果出现的概率. 贝叶斯公式则求已知结果时各个来源的

Kuan and Song

概率，体现了“由果溯因”的逆向思维，因此也往往称为**逆概率公式**. 不过，条件概率本身并不要求两个事件存在时间先后或因果关系. 我们下面用一个具体实例来加深印象：

例 2.1.17 小明上学经常迟到，如果是普通的晴天，那么小明上学迟到的概率为 40%; 如果下雨，则小明上学迟到的概率高达 80%. 根据以往经验及统计数据，小明居住的城市在当下季节有 50% 的概率会下雨. 而小明今天上学果然迟到了. 那么根据小明已经迟到了这一已知信息，我们需要反推今天确实下雨了这件事的概率.

设“下雨”为事件 A, “小明迟到”为事件 B, 则

$$P(A)=0.5,\quad P(B|A)=0.8,\quad P(B|\bar{A})=0.4,$$

根据贝叶斯公式，我们有

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}=\frac{0.8\times 0.5}{0.8\times 0.5+0.4\times 0.5}=\frac{2}{3},$$

因此，在已知小明迟到的前提下，当天确实下雨的概率为 2/3.

从以上例子中我们可以看出，如果没有小明迟到这一信息，我们根据统计规律有当地下雨的概率为 1/2, 而我们已知了小明迟到，这就告诉我们当天小明确实更有可能遇到了下雨天，因此在小明迟到的前提下，下雨的概率上升至了 2/3. 事实上，在贝叶斯理论中，我们把原先完备事件组（如下雨和不下雨）的概率称为**先验概率**，将获知的某个事件（如小明迟到）称为一次**观测**，将该事件发生前提下原先完备事件组的条件概率称为**后验概率**. 在给定概率模型下，后验概率体现了观测信息对原有判断的更新.

例 2.1.18 (识别结果还要结合先验比例) 假设一批邮件中 10% 是垃圾邮件. 一个过滤器会把 90% 的垃圾邮件标记出来，同时误把 5% 的正常邮件也标记出来. 某封邮件被标记后，它确实是垃圾邮件的概率是多少？

设 $A=$“是垃圾邮件”，$B=$“被标记”，则

$$P(A|B)=\frac{0.9\times 0.1}{0.9\times 0.1+0.05\times 0.9}=\frac{2}{3}.$$

也可以用 *1000* 封邮件帮助理解：按上述比例，有 *100* 封垃圾邮件，其中 *90* 封被标记；*900* 封正常邮件中有 *45* 封被误标. 因此被标记的 *135* 封中，垃圾邮件所占比例为 $90/135=2/3$.

注意，90% 描述的是 $P(B|A)$，题目求的是 $P(A|B)$，二者不能直接互换. 若垃圾邮件的先验比例改变，即使过滤器的两项识别概率不变，后验概率也会改变.

2.2　随机变量、分布与数学期望

在上一节中，我们定义了样本空间，它可以是有限的离散的数、可以是无限但可数的离散的数，也可以是在某个区间中连续的数，当然也有可能不是以数字的形式存在的，如比赛胜负的结果等. 然而，对于统计学家来讲，大家更希望将所有的这些形式不同的样本空间都用数字来表达，以便可以使用数学的工具进行更深入的研究. 譬如我们可以将比赛胜利记为 1，失败记为 0, 这样就可以将非数字的样本空间转化为数字的样本空间. 对于更加复杂的样本空间我们也可以相应地设计更加复杂的转化方式，将它们都转换为数字. 对于这些以数字形式存在的随机对象，我们便将其称为**随机变量**（Random variables），通常用大写字母 X, Y, Z 等来表示.

2.2.1 随机变量与分布

根据随机变量的性质，我们可以将其分为两类：一类为**离散随机变量**，它的取值是一些离散的数值，如 $1, 2, 3, \ldots$, 可以是有限个也可以是无限个；另一类为**连续随机变量**, 它的取值为实数轴 $\mathbb{R}$ 上的一段连续区域，如 $[0, 1]$, 其长度可以是有限的，也可以是无限长的，甚至是实数轴本身.

对于一个随机变量 X，我们可以沿用事件的方式，关注随机变量取值范围中的一个子集，并给出它的概率. 然而，子集的数量是庞大的，求出所有子集对应的事件的概率并不容易. 既然我们已经将样本空间转化为了数字，那么就可以很好地利用数字的特征. 数字的一个很好的特点就是可以在其上很方便地定义函数. 我们可以使用**分布**（Distribution）的概念，来描述随机变量具体是以什么样的方式、什么样的概率，在数轴上分布的. 我们需要对离散随机变量与连续随机变量的分布进行分别定义.

离散随机变量的分布定义较为直接，参照古典概型，每个取值均可作为一个基本事件，研究清楚了所有取值对应的概率，我们就可以对该随机变量有一个整体的把握.

定义 2.2.1 (离散随机变量的分布) 给定离散随机变量 X, 设其所有取值构成的集合为 $\mathcal{X}$, 定义以 $\mathcal{X}$ 为定义域的函数

$$f(x) := P(X = x), \quad x \in \mathcal{X},$$

为随机变量 X 的概率质量函数，也称分布列.

例 2.2.2 (随机变量把多个样本点合并为一个取值) 独立地掷两枚均匀硬币，令 X 表示正

面出现的次数. 四个等可能的样本点被映射为

$$(反,反)\mapsto 0,\quad (正,反),(反,正)\mapsto 1,\quad (正,正)\mapsto 2.$$

因此 X 的分布列为

$$f(0)=\frac{1}{4},\qquad f(1)=\frac{1}{2},\qquad f(2)=\frac{1}{4}.$$

这说明样本点等可能，并不意味着随机变量的不同取值也等可能. 求 $P(X\geq 1)$ 时，将相应取值的概率相加即可，得到 $1/2+1/4=3/4$.

对于连续随机变量，因为实数轴上的一个区间包含无数个点，随机变量取一个特定的点的概率基本都是 0. 有意义的是研究随机变量取值落在任意一段的概率，而一个比较好的排列方式就是从左至右以此扫描随机变量的取值范围并赋予相应的概率.

定义 2.2.3 (连续随机变量的分布) 给定连续随机变量 X, 定义以 $\mathbb{R}$ 为定义域的函数

$$F(x):=P(X\leq x),\quad x\in\mathbb{R},$$

为随机变量 X 的累积分布函数，简称分布函数.

小贴士

对于离散随机变量，我们要求所有取值的概率之和为 1, 即 $\sum_{x\in\mathcal{X}} f(x)=1$. 对于连续随机变量，它的累积分布函数是一个单调不减函数，且设其定义域左右端点分别为 a,b, 则有 $F(a)=0, F(b)=1$. 进一步，我们也可以将其延拓为一个定义在 $\mathbb{R}$ 上的函数，只需令 $F(x)=0, x\leq a; F(x)=1, x\geq b$ 即可. 事实上，在概率论中，我们通常认为连续随机变量的取值范围均是 $\mathbb{R}$，而将那些真正的可能取值的部分，即累积分布函数在上涨的区域称为随机变量的支撑（support）.

对于连续随机变量，相较累积分布函数，我们更常用的是**概率密度函数**（Probability Density Function）：

定义 2.2.4 (概率密度函数) 给定连续随机变量 X, 不妨设其取值范围为 $\mathbb{R}$, 已知其分布函数为 $F(x)$, 若存在另一函数 $f(x)$ 满足

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t,$$

且 $f(x)\geq 0$，则称 $f(x)$ 为 X 的概率密度函数，简称密度函数.

多数情况下（譬如累积分布函数可微时），概率密度函数就是累积分布函数的导数，即 $f(x)=F'(x)$. 同时我们不难得出，随机变量支撑之外的区域内有 $f(x)\equiv 0$. 概率密

度函数的意义与物理中密度的意义相似. 物理中密度表示单位体积内的质量大小，而概率密度的含义就是在实数轴上单位长度内的概率大小. 概率密度函数的一个作用是，使用牛顿-莱布尼茨公式，对于任意 $a, b \in \mathcal{X}$, 我们有

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\mathrm{d}x,$$

因此我们可以使用积分来求出连续随机变量落入任何区间内的概率.

例 2.2.5 (从密度求分布函数与区间概率) 设随机变量 X 的密度函数为

$$f(x) = \begin{cases} cx, & 0 \leq x \leq 1, \\ 0, & \text{其他}, \end{cases}$$

求常数 c、分布函数 $F(x)$ 及 $P(1/2 < X \leq 1)$. 密度的总积分必须为 *1*，故

$$1 = \int_0^1 cx\mathrm{d}x = \frac{c}{2}, \qquad c = 2.$$

从左向右累积积分，得到

$$F(x) = \begin{cases} 0, & x < 0, \\ x^2, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

所以 $P(1/2 < X \leq 1) = F(1) - F(1/2) = 3/4$. 虽然 $f(1) = 2 > 1$，但它是密度值，并不是事件的概率；任意单点仍有 $P(X = x) = 0$，区间概率则由密度的积分给出.

2.2.2 常用分布

人们的长相千姿百态，而我们在生活经验中总结出了“国字脸”、“瓜子脸”等具有代表性的脸型. 同样，随机变量的分布千姿百态，而统计学家们也总结出了一些常用的随机变量的分布的模式类型. 本小节我们来列举若干常用的分布类型，我们先介绍离散随机变量的分布，再介绍连续随机变量的分布.

我们首先考虑最简单的情形，当随机变量的取值只有两个时，我们只需要一个参数 p, 就可以完全确定随机变量的分布情况.

例 2.2.6 (伯努利分布/0-1 分布) 离散随机变量 X 取值为 $\{0, 1\}$, 给定常数 $p \in [0, 1]$, X 的分布为

$$f(1) = p, \quad f(0) = 1 - p.$$

称该分布为伯努利分布（*Bernoulli Distribution*），或 *0-1* 分布.

Yuan and Song

伯努利分布可以表示一个实验或是一场比赛成功或失败的概率. 而对于相互独立的 n 个实验或是 n 场比赛，每次的成功概率均是 p, 记累计成功的次数为 X, 则如何求 X 的分布呢？假设成功 x 次，这 x 次在 n 次中共计有 C_n^x 种，由于每次实验或比赛相互独立，我们可以直接使用相互独立事件的乘法公式求出每种情况的概率，并将 C_n^x 种情况加在一起求得总共的概率. 这就得到了二项分布.

例 2.2.7 (二项分布) 离散随机变量 X 取值为 $\{0,1,2,\ldots,n\}$, 给定常数 $p\in[0,1]$, X 的分布为

$$f(x)=C_n^x p^x(1-p)^{n-x},\quad x=0,1,2,\ldots,n.$$

称该分布为二项分布（*Binomial Distribution*），记为 $B(n,p)$, 其中 $n=1$ 时便退化为了伯努利分布.

例 2.2.8 (一次是否成功与多次成功的次数) 假设某服务每次请求成功的概率为 0.8，各次结果相互独立. 令 I 在一次请求成功时取 *1*，失败时取 *0*，则 I 服从参数 $p=0.8$ 的伯努利分布.

若连续发送 *5* 次请求，令 X 为成功次数，则 X 服从 $B(5,0.8)$. 恰好成功 *4* 次的概率为

$$P(X=4)=C_5^4(0.8)^4(0.2)=0.4096.$$

其中 $C_5^4=5$ 表示唯一一次失败可能出现在 *5* 个不同位置. 若问“至少成功一次”，用对立事件计算更简便：

$$P(X\geq 1)=1-P(X=0)=1-(0.2)^5=0.99968.$$

对于二项分布，如果 p 为有理数，即存在非负整数 m,M, 使得 $p=m/M$, 则我们可以将二项分布视为在含有 M 个小球，其中有 m 个是白球的箱子中，有放回地抽取 n 次，其中抽到白球的次数是多少. 相对应地，我们自然可以考虑如果不放回，那么抽取 n 次，抽到白球的个数有多少. 我们同样是使用古典概型的基本事件的方法，考虑抽到 x 个白球的事件包含多少个基本事件，这里 x 的取值最小是 $\max\{0,n-M+m\}$, 最大是 $\min\{m,n\}$. 总共的基本事件个数为组合数 C_M^n, 而抽到 x 个白球的基本事件有 $C_m^x\times C_{M-m}^{n-x}$ 个. 因此我们可以定义如下分布：

例 2.2.9 (超几何分布) 离散随机变量 X 取值为 $\mathcal{X}=\{a,a+1,\ldots,b\}$, 其中 $a=\max\{0,n-M+m\}$, $b=\min\{m,n\}$, X 的分布为：

$$f(x)=\frac{C_m^x\times C_{M-m}^{n-x}}{C_M^n},\quad x\in\mathcal{X}.$$

Yuan and Song

称该分布为超几何分布（*Hypergeometric Distribution*），记为 $H(n, M, m)$.

例 2.2.10 (有放回与不放回应选不同分布) *10* 个零件中有 *3* 个次品，等可能地抽取 *4* 个且不放回，令 X 为抽到的次品个数，则 X 服从 $H(4, 10, 3)$. 恰好抽到 *1* 个次品的概率为

$$P(X=1)=\frac{C_3^1C_7^3}{C_{10}^4}=\frac{3\times 35}{210}=\frac{1}{2}.$$

分子分别选择 *1* 个次品和 *3* 个正品，分母选择任意 *4* 个零件.

若改为每次抽后放回并充分混合，*4* 次抽取相互独立，此时 X 服从 $B(4, 0.3)$，相应概率变为

$$P(X=1)=C_4^1(0.3)(0.7)^3=0.4116.$$

是否放回改变了各次抽取之间的关系，因此也改变了所用的分布.

二项分布关注 n 次实验中总计成功的次数，而如果我们只关注第一次实验成功出现在什么时候，就得到了几何分布. 如果第一次实验成功出现在第 x 次，那么它的前 $x-1$ 次实验必定失败，因此相应的概率为 $p(1-p)^{x-1}$. 我们定义几何分布如下：

例 2.2.11 (几何分布) 离散随机变量 X 取值为 $\mathbb{N}_+ := \{1, 2, \dots\}$, 给定常数 $p \in (0, 1]$, X 的分布为：

$$f(x)=p(1-p)^{x-1},\quad x\in\mathbb{N}_+,$$

称该分布为几何分布（*Geometric Distribution*）.

小贴士

几何分布名称的由来为：几何分布的分布列概率相加构成了一个几何级数（即等比数列的无穷推广）. 而超几何分布之所以如此命名，是因为其分布函数的形式与超几何级数相似，而超几何级数的形式十分复杂，我们不进行讨论.

例 2.2.12 (第一次成功要等到什么时候) 每次独立尝试连接某服务器，成功概率均为 0.2，令 X 表示首次连接成功时的尝试次数，则 X 服从几何分布. 首次在第 *3* 次成功，要求前两次失败而第三次成功，所以

$$P(X=3)=(0.8)^2\times 0.2=0.128.$$

“尝试 *3* 次后仍未成功”则要求前三次全部失败，故 $P(X>3)=(0.8)^3=0.512$. 进一步，

$$P(X>5\mid X>3)=\frac{(0.8)^5}{(0.8)^3}=(0.8)^2=P(X>2).$$

在各次尝试独立且成功概率不变的模型中，已经失败 3 次，不会改变接下来两次仍然失败的概率. 这一性质称为无记忆性.

我们下面考虑一个具体场景：消防站一天内可能会接到若干次 119 火警电话，我们想探究一天内接到火警电话次数的随机规律. 我们理论上可以将其建模为二项分布，即将全城每个人视为一次独立实验，而每个人在一天内以一定的概率拨打 119，我们假设一个人不会在一天内拨打两次 119. 然而，这样建模的二项分布，n 极大，可能是几万到几百万不等，而 p 极小，因为险情的发生是小概率事件. 因此这样建模的二项分布随机变量 X, 只有极少一部分接近于 0 的取值概率是比较显著的，其他较大的取值对应的概率基本十分趋近于 0，但是它们的计算依然需要用到组合数，十分消耗计算资源. 因此，我们需要寻求近似的、且计算较为简单的另一个分布. 我们注意到，$\lambda := np$ 可能是一个比较适中的数字，而当 x 较大时，二项分布的分布函数为

$$\begin{aligned} f(x) = C_n^x p^x (1-p)^{n-x} &= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\cdot\left(1-\frac{\lambda}{n}\right)^{n-x}. \end{aligned}$$

当 x 远小于 n 时，上式中除了首尾两个因子，中间所有的因子值都十分接近于 1，而对于最后一个因子，我们根据微积分的基本结论：

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x} = e^{-\lambda},$$

因此我们可以得出一个分布函数形如 $f(x) = \dfrac{\lambda^x}{x!}e^{-\lambda}$, 而我们使用泰勒公式不难验证，上述分布函数满足完备性，即所有取值的概率之和为 1. 因此我们给出如下的分布定义：

例 2.2.13 (泊松分布) 离散随机变量 X 取值为 $\mathbb{N} := \{0, 1, 2, \dots\}$, 给定常数 $\lambda \geq 0$, X 的分布为：

$$f(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x \in \mathbb{N},$$

称该分布为泊松分布（*Poisson Distribution*）.

例 2.2.14 (把平均到达次数代入泊松分布) 假设某窗口每分钟收到的请求数 X 服从参数 $\lambda = 2$ 的泊松分布，则一分钟内恰好收到 3 次请求的概率为

$$P(X=3) = \frac{2^3}{3!}e^{-2} = \frac{4}{3}e^{-2} \approx 0.1804,$$

至少收到一次请求的概率为 $1 - P(X=0) = 1 - e^{-2} \approx 0.8647$.

参数 λ 对应指定时间段的平均次数. 如果在相应的泊松到达模型中，半分钟内的请求数服从参数 $\lambda = 1$ 的泊松分布，那么半分钟内没有请求的概率为 e^{-1}，不能继续代入

一分钟所用的参数 *2*.

我们下面考虑连续随机变量的分布. 首先我们考虑最简单的分布——均匀分布. 它是几何概型在随机变量语境下的另一种表述.

例 2.2.15 (均匀分布) 连续随机变量 X 取值为 $[a,b]$, a,b 为有限的常数，X 的密度函数为

$$f(x)=\frac{1}{b-a},\quad x\in[a,b],$$

称该分布为均匀分布（*Uniform Distribution*），记为 $U[a,b]$.

例 2.2.16 (均匀分布中的区间长度) 假设某班车到站的剩余等待时间 X 在 $[0,10]$ 分钟内均匀分布，则 $f(x)=1/10$，$0\le x\le 10$. 等待时间介于 *2* 分钟和 *5* 分钟之间的概率为

$$P(2\le X\le 5)=\int_2^5\frac{1}{10}\mathrm{d}x=\frac{3}{10}.$$

若已知至少还要等 *4* 分钟，则新的可能范围是 $[4,10]$，此时

$$P(X\le 7\mid X\ge 4)=\frac{P(4\le X\le 7)}{P(X\ge 4)}=\frac{3/10}{6/10}=\frac{1}{2}.$$

计算条件概率时，作为分母的总长度已经由 *10* 变为 *6*.

在讨论离散随机变量时，我们给出了几何分布，它表示独立地执行同样的实验时，首次成功出现的时间. 我们想要将其推广到连续的情形. 我们知道等比数列在连续情形下的推广就是指数函数，因此一个指数函数有望成为一个分布，只要其满足完备性条件. 我们知道，$\int_0^{+\infty}e^{-\lambda x}\mathrm{d}x=\lambda^{-1}$, 因此只需配备一个归一化系数 λ, 我们就得到了指数分布

例 2.2.17 (指数分布) 连续随机变量 X 取值为 $[0,+\infty)$, $\lambda>0$ 为常数，X 的密度函数为

$$f(x)=\lambda e^{-\lambda x},\quad x\in[0,\infty),$$

称该分布为指数分布（*Exponential Distribution*）.

指数分布时常用来描述电器等耐用器件的寿命. 因为这些器件在单位时间内都有一定的概率会损坏，而作为几何分布在连续意义下的推广的指数分布，恰好描绘了首次损坏出现的时间，也正是器件的寿命.

例 2.2.18 (指数分布中的生存概率) 假设某器件的寿命 T 以小时为单位，服从参数 $\lambda=0.01$ 的指数分布. 它能工作超过 *100* 小时的概率为

$$P(T>100)=\int_{100}^{+\infty}0.01e^{-0.01t}\mathrm{d}t=e^{-1}\approx 0.3679.$$

若器件已经正常工作了 *100* 小时，它还能再工作超过 *50* 小时的条件概率为

$$P(T>150\mid T>100)=\frac{e^{-1.5}}{e^{-1}}=e^{-0.5}=P(T>50).$$

Yuan-qing Song

这就是指数分布的无记忆性，与几何分布的例子相对应. 该结论依赖指数寿命模型；若器件会明显老化，便不能直接作此推断.

我们上文中讨论的连续分布的取值范围均为 $\mathbb{R}$ 的一个真子集. 我们下面想要寻找一个定义在 $\mathbb{R}$ 上的随机变量. 因为完备性，即 $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$, 我们就不能考虑像均匀分布那样的简易的情形. 我们希望这个分布有一个中心，在这里的概率密度最高，而向两边分散时，概率密度逐渐趋向于 0，整个密度函数类似于一个钟形曲线. 在所有的初等函数中，我们锁定了 e^{-ax^2} 这一较为简单的函数满足上述条件. 下面要做的是根据完备性条件，求得上述函数的归一化系数. 这一计算可以使用两个相同积分耦合以及极坐标变换完成，这里不展开推导. 相应的归一化系数是 $\sqrt{a/\pi}$. 在实际问题中，我们令 $a = 1/(2\sigma^2)$, σ 有特殊的含义；而进一步由于平移不影响积分的结果，我们可以使用 $(x-\mu)^2$ 代替 x^2, μ 同样有特殊的含义. 我们将其总结为如下分布：

例 2.2.19 (正态分布/高斯分布) 连续随机变量 X 取值为 $\mathbb{R}$, 给定 $\mu \in \mathbb{R}, \sigma > 0$ 为常数，X 的密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

称该分布为正态分布（*Normal Distribution*），或高斯分布（*Gaussian Distribution*），记为 $N(\mu, \sigma^2)$.

例 2.2.20 (正态分布先标准化再计算) 假设某传感器的读数 X 服从 $N(100, 4)$，其中均值为 *100*，标准差为 $\sqrt{4} = 2$. 令 $Z = (X - 100)/2$，通过变量代换可得 Z 服从 $N(0, 1)$. 因此

$$P(98 \le X \le 102) = P(-1 \le Z \le 1).$$

记标准正态分布函数为 $\Phi(z) = P(Z \le z)$，通过数值积分可得 $\Phi(1) \approx 0.8413$，又由对称性有 $\Phi(-1) = 1 - \Phi(1)$，所以

$$P(98 \le X \le 102) = 2\Phi(1) - 1 \approx 0.6827.$$

注意 $N(100, 4)$ 中的第二个参数是方差 *4*，标准化时应除以标准差 *2*.

2.2.3 数学期望

对于一列数 $x_1, \ldots, x_n$, 我们有平均数的概念，即 $\bar{x} := \frac{1}{n}(x_1 + \cdots + x_n)$. 而如果我们想要给所有数字分配不等的权重，我们可以额外定义权重 $w_1, \ldots, w_n$ 满足 $\sum_{i=1}^n w_i = 1$, 以及加权平均数 $\bar{x}_w := \sum_{i=1}^n w_i x_i$. 我们可以将权重的概念推广到随机变量与概率的语境中，用概率来代表不同取值的权重，从而也得到一个某种意义上的关于某个随机变量

的平均的数值，我们称这个数值为随机变量的均值（Mean）或数学期望（Expectation），简称期望. 离散随机变量与连续随机变量的数学期望的定义分别如下：

定义 2.2.21 (离散随机变量的数学期望) 对于离散随机变量 X，设其取值范围为 $\mathcal{X} = \{x_1, \ldots, x_n, \ldots\}$, 分布列为 $f(x)$, 如果级数 $\sum_{i=1}^{\infty} |x_i| f(x_i)$ 收敛，则称级数

$$\sum_{i=1}^{\infty} x_i f(x_i)$$

为随机变量 X 的数学期望，记作 $\mathbb{E}X$.

小贴士

考虑级数 $\sum_{i=1}^{\infty} |x_i| f(x_i)$ 的目的是，有些随机变量可能满足 $\sum_{i=1}^{\infty} x_i f(x_i)$ 收敛，然而 $\sum_{i=1}^{\infty} |x_i| f(x_i)$ 可能发散. 通常来讲，这是由于随机变量的分布过于分散，此时考虑它的均值这一衡量总体性质的量意义不大，此时我们称它的期望不存在. 不过，如果离散随机变量的取值仅为有限个，那么它的期望必定存在.

定义 2.2.22 (连续随机变量的数学期望) 对于连续随机变量 X, 设其取值范围为 $[a, b] \subseteq \mathbb{R}$, 密度函数为 $f(x)$, 如果积分 $\int_a^b |x| f(x) \mathrm{d}x$ 收敛，则称积分

$$\int_a^b x f(x) \mathrm{d}x$$

为随机变量 X 的数学期望，记作 $\mathbb{E}X$.

例 2.2.23 (期望不一定是可能取到的值) 掷一次均匀骰子，令 X 表示点数，则

$$\mathbb{E}X = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5.$$

骰子不会掷出 *3.5* 点，但多次独立掷骰子所得点数的平均值会趋近这一数值. 期望描述按概率加权的平均水平，并不保证某次试验能取到它.

连续情形也一样. 例如前面密度为 $f(x) = 2x$、$0 \le x \le 1$ 的随机变量，其期望为

$$\mathbb{E}X = \int_0^1 x \cdot 2x \mathrm{d}x = \frac{2}{3}.$$

这比区间中点 $1/2$ 更大，因为较大的取值附近具有更大的概率密度.

我们下面试着求一部分常用分布的数学期望，其他分布的期望留作习题.

例 2.2.24 对于二项分布，我们有：

$$\begin{aligned}\mathbb{E}X &= \sum_{x=0}^{n} xC_n^x p^x(1-p)^{n-x} = \sum_{x=0}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}\\ &= \sum_{x=1}^{n} nC_{n-1}^{x-1}p\cdot p^{x-1}(1-p)^{n-x} = np\sum_{y=0}^{n-1} C_{n-1}^{y}p^y(1-p)^{n-1-y} = np,\end{aligned}$$

其中最后一个等式利用了一个新的二项分布的完备性. 因此，我们有二项分布的均值即为 np, 特别地，当 $n=1$ 时，相应的伯努利分布的均值为 p.

例 2.2.25 对于正态分布，我们有：

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x\\ &= \int_{-\infty}^{+\infty}(x-\mu)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x + \int_{-\infty}^{+\infty}\frac{\mu}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x\\ &\xlongequal{y=x-\mu} \int_{-\infty}^{+\infty}\frac{y}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y + \mu = \mu,\end{aligned}$$

其中最后一个等式利用了奇函数的中心对称性. 因此，我们有正态分布的均值即为 μ, 这也是正态分布中参数 μ 的实际意义.

我们下面来讨论一下数学期望的性质. 首先，数学期望具有线性性，即：

性质 2.2.26 (线性性) 对于任意两个期望存在的随机变量 X, Y, 与任意三个常数 $a, b, c \in \mathbb{R}$, 我们有

$$\mathbb{E}(aX+bY+c) = a\mathbb{E}X + b\mathbb{E}Y + c.$$

我们将数学期望展开为级数或是积分的形式，并利用级数或是积分的线性性，就可验证上述关于期望的线性性质的正确性. 线性性可以帮助我们更方便地求得更多分布的均值.

例 2.2.27 (线性性不要求相互独立) 从 *1* 至 *10* 号共 *10* 张签中不放回地抽取两张，其中 *1* 至 *3* 号为中奖签. 令 X_1, X_2 分别表示两次是否中奖，中奖取 *1*，否则取 *0*. 由抽签位置的对称性，两者的中奖概率均为 $3/10$，所以

$$\mathbb{E}(X_1+X_2) = \mathbb{E}X_1 + \mathbb{E}X_2 = \frac{3}{10}+\frac{3}{10} = \frac{3}{5}.$$

虽然第二次是否中奖受第一次结果影响，期望仍然可以相加. 若每次中奖获得 *10* 元，参与两次抽签共花费 *4* 元，则净收益为 $R = 10(X_1+X_2)-4$，有

$$\mathbb{E}R = 10\times\frac{3}{5} - 4 = 2(\text{元}).$$

这里求的是平均净收益，并不是说每次参与都能获得 *2* 元.

Yuan and Song

例 2.2.28 对于超几何分布 $H(n,M,m)$, 我们引入 n 个辅助变量 $X_1,\dots,X_n$, 分别表示第 $1,\dots,n$ 次摸到的球的指示量，当摸到白球时取 *1*，否则取 *0*. 此时这 n 个辅助变量均满足伯努利分布，而且由于抽签结果与抽签顺序无关的原理，它们的成功概率都是 m/M. 而超几何分布随机变量 X 就是这些辅助随机变量的和，因此根据数学期望的线性性，我们有：

$$\mathbb{E}X=\mathbb{E}(X_1+\cdots+X_n)=\mathbb{E}X_1+\cdots+\mathbb{E}X_n=n\frac{m}{M},$$

此即为超几何分布的数学期望.

数学期望的另一个性质考虑的是将一个函数 $g(x)$ 作用在随机变量上，得到另一个随机变量 $g(X)$, 并研究 $g(X)$ 的期望. 我们自然可以先求出 $g(X)$ 的分布然后再求它的期望，然而我们也有更加直接的办法，即如下的性质：

性质 2.2.29 (佚名统计学家公式) 以下假设 $\mathbb{E}|g(X)|<\infty$.

1. 已知离散随机变量 X 取值列为 $x_1,x_2,\dots$, 相应的分布列为 $f(x)$. 对于一个给定的博雷尔函数 $g(x)$, 我们有：

$$\mathbb{E}\big(g(X)\big)=\sum_{i=1}^{\infty}g(x_i)f(x_i).$$

2. 已知连续随机变量 X 取值范围为 $[a,b]$, 相应的密度函数为 $f(x)$. 对于一个给定的博雷尔函数 $g(x)$, 我们有：

$$\mathbb{E}\big(g(X)\big)=\int_a^b g(x)f(x)\mathrm{d}x.$$

小贴士

博雷尔函数（Borel Functions）的严格要求是，每个博雷尔集的原像仍是博雷尔集. 我们常用的连续函数、分段连续函数等都满足这一要求.

佚名统计学家公式在离散情形下，可以通过把映射到同一 $g(x)$ 值的概率合并，再重新求和来理解；它不要求 g 是一一映射. 对于连续情形，求 $g(X)$ 的密度函数有时候并不容易，而这一公式告诉我们，如果只是想求 $g(X)$ 的期望，仍可直接按 X 原有的分布加权，只需把期望公式中的被积因子 x 替换为 $g(x)$ 即可. 这为计算带来了方便. 连续情形的严格证明需要用到测度论知识，我们在此略去.

例 2.2.30 (直接求函数的期望) 设 X 以相同概率取 $-1,0,1$，令 $Y=X^2$. 此时 -1 与 1 都映射到 *1*，所以 Y 的分布为

$$P(Y=0)=\frac{1}{3},\qquad P(Y=1)=\frac{2}{3},$$

从而 $\mathbb{E}Y = 2/3$. 用佚名统计学家公式直接计算，也得到

$$\mathbb{E}(X^2) = (-1)^2 \cdot \frac{1}{3} + 0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} = \frac{2}{3}.$$

但 $\mathbb{E}X = 0$，故 $\mathbb{E}(X^2) \neq (\mathbb{E}X)^2$；一般不能把非线性函数移到期望符号外.

若改为 X 在 $[-1, 1]$ 上均匀分布，则无需求 X^2 的密度，便有

$$\mathbb{E}(X^2) = \int_{-1}^{1} x^2 \cdot \frac{1}{2} \mathrm{d}x = \frac{1}{3}.$$

2.3 多元随机变量与方差

在上一节中，我们研究了单个随机变量的性质及其均值，而在本节中我们要讨论多个随机变量组合起来的联合随机变量的性质. 同时我们也进一步讨论度量随机变量偏离中心程度的指标——方差，并引入它在多元随机变量情形下的变体.

2.3.1 多元随机变量

有些随机对象可能难以用一个数字来描述，如同时掷多轮骰子. 那么这时我们就需要使用一系列的数 $X_1, \ldots, X_n$ 来共同描述一个随机对象. 上述的 n 个数便构成了一组 n 元随机变量，或称随机向量（Random vectors）. 此时它们的样本空间同样需要对每个样本空间做笛卡尔积，即 $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. 每个随机变量既可以都是离散型的，也可以都是连续型的，还可以是有些离散、有些连续的. 然而离散与连续混杂的形式讨论起来较为复杂，我们在本节中只讨论全部离散或是全部连续的情形.

了解一组 n 元随机变量的性质，首先要定义它们的分布，离散与连续的 n 元随机变量的分布分别定义如下：

定义 2.3.1 (n 元离散随机变量的分布) 给定 n 元离散随机变量 $X_1, \ldots, X_n$，设其所有取值构成的集合为 $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$，其中每一个取值为一个 n 元向量，定义以 $\mathcal{X}$ 为定义域的函数

$$f(x_1, \ldots, x_n) := P(X_1 = x_1, \ldots, X_n = x_n), \quad (x_1, \ldots, x_n) \in \mathcal{X},$$

为 n 元随机变量 $X_1, \ldots, X_n$ 的联合概率质量函数，也称联合分布列.

小贴士

概率记号 $P(\cdot, \cdot)$ 中的逗号代表交事件.

Yuan and Song

例 2.3.2 (用联合分布表记录两个量) 独立地掷两枚均匀硬币，令 X 表示第一枚是否为正面，正面取 *1*，反面取 *0*；令 Y 表示两枚硬币中正面的总数. 四个等可能的结果分别给出

$$(X,Y)=(0,0),(0,1),(1,1),(1,2).$$

因此联合分布列可以写成表格：

$X\backslash Y$	0	1	2
0	1/4	1/4	0
1	0	1/4	1/4

表中每个数表示“行条件和列条件同时满足”的概率. 例如 $P(X=1,Y=2)=1/4$，而 $P(X=0,Y=2)=0$，因为第一枚为反面时不可能共出现两个正面. 表中所有概率之和为 *1*，但不是每个数对都可能出现.

定义 2.3.3 (n 元连续随机变量的分布) 给定 n 元连续随机变量 $X_1,\ldots,X_n$，设其取值范围为 $\mathcal{X}:=\mathcal{X}_1\times\cdots\times\mathcal{X}_n\subseteq\mathbb{R}^n$，定义以 $\mathcal{X}$ 为定义域的函数

$$F(x_1,\ldots,x_n):=P(X_1\le x_1,\ldots,X_n\le x_n),\quad (x_1,\ldots,x_n)\in\mathcal{X},$$

为 n 元随机变量 $X_1,\ldots,X_n$ 的联合累积分布函数，简称联合分布函数.

同样，我们也可以将 n 元连续随机变量的联合分布函数延拓至整个 $\mathbb{R}^n$ 空间，那些支撑以外的位置联合分布函数值可能是 0, 可能是 1, 也可能是中间的某个数值，我们只需代入定义计算即可. 随后我们可以引入 n 元连续随机变量的联合概率密度函数：

定义 2.3.4 (n 元连续随机变量的联合概率密度函数) 给定 n 元连续随机变量 $X_1,\ldots,X_n$，不妨设其取值范围为 $\mathbb{R}^n$，已知其联合累积分布函数为 $F(x_1,\ldots,x_n)$，若存在另一函数 $f(x_1,\ldots,x_n)$ 满足

$$F(x_1,\ldots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}f(t_1,\cdots,t_n)\mathrm{d}t_1\cdots\mathrm{d}t_n,$$

则称 $f(x_1,\ldots,x_n)$ 为 $X_1,\ldots,X_n$ 的联合概率密度函数，简称联合密度函数.

如果联合分布函数可微，那么联合密度函数应当是其关于每个分量求一次导的交叉导数，即 $f(x_1,\ldots,x_n)=\dfrac{\partial^n}{\partial x_1\cdots\partial x_n}F(x_1,\ldots,x_n)$. 在支撑以外的部分，联合密度函数恒为 0.

例 2.3.5 (联合密度要在整个支撑区域上归一化) 在三角形 $D=\{(x,y):x\ge0,y\ge0,x+y\le1\}$ 上均匀取一点 (X,Y)，设其联合密度在 D 上为常数 c，在 D 外为 *0*. 因为

D 的面积为 $1/2$，故 $c=2$；用二重积分验证也是

$$\int_0^1\int_0^{1-x} c\mathrm{d}y\mathrm{d}x=\frac{c}{2}=1.$$

事件 $X+Y\leq 1/2$ 对应一个直角边长为 $1/2$ 的小三角形，所以

$$P(X+Y\leq 1/2)=\int_0^{1/2}\int_0^{1/2-x} 2\mathrm{d}y\mathrm{d}x=\frac{1}{4}.$$

虽然 X,Y 各自的可能取值都在 $[0,1]$ 中，但它们的联合支撑是三角形，积分上限应随另一个变量变化，不能直接在整个单位正方形上积分.

我们下面给出一些具体的多元随机变量的例子：

例 2.3.6 (多元独立伯努利分布) 设 n 个相互独立的随机变量 $X_1,\ldots,X_n$ 均服从成功概率为 p 的伯努利分布，则它们可以构成一个多元独立伯努利分布，且它们的联合分布列即为每个分量的分布列的乘积：

$$f(x_1,\ldots,x_n)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i},\quad x_i\in\{0,1\}.$$

我们不难得出，$X_1+\cdots+X_n$ 服从二项分布. 这里随机变量相互独立的含义是，对任意取值范围 $B_1,\ldots,B_n$，事件 $\{X_i\in B_i\}$ 相互独立. 在离散情形下，这等价于联合分布列在所有取值组合上等于各分量分布列的乘积；在有联合密度的连续情形下，则等价于联合密度几乎处处等于各分量密度的乘积. 独立性还保证：若有关期望存在，则 $\mathbb{E}(XY)=\mathbb{E}X\,\mathbb{E}Y$，这可以通过将联合分布的乘积代入求和或积分公式得到.

例 2.3.7 (多维均匀分布) 设 $\mathcal{X}\in\mathbb{R}^n$ 为一个体积有限的区域，记其体积为 V, 则定义多维均匀分布的密度函数为

$$f(x_1,\ldots,x_n)=\frac{1}{V},\quad (x_1,\ldots,x_n)\in\mathcal{X}.$$

多维均匀分布就是几何概型在高维度下的随机变量表达.

我们下面介绍多元随机变量中最重要的实例——多元正态分布，先考虑二元的情形. 我们先来思考一下二元正态分布的联合密度函数可能是什么形式. 首先我们自然可以考虑两个正态分布随机变量相互独立的情况. 将两个随机变量的密度函数直接相乘，得到：

$$f(x,y)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\cdot\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

不过更一般的情形下，X,Y 并不独立. 我们的做法是引入一个表示二者之间关系的因子 $\rho\in(-1,1)$. 当 $\rho=\pm1$ 时分布退化，不再具有下面这种二维密度. 引入 ρ 后的二元正态分布的联合密度函数如下：

Yuan and Song

例 2.3.8 (二元正态分布) 定义服从二元正态分布的随机变量 X, Y 的密度函数为

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2}+\frac{\rho(x-\mu_1)(y-\mu_2)}{(1-\rho^2)\sigma_1\sigma_2}-\frac{(y-\mu_2)^2}{2(1-\rho^2)\sigma_2^2}\right\}.$$

可以看出，当 $\rho=0$ 时，就是我们上面给出的独立的情形. 而 $\rho\neq 0$ 时，归一化因子中出现了 $\sqrt{1-\rho^2}$, 同时指数中出现了交叉项，而交叉项的系数之所以如此设定，我们留待后续讨论边缘密度与条件密度时加以解决.

2.3.2 方差

我们知道，数学期望可以描述一个随机变量的整体平均特征，然而在有些场合，我们想要研究一个随机变量究竟偏离中心的程度如何，我们就需要引入一个全新的统计量.

一个直觉是，我们可以引入每个取值偏离期望的绝对值，并按概率加权，即

$$\sum_{i=1}^{\infty}p_i\left|x_i-\mathbb{E}X\right|,$$

但这样做引入了分段绝对值函数，这个函数具有许多不可导点，研究起来并不容易，因此我们转而研究另一个更加平滑的形式：

$$\sum_{i=1}^{\infty}p_i\left(x_i-\mathbb{E}X\right)^2.$$

从而我们引入了方差的定义，它可以同时适用于离散情形与连续情形.

定义 2.3.9 (方差) 对于一个随机变量 X 及其期望 $\mathbb{E}X$, 若

$$\mathbb{E}(X-\mathbb{E}X)^2$$

存在，则称其为随机变量 X 的方差（*variance*），记为 $\mathbb{D}X$ 或 $\mathrm{Var}(X)$. 同时称 $\sqrt{\mathbb{D}X}$ 为 X 的标准差（*standard deviation*）.

例 2.3.10 (均值相同，波动却不同) 两个随机变量 X, Y 各自以相同概率取两个值：X 取 *4* 和 *6*，Y 取 *0* 和 *10*. 它们的期望都是 *5*，但方差分别为

$$\mathbb{D}X=\frac{1}{2}(4-5)^2+\frac{1}{2}(6-5)^2=1,$$

$$\mathbb{D}Y=\frac{1}{2}(0-5)^2+\frac{1}{2}(10-5)^2=25.$$

对应标准差分别为 *1* 和 *5*. 因此，平均水平相同并不表示稳定程度相同；Y 的取值离均值更远，所以方差更大.

Xuan and Song

利用数学期望的线性性质，我们可以有如下推导：

$$\begin{aligned}\mathbb{D}X &= \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{E}\big(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2\big)\\ &= \mathbb{E}(X^2) - 2\mathbb{E}X\cdot\mathbb{E}X + (\mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2,\end{aligned}$$

这也是另一种计算方差的常用方式. 注意到 $\mathbb{E}(X^2)$ 实际上是随机变量 X 在函数 $y=x^2$ 作用下的新的随机变量的期望，而我们使用佚名统计学家公式就可以无需求新的随机变量的分布便可直接求得 $\mathbb{E}(X^2)$.

由于方差实际上也是一种期望，我们可以根据期望的性质直接得到方差的若干性质：

性质 2.3.11 *给定随机变量 X 及确定常数 c, 则我们有*

1. $\mathbb{D}c = \mathbb{E}(c-c)^2 = 0$.
2. $\mathbb{D}(X+c) = \mathbb{E}(X+c-\mathbb{E}X-c)^2 = \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{D}X$.
3. $\mathbb{D}(cX) = \mathbb{E}(cX-\mathbb{E}(cX))^2 = c^2\mathbb{E}(X-\mathbb{E}X)^2 = c^2\mathbb{D}X$.

例 2.3.12 (平移与缩放对方差的影响) *某随机误差 X 满足 $\mathbb{E}X=0$、$\mathbb{D}X=4$. 若仪器输出 $Y=3X+10$，则*

$$\mathbb{E}Y = 10, \qquad \mathbb{D}Y = 3^2\times 4 = 36, \qquad \sqrt{\mathbb{D}Y} = 6.$$

常数 10 只平移了读数，不改变波动大小；放大 3 倍使标准差放大 3 倍、方差放大 9 倍. 因此“方差也按 3 倍增长”是不正确的.

我们下面依然试着求一些随机变量的方差.

例 2.3.13 *考虑成功概率为 p 的伯努利分布 X, 其方差为*

$$\mathbb{D}X = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)$$

我们可以看出，p 越接近于 $0,1$ 中的某一个值，伯努利分布的方差越小，表示此时该随机变量更偏向一个确定性的结果，反之当 $p=1/2$ 时，方差达到最大，此时我们最难对结果做出预测.

Yuan and Song

例 2.3.14 考虑 $[a,b]$ 上的均匀分布 X, 首先易得其均值为 $(a+b)/2$, 随后方差计算为

$$\begin{aligned}\mathbb{D}X &= \int_a^b \left(x-\frac{a+b}{2}\right)^2 \cdot \frac{1}{b-a}\mathrm{d}x \\ &= \frac{1}{b-a}\left(\frac{x^3}{3}-\frac{a+b}{2}x^2+\frac{(a+b)^2}{4}x\Big|_a^b\right) \\ &= \frac{1}{b-a}\left(\frac{1}{12}b^3-\frac{1}{4}ab^2+\frac{1}{4}a^2b-\frac{1}{12}a^3\right) \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

我们最后再考虑正态分布，并指出其密度函数中的另一个参数 σ^2 正是它的方差.

例 2.3.15 考虑正态分布的方差，直接使用定义计算：

$$\begin{aligned}\mathbb{D}X &= \int_{-\infty}^{+\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \\ &\xlongequal{y=x-\mu} \int_{-\infty}^{+\infty} y^2\frac{1}{\sqrt{2\pi}\sigma}e^{-y^2/(2\sigma^2)}\mathrm{d}y \\ &= -\int_{-\infty}^{+\infty} y\frac{\sigma}{\sqrt{2\pi}}\mathrm{d}e^{-\frac{y^2}{2\sigma^2}} \\ &= -\frac{\sigma y}{\sqrt{2\pi}}e^{-\frac{y^2}{2\sigma^2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y \\ &= 0+\sigma^2=\sigma^2\end{aligned}$$

自此，正态分布 $N(\mu,\sigma^2)$ 中的两个参数的实际含义均已明晰，它们分别是正态分布的均值与方差.

2.3.3 边缘分布与条件分布

在本小节中，为了简便起见，我们均只考虑二元随机变量的情形，因为二元情形足以展示多元随机变量之间的相互关系，更多元的情形可以类推得到.

我们下面考虑二元随机变量的期望，它的期望自然也是一个二维向量. 而无论是级数还是积分，我们计算期望的公式得到的都只是一个标量，因此，考虑二元随机变量的期望需要每个分量单独进行考虑. 一个合理的思路是，考虑一个函数：$g(X,Y)=X$, 并使用佚名统计学家公式. 对于连续随机变量的情形，我们可以定义二元期望的第一个分量为

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)\mathrm{d}x\mathrm{d}y.$$

然而，我们从另一角度来想，二元随机变量中的第一个分量 X, 本身也是一个随机变量，它也有自己的分布与期望. 我们希望上述二元期望的第一个分量就恰好是 X 自身的期

Kuan and Song

望，这样就为我们研究二元多元随机变量带来了极大的方便. 而我们进一步观察上述式子即可得到：

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{+\infty} x\left(\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y\right)\mathrm{d}x,$$

式子中出现的 $\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y$ 是一个关于 x 的函数，它实际上正是 X 单独作为随机变量时的密度函数. 这样一来，二元期望的第一个分量就与 X 自己的期望是等价的了，而相应的 X 自身的密度函数我们称其为边缘密度，以便和联合密度区分开来.

离散情形的边缘分布与连续情形的推导基本类似，我们分别给出二者的边缘分布的定义：

定义 2.3.16 (二元离散随机变量的边缘分布) *给定二元离散随机变量 X,Y, 设其所有取值构成的集合为 $\mathcal{X}\times\mathcal{Y}$, 定义关于 x 的函数*

$$f_X(x) := \sum_{y\in\mathcal{Y}} P(X=x, Y=y), \quad x\in\mathcal{X},$$

为 X 的边缘分布. 同理可得 Y 的边缘分布.

我们从事件的分解角度也可以对上述定义给出合理的解释，Y 的不同取值可以作为一个完备事件组 $B_1, B_2, \ldots$, 而 $X=x$ 为事件 A, 则我们有

$$A = A\cap\Omega = A\cap(B_1+B_2+\cdots) = AB_1+AB_2+\cdots,$$

从而有

$$P(A) = P(AB_1)+P(AB_2)+\cdots.$$

因此，X 的边缘分布其实就是 $P(X=x)$.

定义 2.3.17 (二元连续随机变量的边缘密度与边缘分布) *给定二元连续随机变量 X,Y, 不妨设其取值范围为 $\mathbb{R}^2$, 已知其联合密度函数为 $f(x,y)$, 定义关于 x 的函数*

$$f_X(x) := \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y,$$

为 X 的边缘密度，同时定义 $F_X(x) := \int_{-\infty}^{x} f_X(t)\mathrm{d}t$ 为 X 的边缘分布. 同理可得 Y 的边缘密度与边缘分布.

在事件与概率的语境中，我们有条件概率 $P(B|A) = P(AB)/P(A)$, 我们想把它推广到两个随机变量的情形. 对于离散的随机变量，这一推广是比较自然的，我们可以定义 Y 关于 X 的条件分布为

$$f_{Y|X} = \frac{P(X=x,Y=y)}{P(X=x)} = \frac{f(x,y)}{f_X(x)}.$$

Yuan and Song

对于连续随机变量的情形，因为变量取一个特定的值的概率大部分情形下是 0，因此上述式子是可能需要从极限的角度去定义. 然而，我们可以直接形式地将联合密度与边缘密度相除得到条件密度，这实际上就是使用极限方式定义后推导得到的结果. 总之，无论离散情形还是连续情形，我们都可以使用二者的商来定义一个条件概率.

定义 2.3.18 (二元离散随机变量的条件分布) 给定二元离散随机变量 X,Y, 设其所有取值构成的集合为 $\mathcal{X}\times\mathcal{Y}$, 已知其联合分布列 $f(x,y)$ 与 X 的边缘分布列 $f_X(x)$, 对 $f_X(x)>0$ 的取值 x，定义

$$f_{Y|X}(y|x):=\frac{f(x,y)}{f_X(x)}=\frac{P(X=x,Y=y)}{P(X=x)}$$

为 Y 关于 X 的条件分布函数. 同理可得 X 关于 Y 的条件分布函数.

例 2.3.19 (对联合分布表求和，再按条件归一化) 仍考虑例2.3.2中的联合分布表. 按行求和，得到 X 的边缘分布；按列求和，得到 Y 的边缘分布：

$$P(X=0)=P(X=1)=\frac{1}{2},$$

$$P(Y=0)=\frac{1}{4},\qquad P(Y=1)=\frac{1}{2},\qquad P(Y=2)=\frac{1}{4}.$$

若已知 $X=1$，就在该行内重新归一化，得到

$$P(Y=0|X=1)=0,\quad P(Y=1|X=1)=\frac{1/4}{1/2}=\frac{1}{2},\quad P(Y=2|X=1)=\frac{1}{2}.$$

条件分布的三个概率之和仍为 *1*. 因为 $P(Y=2|X=1)=1/2\neq 1/4=P(Y=2)$，所以 X,Y 并不独立. 也可用 $P(X=0,Y=2)=0\neq(1/2)(1/4)$ 检验这一点.

定义 2.3.20 (二元连续随机变量的条件密度与条件分布) 给定二元连续随机变量 X,Y, 不妨设其取值范围为 $\mathbb{R}^2$, 已知其联合密度函数 $f(x,y)$ 与 X 的边缘密度函数 $f_X(x)$, 对 $f_X(x)>0$ 的取值 x，定义

$$f_{Y|X}(y|x):=\frac{f(x,y)}{f_X(x)}$$

为 Y 关于 X 的条件密度函数，同时定义 $F_{Y|X}=\int_{-\infty}^{y}f_{Y|X}(t|x)\mathrm{d}t$ 为 Y 关于 X 的条件分布函数. 同理可得 X 关于 Y 的条件密度函数与条件分布函数.

例 2.3.21 (三角形支撑上的边缘密度与条件密度) 仍考虑例2.3.5，联合密度在 D 上为 *2*. 对于 $0\leq x\leq 1$，固定 x 后，y 只能在 $[0,1-x]$ 内取值，所以

$$f_X(x)=\int_0^{1-x}2\mathrm{d}y=2(1-x).$$

同理，$f_Y(y)=2(1-y)$，$0\leq y\leq 1$；相应区间外的边缘密度均为 *0*. 因此在三角形内均匀取点，并不意味着每个坐标都在 $[0,1]$ 上均匀分布.

Yuan and Song

对于 $0 \le x < 1$，条件密度为

$$f_{Y|X}(y|x) = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad 0 \le y \le 1-x,$$

区间外为 0. 也就是说，给定 $X = x$ 后，Y 在 $[0, 1-x]$ 上均匀分布. 例如

$$P(Y > 1/4 \mid X = 1/2) = \int_{1/4}^{1/2} 2\mathrm{d}y = \frac{1}{2}.$$

这里使用的是条件密度，不能把它写成两个单点事件概率的比值，因为 $P(X = 1/2) = 0$.

在例2.3.8讨论二元正态分布时，我们留有一个悬念，即指数中交叉项的系数为什么如此设定. 我们下面来计算一下二元正态分布的边缘密度与条件密度，并指出上述系数的设定可以保证二者依然还是正态分布.

例 2.3.22 考虑 X 的边缘密度，即将联合密度函数中的 y 积分掉. 直接积分毫无疑问会非常复杂，我们需要进行巧妙的观察. 我们将指数中的所有项在此列出：

$$-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2} + \frac{\rho(x-\mu_1)(y-\mu_2)}{(1-\rho^2)\sigma_1\sigma_2} - \frac{(y-\mu_2)^2}{2(1-\rho^2)\sigma_2^2}.$$

我们看到，如果只关注 y, 那么指数中是一个关于 y 的二次函数，我们需要对其进行配方，就可以得到一个一元正态分布的形式. 具体来讲，y 的二次项系数是

$$-\frac{1}{2(1-\rho^2)\sigma_2^2},$$

一次项系数是

$$\frac{\mu_2}{(1-\rho^2)\sigma_2^2} + \frac{\rho(x-\mu_1)}{(1-\rho^2)\sigma_1\sigma_2} = \frac{1}{(1-\rho^2)\sigma_2}\left(\frac{\mu_2}{\sigma_2} + \frac{\rho(x-\mu_1)}{\sigma_1}\right),$$

因此 y 的对称轴为

$$\mu_2 + \frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}$$

同时需要配出的常数项是

$$-\frac{1}{2(1-\rho^2)}\left(\frac{\mu_2}{\sigma_2} + \frac{\rho(x-\mu_1)}{\sigma_1}\right)^2.$$

将剩余与 y 无关的部分，连同配出的常数项结合，有：

$$\begin{aligned}
&-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2} - \frac{\rho\mu_2(x-\mu_1)}{(1-\rho^2)\sigma_1\sigma_2} - \frac{\mu_2^2}{2(1-\rho^2)\sigma_2^2} + \frac{1}{2(1-\rho^2)}\left(\frac{\mu_2}{\sigma_2} + \frac{\rho(x-\mu_1)}{\sigma_1}\right)^2 \\
&= (\rho^2-1)\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2} = -\frac{(x-\mu_1)^2}{2\sigma_1^2},
\end{aligned}$$

余下的这正是 X 单独作为一元正态分布时指数中的项. 因此联合密度的指数可以做如下分解：

$$-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{1}{2(1-\rho^2)\sigma_2^2}\left(y - \mu_2 - \frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}\right)^2.$$

从而我们就可以很方便地求出关于 X 的边缘密度，因为可以利用正态分布的密度函数

Yuan and Song

将积分化为 *1*，即：

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\mu_2-\frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}\right)^2} \mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \end{aligned}$$

这恰好就是 X 自己作为正态分布时的密度函数. 进一步，我们也可以求得 Y 关于 X 的条件分布，事实上正是上述式子的后半部分，即：

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\mu_2-\frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}\right)^2},$$

这同样也是一个正态分布，是一个均值为 $\mu_2+\frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}$（与 x 有关），方差为 $(1-\rho^2)\sigma_2^2$ 的正态分布. 从而我们有结论：二元正态分布的边缘分布与条件分布都仍然是正态分布.

例 2.3.23 (给二元正态分布代入具体参数) 设 (X,Y) 服从二元正态分布，其中 $\mu_1=10$、$\mu_2=20$、$\sigma_1=2$、$\sigma_2=3$、$\rho=1/2$. 根据上例，X 的边缘分布为 $N(10,4)$，而给定 $X=12$ 后，Y 的条件均值与条件方差分别为

$$20+\frac{(1/2)\times 3}{2}(12-10)=21.5, \qquad (1-(1/2)^2)\times 3^2 = 6.75.$$

因此 Y 的条件分布为 $N(21.5,6.75)$. 相较于 Y 自身的边缘分布 $N(20,9)$，观测到高于均值的 X 后，对 Y 的预测均值上升，预测的方差减小.

在结束本小节之前，我们最后再来探讨另一个概念——条件期望（conditional expectation）. Y 关于 X 的条件期望就是将对应的条件分布作为新的概率去计算 Y 的期望，计算出的结果是一个与 X 取值有关的函数. 我们只给出连续情形的条件期望定义，离散情形的可以类似得出.

定义 2.3.24 (条件期望) 给定二元连续随机变量 X,Y，不妨设其取值范围为 $\mathbb{R}^2$，已知 Y 关于 X 的条件密度函数为 $f_{Y|X}(y|x)$，定义关于 x 的函数

$$\mathbb{E}(Y|X=x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\mathrm{d}y$$

为 Y 关于 X 的条件期望，同理可得 X 关于 Y 的条件期望.

条件期望的一个重要性质是，将 Y 关于 X 的条件期望作为一个新的随机变量，对它求期望就等于 Y 自身的期望. 其严格叙述如下：

性质 2.3.25 (重期望公式) 设 Y 关于 X 的条件期望为 $\mathbb{E}(Y|X=x)$，Y 自身的期望是 $\mathbb{E}Y$，

则有：

$$\mathbb{E}\big(\mathbb{E}(Y|X)\big) = \mathbb{E}Y.$$

证明. 我们注意到，$\mathbb{E}(Y|X=x)$ 是一个关于 x 的函数，那么 $\mathbb{E}(Y|X)$ 就是随机变量 X 在该函数作用下的一个新的随机变量，因此求它的期望依然使用的是佚名统计学家公式：

$$\begin{aligned}\mathbb{E}\big(\mathbb{E}(Y|X)\big) &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\mathrm{d}y\right) f_X(x)\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} y\left(\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{-\infty}^{+\infty} y f_Y(y)\mathrm{d}y = \mathbb{E}Y.\end{aligned}$$

■

例 2.3.26 (先按条件求平均，再对条件求平均) 在例2.3.5的三角形均匀分布中，我们已经知道给定 $X=x$ 后，Y 在 $[0,1-x]$ 上均匀分布，故

$$\mathbb{E}(Y|X=x) = \frac{1-x}{2}, \qquad \mathbb{E}(Y|X) = \frac{1-X}{2}.$$

前者在给定 x 时是一个数，后者随随机变量 X 变化，仍是一个随机变量. 利用重期望公式与 $f_X(x) = 2(1-x)$，有

$$\mathbb{E}Y = \int_0^1 \frac{1-x}{2}\,2(1-x)\mathrm{d}x = \int_0^1 (1-x)^2\mathrm{d}x = \frac{1}{3}.$$

用 Y 的边缘密度直接计算，也得到

$$\mathbb{E}Y = \int_0^1 y\,2(1-y)\mathrm{d}y = \frac{1}{3}.$$

两种方法给出相同结果，展示了“先在每个条件下平均，再按条件出现的概率平均”的过程.

我们下面使用二元正态分布来验证重期望公式的正确性. 我们已经求得 Y 关于 X 的条件分布为均值 $\mu_2 + \frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}$ 的正态分布，因此

$$\mathbb{E}(Y|X) = \mu_2 + \frac{\rho\sigma_2(X-\mu_1)}{\sigma_1},$$

我们根据期望的线性性质很容易得到

$$\mathbb{E}\big(\mathbb{E}(Y|X)\big) = \mathbb{E}\left(\mu_2 + \frac{\rho\sigma_2(X-\mu_1)}{\sigma_1}\right) = \mu_2 + \frac{\rho\sigma_2(\mu_1-\mu_1)}{\sigma_1} = \mu_2,$$

恰好为 Y 自己的均值 $\mathbb{E}Y$. 因此重期望公式得以验证.

在我们后续有关随机梯度下降算法的讨论中，重期望公式是一个很有利的工具，其帮助我们分步求取随机变量的均值.

Yuan and Song

2.3.4 协方差与相关系数

我们知道数学期望具有线性性，即 $\mathbb{E}(X+Y)=\mathbb{E}X+\mathbb{E}Y$, 然而方差是否也具有线性性呢？我们尝试将其展开：

$$\begin{aligned}\mathbb{D}(X+Y)&=\mathbb{E}\big(X+Y-\mathbb{E}(X+Y)\big)^2=\mathbb{E}\big(X-\mathbb{E}X+Y-\mathbb{E}Y\big)^2\\&=\mathbb{E}(X-\mathbb{E}X)^2+\mathbb{E}(Y-\mathbb{E}Y)^2+2\mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big)\\&=\mathbb{D}X+\mathbb{D}Y+2\mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big).\end{aligned}$$

可以看到，两个随机变量之和的方差拆成两个方差之和之后还多出来了一个交叉项，我们自然地将这一项定义为两个随机变量之间的协方差（covariance）.

定义 2.3.27 (协方差) *对于两个随机变量 X,Y 及各自的期望 $\mathbb{E}X,\mathbb{E}Y$, 记*

$$\operatorname{cov}(X,Y):=\mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big)$$

为二者的协方差.

协方差实际上是一个二元函数作用在 X,Y 二元随机变量上得到的新的随机变量的期望，计算方法依然是使用佚名统计学家公式. 我们下面以二元正态分布为例，计算两个变量之间的协方差. 首先，两个随机变量的各自的期望分别是 μ_1,μ_2, 随后我们使用上一小节使用过的分解式，得到：

$$\begin{aligned}&\operatorname{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)f(x,y)\mathrm{d}x\mathrm{d}y\\&=\int_{-\infty}^{+\infty}\frac{(x-\mu_1)}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\left(\int_{-\infty}^{+\infty}\frac{y-\mu_2}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\mu_2-\frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}\right)^2}\mathrm{d}y\right)\mathrm{d}x\\&=\int_{-\infty}^{+\infty}\frac{(x-\mu_1)}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\cdot\frac{\rho\sigma_2(x-\mu_1)}{\sigma_1}\mathrm{d}x\\&=\frac{\rho\sigma_2}{\sigma_1}\mathbb{E}(X-\mu_1)^2=\frac{\rho\sigma_2}{\sigma_1}\mathbb{D}X\\&=\frac{\rho\sigma_2}{\sigma_1}\sigma_1^2=\rho\sigma_1\sigma_2,\end{aligned}$$

因此二元正态分布中 X,Y 之间的协方差为 $\rho\sigma_1\sigma_2$.

回过头来，方差可以视为协方差的一种特例，即自己与自己的协方差. 进一步，我们在上一章中介绍的柯西-施瓦茨不等式可以从向量形式推广到离散随机变量分布列，乃至连续随机变量的场景，我们可以得出如下结论：

$$\big|\mathbb{E}(XY)\big|\le\sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)},$$

Yuanzhi Song

稍加转换即可得到

$$\big|\operatorname{cov}(X,Y)\big| \le \sqrt{\mathbb{D}X}\sqrt{\mathbb{D}Y},$$

从而我们可以定义如下的相关系数（correlation coefficient）：

定义 2.3.28 (相关系数) *对于两个随机变量 X,Y, 已知各自的方差 $\mathbb{D}X,\mathbb{D}Y$, 及协方差 $\operatorname{cov}(X,Y)$, 定义相关系数为*

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbb{D}X}\sqrt{\mathbb{D}Y}},$$

当 $\mathbb{D}X,\mathbb{D}Y$ 其中一个为 0 时，即其中一个随机变量为确定的常数，我们规定此时相关系数为 0.

因此，我们可以看到，二元正态分布中参数 ρ 的实际意义其实就是相关系数.

例 2.3.29 (从联合分布计算协方差与相关系数) *再考虑例2.3.2中的 X,Y. 由边缘分布可得*

$$\mathbb{E}X = \frac{1}{2},\quad \mathbb{E}Y = 1,\quad \mathbb{D}X = \frac{1}{4},\quad \mathbb{D}Y = \frac{1}{2}.$$

将联合分布表中的 xy 按概率加权，得到 $\mathbb{E}(XY) = 1\cdot 1\cdot(1/4) + 1\cdot 2\cdot(1/4) = 3/4$. 展开协方差的定义可得 $\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}X\,\mathbb{E}Y$，所以

$$\operatorname{cov}(X,Y) = \frac{3}{4} - \frac{1}{2}\times 1 = \frac{1}{4},\qquad \rho = \frac{1/4}{\sqrt{1/4}\sqrt{1/2}} = \frac{1}{\sqrt{2}}.$$

第一枚出现正面时，总正面数往往也更大，所以协方差为正. 若将 Y 替换为 $-Y$，协方差与相关系数都会变号. 若令 $Z = 2X+3$，则 $\operatorname{cov}(X,Z) = 2\mathbb{D}X = 1/2$、$\mathbb{D}Z = 1$，相关系数为 1，体现了完全的正线性关系.

例 2.3.30 (为什么独立测量取平均能减小方差) *设 $X_1,\dots,X_n$ 相互独立，每个变量的期望为 μ、方差为 σ^2. 独立性保证不同变量的协方差为 0. 将前面的方差展开式推广到 n 项，并令 $\bar{X} = (X_1+\cdots+X_n)/n$，得到*

$$\mathbb{E}\bar{X} = \mu,\qquad \mathbb{D}\bar{X} = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}X_i = \frac{\sigma^2}{n}.$$

例如单次测量的标准差为 2，4 次独立测量的平均值的标准差就是 $\sqrt{4/4} = 1$.

若只是把同一次测量 X 重复记录 4 遍，则平均值仍为 X，方差仍为 4. 减小方差需要利用测量之间的关系，不能只看记录条数. 我们后续研究小批量随机梯度时，也会用到这一思路.

Yuan and Song

例 2.3.31 (不相关不一定独立) 设 X 以相同概率取 $-1,0,1$，令 $Y=X^2$. 前一节已经算得 $\mathbb{E}X=0$、$\mathbb{E}Y=2/3$. 又因为

$$\mathbb{E}(XY)=\mathbb{E}(X^3)=\frac{-1+0+1}{3}=0,$$

所以 $\operatorname{cov}(X,Y)=0$，且 $\mathbb{D}X=2/3$、$\mathbb{D}Y=2/9$ 均为正，因此相关系数为 *0*. 但

$$P(Y=0|X=0)=1\neq\frac{1}{3}=P(Y=0),$$

故二者并不独立. 相关系数为 *0* 只排除了线性相关，不能排除 $Y=X^2$ 这样的非线性关系.

对于具有前述密度的二元正态分布，情形有所不同：当 $\rho=0$ 时，联合密度恰好分解为两个边缘密度的乘积，因此 X,Y 独立. 这个结论要求 (X,Y) 联合服从二元正态分布，不能推广到任意两个随机变量.

Yuan and Song

第二部分

最优化理论

Yuan and Song

第三章　凸优化理论

Yuan and Song

第四章　深度学习优化理论

4.1　神经网络的计算与性质

相比于传统凸优化理论与算法，深度学习问题中目标函数具有其独特的性质，其最显著的特征为：绝大多数神经网络都是高度非凸的[** 习题 **]. 其次，深度神经网络的结构往往十分复杂，梯度的计算需要合适的算法，来尽可能减少所需计算量，而前反向传播算法正是最标准的处理方式. 最后，由于神经网络高度非凸，诸如牛顿法等高阶的凸优化方法并不能带来更好的收益，因此最为基础的梯度下降法反而是最常用的优化算法. 在凸性缺失的条件下，光滑性往往是确保神经网络的训练收敛的基本假设，并指导梯度下降法的参数选取. 本节中我们将就上述问题展开探究.

4.1.1 前反向传播

我们首先考虑一个简易的三层线性神经网络：

$$\boldsymbol{z}_1 = W_1\boldsymbol{x}; \quad \boldsymbol{z}_2 = W_2\boldsymbol{z}_1; \quad \hat{\boldsymbol{y}} = W_3\boldsymbol{z}_2; \quad L = \ell(\hat{\boldsymbol{y}}, \boldsymbol{y}),$$

其中 $\boldsymbol{x} \in \mathbb{R}^m$ 为输入向量；$\boldsymbol{z}_1 \in \mathbb{R}^n, \boldsymbol{z}_2 \in \mathbb{R}^p$ 为中间值向量；$W_1 \in \mathbb{R}^{n\times m}, W_2 \in \mathbb{R}^{p\times n}, W_3 \in \mathbb{R}^{q\times p}$ 为三个待学习的参数矩阵；$\hat{\boldsymbol{y}} \in \mathbb{R}^q$ 为神经网络的输出向量，$\boldsymbol{y} \in \mathbb{R}^q$ 为实际的标签向量（二者也可以为标量，只需 $q = 1$ 即可）；$\ell(\cdot,\cdot)$ 为一个可求导的损失函数（其具体形式可以为平方损失，即 $\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{2}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$）.

在实际训练过程中，$(\boldsymbol{x}, \boldsymbol{y})$ 通常为一对给定的数据及相应的标签，而待学习的参数 W_1, W_2, W_3 则为我们关注的优化变量. 我们将目标函数表示为：

$$L = f(W_1, W_2, W_3) = \ell(W_3W_2W_1\boldsymbol{x}, \boldsymbol{y}),$$

我们的任务是计算上述目标函数关于 W_1, W_2, W_3 的分块偏导数，从而得到该函数整体

Yuan and Sun

的梯度并进行梯度下降. 我们首先考虑损失函数恰为平方损失的情形，即

$$L = \frac{1}{2}\|W_3W_2W_1\boldsymbol{x} - \boldsymbol{y}\|^2,$$

我们分别试求 $\dfrac{\partial L}{\partial W_1}, \dfrac{\partial L}{\partial W_2}, \dfrac{\partial L}{\partial W_3}$.

小贴士

在求解深度学习问题的分块偏导数时，我们将会反复使用到推论1.5.23, 读者可返回进行复习.

1. $\dfrac{\partial L}{\partial W_1}$: 当我们关注 W_1 时，W_2, W_3 将视为常数.
   首先关于 $\hat{\boldsymbol{y}} = W_3W_2W_1\boldsymbol{x}$, 有 $L = \frac{1}{2}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$, 我们可直接计算梯度为
   $$\frac{\partial L}{\partial \hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}} - \boldsymbol{y},$$
   注意此时分块偏导数记号 $\partial L/\partial \hat{\boldsymbol{y}}$ 实则表示全梯度，因为此时 $\hat{\boldsymbol{y}}$ 整合了未知量 W_1.
   随后考虑 $\boldsymbol{z}_1 = W_1\boldsymbol{x}$, 使用推论1.5.23, 有
   $$\frac{\partial L}{\partial \boldsymbol{z}_1} = (W_3W_2)^\top \frac{\partial L}{\partial \hat{\boldsymbol{y}}} = W_2^\top W_3^\top(\hat{\boldsymbol{y}} - \boldsymbol{y}),$$
   此时 $\partial L/\partial \boldsymbol{z}_1$ 同样为全梯度，在使用推论1.5.23时要始终确保外层函数的分块偏导数实际为全梯度，即外层函数之外不再包含我们关注的变量.
   再考虑 W_1, 再次使用推论1.5.23, 得
   $$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial \boldsymbol{z}_1}\boldsymbol{x}^\top = W_2^\top W_3^\top(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{x}^\top = W_2^\top W_3^\top(W_3W_2W_1\boldsymbol{x} - \boldsymbol{y})\boldsymbol{x}^\top.$$
2. $\dfrac{\partial L}{\partial W_2}$: 同样我们有 $\dfrac{\partial L}{\partial \hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}} - \boldsymbol{y}$, 其次考虑 $\boldsymbol{z}_2 = W_2\boldsymbol{z}_1$, 使用推论1.5.23, 得到
   $$\frac{\partial L}{\partial \boldsymbol{z}_2} = W_3^\top \frac{\partial L}{\partial \hat{\boldsymbol{y}}} = W_3^\top(\hat{\boldsymbol{y}} - \boldsymbol{y}).$$
   随后
   $$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \boldsymbol{z}_2}\boldsymbol{z}_1^\top = W_3^\top(W_3W_2W_1\boldsymbol{x} - \boldsymbol{y})\boldsymbol{x}^\top W_1^\top.$$
3. $\dfrac{\partial L}{\partial W_3}$: 同样以 $\dfrac{\partial L}{\partial \hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}} - \boldsymbol{y}$ 为基础，再次使用推论1.5.23,
   $$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial \hat{\boldsymbol{y}}}\boldsymbol{z}_2^\top = (W_3W_2W_1\boldsymbol{x} - \boldsymbol{y})\boldsymbol{x}^\top W_1^\top W_2^\top.$$

我们将三个分块偏导数的求解结果总结如下表：

[** 插入一个好看的表 **]

Yuan-Hao Song

小贴士

注意到，对于神经网络的目标函数，其自变量的形式为 (W_1, W_2, W_3), 是由三个形状不同的矩阵拼合成的变量结构. 相应地，其对应的梯度也可以写成一个相同形状的变量结构 $\left(\dfrac{\partial L}{\partial W_1}, \dfrac{\partial L}{\partial W_2}, \dfrac{\partial L}{\partial W_3}\right)$. 这些变量结构本质上等价于**其展平为向量形式**的变量结构，但是出于计算表示方便的考量，我们一般保持上述的复杂变量结构，而不将其展平为向量.

在求出了关于所有优化变量的分块偏导数后，我们思考如何安排计算的顺序，使得总的计算量最少. 首先，我们发现所有偏导数中均包含 $W_3W_2W_1\boldsymbol{x}-\boldsymbol{y}$ 这一因子，这实际上就是神经网络由输入得到输出的主体过程，我们称这一过程为**前向传播**（forward propagation）. 其次，我们发现三个偏导数的右侧依次出现了 $\boldsymbol{x}^\top$, $\boldsymbol{z}_1^\top = \boldsymbol{x}^\top W_1^\top$, $\boldsymbol{z}_2^\top = \boldsymbol{x}^\top W_1^\top W_2^\top$, 三者在前向传播过程中已经计算过了，可以免去重复计算，相应地我们需要将 $\boldsymbol{x}, \boldsymbol{z}_1, \boldsymbol{z}_2$ 的计算结果存储起来，以用于偏导数的计算，我们称这一操作为前向传播的**激活值存储**（activations' memory）. 最后，我们发现三个偏导数的左侧也会出现一些因子，且层数越靠前，相应的偏导数的因子就越多，因此一个合适的计算方式是：先计算 $d_3 = (W_3W_2W_1\boldsymbol{x}-\boldsymbol{y})$, 再计算 $d_2 = W_3^\top d_3$, 最后计算 $d_1 = W_2^\top d_2$. 这一计算顺序正好是与前向传播相反方向的计算顺序，因此我们称这一过程为**反向传播**（backward propagation），在计算过程中，d_3, d_2 等通常为关于激活值的导数，需要供接续的计算使用，因此同样需要存储，但是由于下一个激活值导数存储前，前一个激活值导数可以使用后直接释放，所以其带来的存储开销是常数级的，相比前文的激活值存储可以忽略.

综上所述，我们得到一个最佳的计算顺序为：

1. **前向传播**：
   (a) 存储 $\boldsymbol{x}$;
   (b) 计算并存储 $\boldsymbol{z}_1 = W_1\boldsymbol{x}$;
   (c) 计算并存储 $\boldsymbol{z}_2 = W_2\boldsymbol{z}_1$;
   (d) 计算并存储 $\hat{\boldsymbol{y}} = W_3\boldsymbol{z}_2$;
   (e) 计算 $L = \frac{1}{2}\|\hat{\boldsymbol{y}}-\boldsymbol{y}\|^2$, 如不需得到损失具体数值，可省略此步骤.
2. **反向传播**：
   (a) 计算并存储激活值导数 $\dfrac{\partial L}{\partial \hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}}-\boldsymbol{y}$, 释放 $\hat{\boldsymbol{y}}$ 的内存；
   (b) 计算偏导数 $\dfrac{\partial L}{\partial W_3} = \dfrac{\partial L}{\partial \hat{\boldsymbol{y}}}\boldsymbol{z}_2^\top$, 释放 $\boldsymbol{z}_2$ 内存；
   (c) 计算并存储激活值导数 $\dfrac{\partial L}{\partial \boldsymbol{z}_2} = W_3^\top \dfrac{\partial L}{\partial \hat{\boldsymbol{y}}}$, 释放 $\dfrac{\partial L}{\partial \hat{\boldsymbol{y}}}$ 内存；

(d) 计算偏导数 $\dfrac{\partial L}{\partial W_2} = \dfrac{\partial L}{\partial \boldsymbol{z}_2}\boldsymbol{z}_1^\top$, 释放 $\boldsymbol{z}_1$ 内存；

(e) 计算并存储激活值导数 $\dfrac{\partial L}{\partial \boldsymbol{z}_1} = W_2^\top \dfrac{\partial L}{\partial \boldsymbol{z}_2}$, 释放 $\dfrac{\partial L}{\partial \boldsymbol{z}_2}$ 内存；

(f) 计算偏导数 $\dfrac{\partial L}{\partial W_1} = \dfrac{\partial L}{\partial \boldsymbol{z}_1}\boldsymbol{x}^\top$, 释放 $\boldsymbol{x}$ 内存.

以上即为神经网络梯度计算的**前反向传播算法**（forward backward propagation）. 由于反向传播相比前向传播通常需要额外计算一系列关于激活值的导数，粗略来讲，**反向传播的计算量大致为前向传播的二倍**. [** 一个计算量习题 **]

4.1.2 激活函数及其计算

上一小节我们考虑了一个简单的线性网络 $\boldsymbol{y} = W_3W_2W_1\boldsymbol{x}$. 然而，如果一个神经网络中只有线性层的话，我们自然可以取一个整合的矩阵 $W = W_3W_2W_1$ 使得 $\boldsymbol{y} = W\boldsymbol{x}$, 无论网络的层数是多少均是如此. 因此，如果只有线性层，无论多少层，神经网络均只能表示最简单的线性关系. 为了解决这一问题，就必须引入非线性的操作. 神经网络的设计者们通常是这么做的：引入**每个元素可分**的**一致**的非线性一元函数作为非线性操作，以求在非线性前提下尽可能保持问题足够简单. 所谓每个元素可分且一致的操作即为：

$$\boldsymbol{y} = \boldsymbol{\sigma}(\boldsymbol{x}) = \big(\sigma(x_1), \sigma(x_2), \dots, \sigma(x_n)\big)^\top,$$

即对输入向量的每个元素实施相同的一元函数操作，并输出一个相同形状的向量. 最常见的用于神经网络的一元非线性函数为**逻辑斯谛**（Logistic）函数，或称 Sigmoid 函数，其形式如下：

$$\sigma(x) = \frac{1}{1+e^{-x}},$$

它是定义域为 $\mathbb{R}$, 值域为 $(0,1)$ 的单调递增函数，其图像如下：[** 插入一个图 **] 我们易求出它的导数为

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x)\big(1-\sigma(x)\big).$$

这类每个元素可分的一致的非线性函数，在深度学习中一般称为**激活函数**（activation functions），意指输入的数值经过这一函数就激活了非线性的性质. 相应地，激活函数输出的值被称为激活值（activation value / activations）；然而，当下激活值的概念发生了泛化，人们倾向于将所有前向传播时需要存储的值均称为激活值.

我们下面来考虑一个引入激活函数后的三层神经网络的例子：

$$\boldsymbol{h}_1 = W_1\boldsymbol{x}; \quad \boldsymbol{z}_1 = \boldsymbol{\sigma}(\boldsymbol{h}_1); \quad \boldsymbol{h}_2 = W_2\boldsymbol{z}_1; \quad \boldsymbol{z}_2 = \boldsymbol{\sigma}(\boldsymbol{h}_2); \quad \hat{\boldsymbol{y}} = W_3\boldsymbol{z}_2; \quad L = \ell(\hat{\boldsymbol{y}}; \boldsymbol{y}),$$

其中 $\boldsymbol{x} \in \mathbb{R}^m$ 为输入向量；$\boldsymbol{h}_1 \in \mathbb{R}^n, \boldsymbol{h}_2 \in \mathbb{R}^p$ 为激活前的中间值向量；$\boldsymbol{z}_1 \in \mathbb{R}^n, \boldsymbol{z}_2 \in \mathbb{R}^p$

为激活后的中间值向量；$W_1 \in \mathbb{R}^{n\times m}, W_2 \in \mathbb{R}^{p\times n}, W_3 \in \mathbb{R}^{q\times p}$ 为三个待学习的参数矩阵，$\hat{\boldsymbol{y}} \in \mathbb{R}^q$ 为神经网络的输出向量，$\boldsymbol{y} \in \mathbb{R}^q$ 为实际的标签向量，$\ell(\cdot,\cdot)$ 为一个可求导的损失函数.

我们省略前向传播及激活值存储步骤的叙述，直接进入反向传播的环节. 首先，$\dfrac{\partial L}{\partial \hat{\boldsymbol{y}}}, \dfrac{\partial L}{\partial W_3}, \dfrac{\partial L}{\partial \boldsymbol{z}_2}$ 的计算依然不变. 而由于加入了激活的操作，我们需要考虑当 $\boldsymbol{z}_2 = \boldsymbol{\sigma}(\boldsymbol{h}_2)$ 时，如何计算 $\dfrac{\partial L}{\partial \boldsymbol{h}_2}$, 这时我们需要使用最基础的向量值函数的**链式法则**（定理1.3.10），同时需要尤其注意，使用链式法则时要使用多元函数的**行梯度**. 在该问题中，内层函数即为 $\boldsymbol{z}_2 = \boldsymbol{\sigma}(\boldsymbol{h}_2)$, 外层函数为 $L = f(\boldsymbol{z}_2) = \ell(W_3\boldsymbol{z}_2; \boldsymbol{y})$. 记复合函数 $g(\boldsymbol{h}_2) = f\big(\boldsymbol{\sigma}(\boldsymbol{h}_2)\big)$, 则由链式法则可以得到：

$$g'(\boldsymbol{h}_2) = f'\big(\boldsymbol{\sigma}(\boldsymbol{h}_2)\big)\,\boldsymbol{\sigma}'(\boldsymbol{h}_2),$$

其中 $f'\big(\boldsymbol{\sigma}(\boldsymbol{h}_2)\big) = f'(\boldsymbol{z}_2)$ 是一个行梯度，而且恰为 $\left(\dfrac{\partial L}{\partial \boldsymbol{z}_2}\right)^\top$. $\boldsymbol{\sigma}'(\boldsymbol{h}_2)$ 为一个雅各比矩阵，即 $J_{\boldsymbol{\sigma}}(\boldsymbol{h}_2)$. $g'(\boldsymbol{h}_2)$ 同样是一个行梯度，且即是我们关心的 $\left(\dfrac{\partial L}{\partial \boldsymbol{h}_2}\right)^\top$. 我们将链式法则的左右两侧做转置，以求得列梯度如下：

$$\frac{\partial L}{\partial \boldsymbol{h}_2} = J_{\boldsymbol{\sigma}}(\boldsymbol{h}_2)^\top \frac{\partial L}{\partial \boldsymbol{z}_2}.$$

最后，由于 $\boldsymbol{\sigma}(\cdot)$ 是一个逐元素可分一致的函数，它的雅各比矩阵实际为一个对角矩阵：

$$J_{\boldsymbol{\sigma}}(\boldsymbol{h}_2) = \begin{pmatrix} \sigma'(h_{2,1}) & & \\ & \ddots & \\ & & \sigma'(h_{2,p}) \end{pmatrix}.$$

在深度学习中，我们不严格地记记号 $\sigma'(\boldsymbol{h}_2) = \big(\sigma'(h_{2,1}), \ldots, \sigma'(h_{2,p})\big)^\top$, 即将上述雅各比矩阵中的所有元素排列为了一个列向量. 并且由于一个对角矩阵左乘一个向量，就等价于将对角矩阵的所有元素与右侧的向量对应元素分别相乘，构成一个新的向量，这即为向量的逐元素积，或称哈达玛积. 因此上述求导步骤的最简易写法为：

$$\frac{\partial L}{\partial \boldsymbol{h}_2} = \sigma'(\boldsymbol{h}_2) \odot \frac{\partial L}{\partial \boldsymbol{z}_2}.$$

在求得 $\dfrac{\partial L}{\partial \boldsymbol{h}_2}$ 后，其他后续的反向传播步骤便可按照先前的步骤类似推导出，我们在此不再赘述. [** 给一个习题求出前反向全过程 **]

在结束本小节之前，我们最后再讨论一下逻辑斯谛函数的两个性质.

其一、对于逻辑斯谛函数 $y = \sigma(x)$, 我们知道其导数等于

$$\sigma'(x) = \sigma(x)\big(1-\sigma(x)\big) = y(1-y),$$

Yuan and Song

也就是说，逻辑斯谛函数的导数可以由函数值或者说**由因变量来完全表示**，这在前反向传播过程中是具有一定优势的. 由于逻辑斯谛函数的自变量在其相应的反向传播步骤中可以被因变量替代，如果在其余位置同样也不需要使用到自变量，那么自变量在前向传播时便可不必存储. 具体来讲，对于上述三层神经网络的情形，$\boldsymbol{z}_2 = \boldsymbol{\sigma}(\boldsymbol{h}_2), \sigma'(\boldsymbol{h}_2)$ 可以用 $\boldsymbol{z}_2(1-\boldsymbol{z}_2)$ 完全表示，并且 $\boldsymbol{h}_2$ 在其他步骤中也没有被使用，因此 $\boldsymbol{h}_2$ 在前反向传播过程中可以不必存储，即在计算完 $\boldsymbol{z}_2$ 后便可释放，$\boldsymbol{h}_1$ 同理.

其二、观察逻辑斯谛函数的图像我们会发现，在偏离原点一定的距离后，函数的图像会变得非常平坦，即函数的导数非常接近于零. 而由于反向传播求导数的过程中，逻辑斯谛函数的导数值与其他量之间是逐元素乘积的关系，因此过小的逻辑斯谛函数导数值会导致反向传播时的值十分接近于零. 鉴于此，我们通常讲**逻辑斯谛函数易导致梯度消失问题**，即求出的整体梯度模长很小，使用梯度下降法时，迭代点基本不发生移动. 为解决这一问题，深度学习工作者们提出了一系列改进的激活函数. ReLU（Rectified Linear Unit）函数为最具代表性的一个，其表达式为：

$$y = \max\{x, 0\} = \begin{cases} x, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

其部分地解决了梯度消失问题，且因为计算更加简便而被广泛地使用.

4.1.3 光滑性

在先前的探讨中，为了使得梯度更加具有可解释性且更加便于计算使用，我们保留了梯度的与神经网络参数相一致的复杂变量结构. 而在之后的数学原理的探究中，我们便直接使用展平的向量形式的目标函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, 以简化符号表示.

我们在本节最开始指出，深度学习中神经网络的目标函数都是高度非凸的. 在凸性缺失的情况下，一个最大的问题即为：使用梯度下降等算法可能无法收敛至全局的最优解，而只能收敛至一个局部的最优解.

[** 画一张图 **]

在凸性缺失后，我们自然要退而求其次，寻找其他可能的较好的性质，以求能够使用合适的算法求得较好的收敛结果.

我们首先从直观上可以考虑到，函数 $f(\boldsymbol{x})$ 的函数值变化不应该太剧烈，这样的话我们在使用梯度下降等算法进行迭代时，函数值的变化能够处于可控的范围内. 这一性质即函数的**李普希茨连续性**（Lipschitz continuity），其严格定义如下：

定义 4.1.1 (李普希茨连续性) *对于多元函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, 如果对于定义域 X 上的任*

意两点 $\boldsymbol{x}, \boldsymbol{y}$, 都有

$$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq G\|\boldsymbol{y} - \boldsymbol{x}\|,$$

则称函数 $f(\boldsymbol{x})$ 在定义域 X 上是 G-李普希茨连续的.

对于李普希茨连续性，我们可以得到一个它的基本结论如下：

性质 4.1.2 若多元函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ 在定义域 X 上是可微函数，且是 G-李普希茨连续的，则有：

$$\|\nabla f(\boldsymbol{x})\| \leq G, \quad \forall \boldsymbol{x} \in X.$$

证明. 对于任意 $\boldsymbol{x} \in X$, 我们取任意一个模长为 1 的单位向量 $\boldsymbol{v}$ 以及任意 $t > 0$, 由可微性可得：

$$f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x}) = t\nabla f(\boldsymbol{x})^\top \boldsymbol{v} + o(t\|\boldsymbol{v}\|) = t\nabla f(\boldsymbol{x})^\top \boldsymbol{v} + o(t).$$

由 G-李普希茨连续性得：

$$|f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})| \leq G\|t\boldsymbol{v}\| = tG.$$

将两式整合并同时除以 t 得：

$$\left|\nabla f(\boldsymbol{x})^\top \boldsymbol{v} + o(1)\right| \leq G,$$

令 $t \to 0$ 可直接消去 $o(1)$ 项，进一步，

$$|\nabla f(\boldsymbol{x})^\top \boldsymbol{v}| = \|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{v}\| \cdot |\cos\theta| \leq G,$$

其中 $\|\boldsymbol{v}\| = 1, \cos\theta$ 为两向量夹角的余弦值，由于 $\boldsymbol{v}$ 的任意性，其可以取值为 1，因此为保证上式成立，我们必有

$$\|\nabla f(\boldsymbol{x})\| \leq G.$$

■

因此，一个可微的目标函数如果需要满足李普希茨连续性，那么它的梯度就必须是**有界的**. 这是一个很强的条件，以至于最常见的二次函数 $y = x^2$ 也不满足这一条件. 因此在深度学习中，我们通常考虑另一个能够被更多种函数满足的条件：我们假设目标函数可微，并将原先有关函数值的限制转为对梯度的限制上，就得到了**梯度李普希茨连续性**，或称为**（梯度）李普希茨光滑性**. 由于该条件十分常用，我们常常简称其为**光滑性**（smoothness）. 光滑性的严格定义如下：

定义 4.1.3 (梯度李普希茨连续性/光滑性) 对于可微的多元函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, 如果对

于定义域 X 上的任意两点 $\boldsymbol{x}, \boldsymbol{y}$, 都有

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|,$$

则称函数 $f(\boldsymbol{x})$ 在定义域 X 上是 L-李普希茨光滑的.

小贴士

在数学分析中，“光滑函数”通常指具有无限阶连续导数的函数，记为 C^{∞}. 其确保了函数可以求无限次导数. 而在深度学习领域“光滑性”特指满足上述性质的函数. C^{∞} 的函数不一定是 L-李普希茨光滑的，且 L-李普希茨光滑函数也不一定是 C^{∞} 函数.

与性质4.1.2类似，如果假设函数二阶可微，光滑性也可以得到海瑟矩阵有界的性质，由于该性质的证明需要使用到二阶可微的性质及谱范数的性质，我们略去其证明.

性质 4.1.4 若多元函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ 在定义域 X 上二阶可微，且是 L-李普希茨光滑的，则有：

$$\|\nabla^2 f(\boldsymbol{x})\|_2 \leq L, \quad \forall \boldsymbol{x} \in X,$$

其中 ∇^2 表示海瑟矩阵，$\|\cdot\|_2$ 为矩阵的 2-算子范数或谱范数.

相比于光滑性的原始定义，我们更常用的是其如下的性质，多数优化算法的收敛性分析均以以下光滑性的性质为出发点.

性质 4.1.5 (光滑性的二次上界) 若多元函数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ 在定义域 X 上是 L-光滑的，则对任意 $\boldsymbol{x}, \boldsymbol{y} \in X$, 有

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

证明. 构造以 $f(\boldsymbol{x}), f(\boldsymbol{y})$ 为端点的辅助函数：

$$g(t) = f\big(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\big), \quad t \in [0, 1],$$

则 $g(0) = f(\boldsymbol{x}), g(1) = f(\boldsymbol{y})$. $g(t)$ 为一个一元函数，我们可以使用复合函数的链式法则1.3.10对其求导，得到：

$$g'(t) = \nabla f\big(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\big)^{\top}(\boldsymbol{y} - \boldsymbol{x}).$$

使用牛顿-莱布尼茨公式：

$$g(1) - g(0) = \int_0^1 g'(t)\mathrm{d}t,$$

即

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \nabla f\big(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})\big)^\top (\boldsymbol{y}-\boldsymbol{x}) \mathrm{d}t \\ &= \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \int_0^1 \Big(\nabla f\big(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\Big)^\top (\boldsymbol{y}-\boldsymbol{x}) \mathrm{d}t \\ &\leq \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \int_0^1 \Big\|\nabla f\big(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})\big) - \nabla f(\boldsymbol{x})\Big\| \|\boldsymbol{y}-\boldsymbol{x}\| \mathrm{d}t \\ &\leq \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \int_0^1 Lt\|\boldsymbol{y}-\boldsymbol{x}\|^2 \mathrm{d}t \\ &= \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \end{aligned}$$

■

性质4.1.5表示，以函数 $f(\boldsymbol{x})$ 上的任意一点 $\boldsymbol{x}$ 为基点，其函数值变化均存在一个二次函数的上界，如图[** 图 **] . 当然也对应地存在一个负的二次函数下界，只是优化理论中更多感兴趣上界那部分的性质，我们将另一半性质留作习题[** 习题 **] .

[** 插入图 **]

最后，我们来探讨光滑性这一条件是否已经足够适用于绝大多数的深度学习优化问题. 首先，对于第一小节中的纯线性神经网络，它实际上是一个关于三个变量的六次函数，它的梯度中三个分量都是五次函数，因此不能被一个二次函数所限制，所以不是光滑函数. 而对于第二小节的添加了激活函数的神经网络，由于 Sigmoid 函数的值域及其导数都是有界的，这会使得梯度的计算也倾向于有界，事实上，加入了 Sigmoid 激活函数的神经网络是光滑的，但是严格证明这一点确实不太容易. 而对于更复杂的神经网络，我们就只能从十分粗略的定性角度去解释了. 一般来讲，神经网络不要出现连续三个及以上的线性层，且非线性层要满足李普希茨连续性，或者最好是有界的，那么整体神经网络作为待学习参数的目标函数就大概率是光滑函数. 进一步，在实际问题中，我们通常假设待学习参数保持在一定的界限内，即神经网络的目标函数定义域是一个有界的闭集 X. 在此情形下，如果加入若干的平滑措施来抹除部分的尖锐点，那么我们总能找到一个充分大的 L, 使得在 X 内，目标函数是 L-光滑的. 综上所述，对于深度学习的优化任务，**我们一般假设目标函数均满足光滑性条件**，并进行相应的收敛性分析，这也是我们下一节要讨论的内容.

Yuan and Song

4.2　梯度下降算法的收敛性

我们在很早便接触了梯度下降算法. 而本节中，我们将会对梯度下降算法的性质进行详细的探究. 我们首先考虑非凸目标函数的情形，这适用于大多数的神经网络，随后再加入凸性以及强凸性的条件，并说明凸性的引入提供了更好的收敛性保证.

4.2.1 非凸情形的收敛性

在研究一个迭代序列 $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots$ 的收敛性的时候，一个首先需要讨论的问题是，我们如何刻画一个序列收敛的好坏呢？我们首先想到了可以使用序列中每个点到最优解的距离 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ 来衡量收敛性，如果 $\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\| = 0$, 则说明算法成功收敛到了最优解，且序列 $\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|\}_{k=1}^{\infty}$ 趋近于 0 的速率越快，就表明算法收敛的速率越快.

然而，有些目标函数的最优解并不唯一，或者并不是一个有限值（如 e^{-x}），那么此时 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ 可能不能很好地定义. 我们这时退而求其次，考虑目标函数的最优值 f^*. 只要目标函数有下界，那么它的最优值一定存在（即 $f(\boldsymbol{x}^*)$）或是在极限的意义下存在. 我们进而可以使用 $f(\boldsymbol{x}_k) - f^*$ 来代替 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ 作为收敛指标，来衡量一个算法是否收敛，以及收敛的速率如何.

但是更进一步，非凸函数的图像千沟万壑[** 图片 **]，梯度下降算法绝大多数时候只能找到一个局部的极小值点，而无法找到全局的最小点，即 $f(\boldsymbol{x}_k) - f^*$ 可能不一定能收敛到 0，我们需要进一步更换一个收敛指标. 根据微积分的知识我们知道，在多元函数图像中有一些特殊的点，即极小值点、极大值点和鞍点（在鞍点处，有些方向上看是极大值点，而在另一些方向上看则是极小值点）[** 图片 **]，它们统称为驻点. 判定 $\boldsymbol{x}^*$ 为驻点的依据为 $\nabla f(\boldsymbol{x}^*) = 0$, 即梯度等于 0 向量. 因此我们可以使用 $\|\nabla f(\boldsymbol{x}_k)\|$ 来衡量迭代点是否足够接近一个驻点，从而可以进一步讨论算法的收敛速率.

值得注意的是，虽然对于非凸函数，我们很多时候只能找到一个驻点，但是除了极小值点外，极大值点与鞍点在梯度下降算法中一般是不会作为收敛点而存在的. 因为一旦给予迭代点以某一方向上的扰动，迭代点就会快速脱离驻点重新进发. 因此，在实际问题中，梯度下降算法收敛至的驻点总会是一个局部的极小值点.

我们下面正式来讨论在非凸情形下，梯度下降算法的收敛性.

首先，一个算法如果能够收敛，它一般需要满足迭代点的函数值单调下降，即 $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_1) \geq f(\boldsymbol{x}_2) \geq \cdots$. 我们可以证明，当函数满足光滑性时，若步长满足

一定的条件，我们能够得到目标函数单调下降，即下述引理.

引理 4.2.1 (下降引理) 设目标函数 $f(\boldsymbol{x})$ 为 L-光滑函数，使用梯度下降法，若步长 γ 满足 $\gamma \leq 2/L$, 则有

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k).$$

进一步，若 $\gamma \leq 1/L$, 则有

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{\gamma}{2}\|\nabla f(\boldsymbol{x}_k)\|^2.$$

证明. 取 $\boldsymbol{x}_k, \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma\nabla f(\boldsymbol{x}_k)$ 代入光滑性的二次上界中：

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \left\langle \nabla f(\boldsymbol{x}_k), -\gamma\nabla f(\boldsymbol{x}_k) \right\rangle + \frac{L}{2}\left\| -\gamma\nabla f(\boldsymbol{x}_k)\right\|^2 \\ &= f(\boldsymbol{x}_k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2. \end{aligned}$$

当 $\gamma \leq 2/L$ 时，$1 - L\gamma/2 \geq 0$, 此时 $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$ 成立.

进一步，当 $\gamma \leq 1/L$ 时，$1 - L\gamma/2 \geq 1/2$, 此时我们有

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{\gamma}{2}\|\nabla f(\boldsymbol{x}_k)\|^2$$

成立. ■

下降引理告诉我们，光滑函数保证了函数值在局部不会变化得过于剧烈，因此如果我们取足够小的步长，一定能够保证迭代是朝着函数值减少的方向进行的.

我们先前曾指出，$\|\nabla f(\boldsymbol{x}_k)\|$ 应当作为非凸情形下的收敛指标，我们应当如何使用上述的下降引理，来得到有关 $\|\nabla f(\boldsymbol{x}_k)\|$ 的收敛结论呢？我们注意到，下降引理中第二个不等式的右侧出现了与 $\|\nabla f(\boldsymbol{x}_k)\|$ 有关的项，它就是推导收敛性结论的关键. 而不等式其余的两项则是相邻两次迭代的函数值分立不等式两侧，对于这种形式，我们可以写出有关 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ 的 k 个相同形式的不等式，然后将其加和并消去中间的项，只留下首尾两项，这种做法一般称为叠加法（telescoping）. 以下便是非凸情形下梯度下降法的收敛性结论：

定理 4.2.2 (非凸情形梯度下降法收敛性) 设目标函数 $f(\boldsymbol{x})$ 为 L-光滑函数，且存在下界 f^*, 使用梯度下降法迭代 K 步，若取步长 $\gamma \leq 1/L$, 则有

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2\big(f(\boldsymbol{x}_0) - f^*\big)}{\gamma K}.$$

进一步，若恰好取步长 $\gamma = 1/L$, 则有

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2L\big(f(\boldsymbol{x}_0) - f^*\big)}{K}.$$

Yuan Song

证明. 我们依次对 $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{K-1}$ 使用下降引理，得到

$$\frac{\gamma}{2}\|\nabla f(\boldsymbol{x}_0)\|^2 \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_1)$$
$$\frac{\gamma}{2}\|\nabla f(\boldsymbol{x}_1)\|^2 \leq f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)$$
$$\cdots$$
$$\frac{\gamma}{2}\|\nabla f(\boldsymbol{x}_{K-1})\|^2 \leq f(\boldsymbol{x}_{K-1}) - f(\boldsymbol{x}_K),$$

叠加可得：

$$\frac{\gamma}{2}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_K).$$

不等式两侧同时乘以 $2/(\gamma K)$, 同时注意到 $f(\boldsymbol{x}_K) \geq f^*$, 我们有

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2\big(f(\boldsymbol{x}_0) - f^*\big)}{\gamma K}.$$

特别地取 $\gamma = 1/L$, 我们有

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2L\big(f(\boldsymbol{x}_0) - f^*\big)}{K}.$$

■

该定理中的不等式左侧为 $\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(\boldsymbol{x}_k)\|^2$, 即迭代前 K 步的梯度模长平方的平均值. 相较于只关注最后一次迭代的收敛指标，我们称收敛指标求平均为遍历性（ergodic）的收敛判定方式. 在较弱的条件下，每次迭代的结果可能存在波动，我们难以直接衡量最后一次迭代的收敛性，因此遍历的收敛指标便为我们提供了更加稳定的收敛性判定方式.

事实上，大多数收敛性定理的核心不等式形式均为：不等式取小于等于号，**左侧是一个收敛指标**，可以是只关注最后一次迭代的收敛指标，也可以是遍历性的收敛指标；**右侧是一个与迭代次数 K 有关的式子**. 在数学中，我们一般使用同阶表示法来表示不同序列收敛速率的关系. 即对于两个序列 $\{a_k\}_{k=1}^{\infty}, \{b_k\}_{k=1}^{\infty}$, 我们记 $a_k = \mathcal{O}(b_k)$, 如果存在一个常数 $C > 0$, 使得

$$|a_k| \leq C|b_k|, \quad \forall k = 1, 2, \ldots.$$

在收敛性问题中，$\{a_k\}$ 是我们的收敛指标，而 $\{b_k\}$ 并不一定是不等式右侧的全部内容，我们可以根据需要，将一些常数放入常数 C 中，来简化我们的表示.

具体来讲，上述收敛结论中右侧的因子中出现了 $1/K$, 这便是我们感兴趣的最简单形式的收敛速率. 我们通常称非凸情形下梯度下降的收敛速率是

$$\mathcal{O}(1/K),$$

Kuan and Song

即它的收敛速率至少与调和序列 $\{1/k\}_{k=0}^{\infty}$ 是同阶次的. 当然我们也可以进一步考虑一些常数，譬如考虑步长，则收敛速率是 $\mathcal{O}\big(1/(\gamma K)\big)$ 因此步长越小，收敛速率就会越慢；如果取 $\gamma = 1/L$，此时收敛速率可以是 $\mathcal{O}(L/K)$, 因此光滑性系数越大，收敛速率就越慢.

除了使用收敛指标的同阶次序列描述收敛速率外，有时我们也使用算法复杂度来描述收敛速率：即给定一个接近于 0 的值 ϵ, 我们想要让收敛指标达到 ϵ 的精度，我们至少需要多少轮迭代. 我们的做法是令

$$\epsilon = \mathcal{O}\left(\frac{1}{K}\right),$$

解得

$$K = \Omega\left(\frac{1}{\epsilon}\right),$$

其中 Ω 表示另一个方向的同阶性，即 $a_k = \Omega(b_k) \Leftrightarrow \exists C > 0, \forall k, |a_k| \geq C|b_k|$，换句话说，迭代次数 K 至少是 $1/\epsilon$. 这就是算法复杂度的描述方式. 然而，由于多数情况下我们最先得到的均是有关收敛指标的同阶次序列的结论，因此我们一般讲的收敛速率均指收敛指标的同阶次序列，除非明确声明我们讨论的是算法复杂度.

4.2.2 凸情形的收敛性

我们下面考虑目标函数是凸函数且满足 L-光滑性的情形. 在讨论收敛性之前，我们首先来关注光滑凸函数的性质. 首先第一条性质是凸性与光滑性两个性质的直接叠加.

性质 4.2.3 (上下界) *设函数 $f(\boldsymbol{x})$ 为可导的凸函数，且满足 L-光滑性，则有：*

$$0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

而下述的第二条性质证明起来并不直接，且不会直接用到光滑凸函数梯度下降的收敛性证明中，但其为后续强凸情形的证明提供了基础.

性质 4.2.4 (余强制性) *设函数 $f(\boldsymbol{x})$ 为可导的凸函数，且满足 L-光滑性，则有：*

$$\langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) \rangle \geq \frac{1}{L}\big\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\big\|^2$$

证明. 给定一个 $\boldsymbol{x} \in \mathbb{R}^n$, 我们取辅助函数 $g(\boldsymbol{t}) := f(\boldsymbol{t}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{t} - \boldsymbol{x} \rangle$, 则 $\nabla g(\boldsymbol{t}) = \nabla f(\boldsymbol{t}) - \nabla f(\boldsymbol{x})$. 根据凸函数的性质，满足 $\nabla g(\boldsymbol{t}) = \boldsymbol{0}$ 的点即为全局最小值点，而这个点恰好是 $\boldsymbol{x}$, 且 $g(\boldsymbol{x}) = 0$. 现在取定 $\boldsymbol{y} \in \mathbb{R}^n$, 并令 $\boldsymbol{t} = \boldsymbol{y} - \frac{1}{L}\nabla g(\boldsymbol{y})$, 同时使用光滑

函数的二次上界可以得到：

$$\begin{aligned}0 \le g\Big(\boldsymbol{y} - \frac{1}{L}\nabla g(\boldsymbol{y})\Big) &= f\Big(\boldsymbol{y} - \frac{1}{L}\nabla g(\boldsymbol{y})\Big) - f(\boldsymbol{x}) - \Big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} - \frac{1}{L}\nabla g(\boldsymbol{y})\Big\rangle \\ &\le f(\boldsymbol{y}) - \frac{1}{L}\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \nabla g(\boldsymbol{y})\rangle + \frac{1}{2L}\|\nabla g(\boldsymbol{y})\|^2 - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle \\ &= f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle - \frac{1}{2L}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2.\end{aligned}$$

同理交换 $\boldsymbol{x}, \boldsymbol{y}$ 我们有

$$0 \le f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle - \frac{1}{2L}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2.$$

两式相加即可得到最终结果. ■

我们下面考虑光滑凸函数作为目标函数时梯度下降算法的收敛性. 我们依然要先确定收敛指标. 在先前的讨论中，我们挑选 $\|\nabla f(\boldsymbol{x}_k)\|$ 来作为收敛指标. 而对于凸函数，由于其局部极小值点一定是全局极小值点，因此我们可以使用更好的收敛指标. 事实上，对于光滑凸函数，最常使用的收敛指标就是 $f(\boldsymbol{x}_k) - f^*$.

与非凸情形不同，为使得叠加时不同项的系数能够相同，以得到一个最简单形式的收敛结论，我们通常会精确取定一个步长 $\gamma = 1/(2L)$, 来进行理论分析. 尽管这在实际操作时往往不容易实现，但我们可以在最简单形式的基础上进行一些实用性的放缩，从而使得当步长不恰好取 $1/(2L)$ 时，收敛结论依然能够在同阶次的意义下成立. 光滑凸情形下的梯度下降算法收敛结论如下：

定理 4.2.5 (凸情形梯度下降法收敛性) *设目标函数 $f(\boldsymbol{x})$ 为 L-光滑的凸函数，且存在最优解 $\boldsymbol{x}^*$ 与最优值 $f^* = f(\boldsymbol{x}^*)$, 使用梯度下降法迭代 K 步，若取步长 $\gamma = 1/(2L)$, 则有*

$$f(\boldsymbol{x}_K) - f^* \le \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{K}$$

证明. 我们考虑将 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$ 作为叠加项，因此我们需要考虑将 $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2$ 分离出一个 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$ 来：

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \Big\|\boldsymbol{x}_k - \frac{1}{2L}\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^*\Big\|^2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \frac{1}{L}\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + \frac{1}{4L^2}\|\nabla f(\boldsymbol{x}_k)\|^2.\end{aligned}$$

分别对第二项、第三项使用凸性与下降引理得到：

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \frac{1}{L}\big(f(\boldsymbol{x}_k) - f^*\big) + \frac{1}{L}\big(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})\big) \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \frac{1}{L}\big(f(\boldsymbol{x}_{k+1}) - f^*\big).\end{aligned}$$

取 $k=0,1,\ldots,K-1$, 叠加可得：

$$\sum_{k=1}^{K}\big(f(\boldsymbol{x}_k)-f^*\big)\leq L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-L\|\boldsymbol{x}_K-\boldsymbol{x}^*\|^2.$$

根据下降引理，当 $\gamma\leq 2/L$ 时，函数值单调下降，因此 $\sum_{k=1}^{K}(f(\boldsymbol{x}_k)-f^*)\geq K(f(\boldsymbol{x}_K)-f^*)$, 且由与 $\|\boldsymbol{x}_K-\boldsymbol{x}^*\|\geq 0$, 我们有：

$$f(\boldsymbol{x}_K)-f^*\leq\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{K},$$

定理得证. ■

观察光滑凸函数对应的收敛性定理，我们看到收敛速率依然是 $\mathcal{O}(1/K)$, 相较非凸情形并没有改善速率. 然而，在凸函数的情形中我们使用了更好的收敛指标 $f(\boldsymbol{x}_k)-f^*$ 代替原先的梯度范数，可以保证收敛的点一定是一个全局最优值点. 因此该收敛性结论相比非凸情形更强.

我们注意到不等式的右侧存在因子 $\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$, 事实上我们假设了函数的最优解一定存在. 而某些光滑凸函数的最优解不一定存在（如 $e^{-x},x\geq 0$），这时上述结论就不能使用. 但是实验告诉我们，对这些函数使用梯度下降法，函数值的收敛速率通常也是 $\mathcal{O}(1/K)$, 只是我们尚且不好找到一个比较合适的方式去刻画这一收敛速率.

4.2.3 强凸情形的收敛性

我们下面考虑一个具体的例子. 目标函数为二次函数 $f(x)=x^2$, 取初始点 $x_0=1$, 梯度下降步长 $\gamma=1/4$, 通过计算我们不难得出整个迭代的通项公式为：

$$x_k=\left(\frac{1}{2}\right)^k,\quad k=0,1,2,\ldots$$

这个序列的收敛速率是指数阶次的，显然比 $\mathcal{O}(1/k)$ 要更快. 但是，二次函数具有什么样的良好性质，才能显著地提高收敛速率呢？这一性质就是强凸性（strongly convex）.

定义 4.2.6 (强凸性) *设多元函数 $f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$, 若存在常数 $\mu>0$ 使得*

$$g(\boldsymbol{x})=f(\boldsymbol{x})-\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}$$

为凸函数，则称 $f(\boldsymbol{x})$ 为 μ-强凸函数，其中 μ 称为强凸参数或强凸模量（modulus）.

由于原函数加减线性函数不影响凸性，只要存在 $\boldsymbol{b},c$ 使得 $f(\boldsymbol{x})-\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}-\boldsymbol{b}^\top\boldsymbol{x}-c$ 为凸函数，那么 $f(\boldsymbol{x})$ 就是强凸函数.

下述的性质是强凸函数的一个重要性质，我们发现它与光滑性的结论恰好是相反的两个方向. 光滑性给出了二次上界，而强凸性则给出了二次下界. [** 示意图 **]

性质 4.2.7 (强凸性的二次下界) 若可微的多元函数 $f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$ 在定义域 X 上是 μ-强凸的，则对任意 $\boldsymbol{x},\boldsymbol{y}\in X$, 有

$$f(\boldsymbol{y})\geq f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})+\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

证明. 对任意 $\boldsymbol{x},\boldsymbol{y}$, 对函数 $f(\boldsymbol{x})-\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}$ 使用凸性可得：

$$f(\boldsymbol{y})-\frac{\mu}{2}\boldsymbol{y}^\top\boldsymbol{y}\geq f(\boldsymbol{x})-\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}+\langle\nabla f(\boldsymbol{x})-\mu\boldsymbol{x},\boldsymbol{y}-\boldsymbol{x}\rangle,$$

即

$$\begin{aligned}f(\boldsymbol{y})&\geq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}-\mu\boldsymbol{x}^\top\boldsymbol{y}+\frac{\mu}{2}\boldsymbol{y}^\top\boldsymbol{y}\\&=f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2.\end{aligned}$$

命题得证. ■

下述性质将光滑性与强凸性结合考虑，其形式类似于将光滑凸函数的余强制性与强凸性的二次下界结合起来，但其系数更加紧致，不能简单由两个不等式相加得到.

性质 4.2.8 (扩展的余强制性) 若可微的多元函数 $f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$ 在定义域 X 上是 L-光滑且 μ-强凸的，则对于任意的 $\boldsymbol{x},\boldsymbol{y}\in X$, 有

$$\langle\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle\geq\frac{\mu L}{\mu+L}\|\boldsymbol{y}-\boldsymbol{x}\|^2+\frac{1}{\mu+L}\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|^2.$$

证明. 我们分两种情况讨论.

1. **情形一、** $\mu=L$, 此时有

   $$f(\boldsymbol{y})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})+\frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2,\quad\forall\boldsymbol{y},\boldsymbol{x}\in X,$$

   即 $f(\boldsymbol{x})$ 是一个二次函数，不妨记为 $f(\boldsymbol{x})=\frac{L}{2}\boldsymbol{x}^\top\boldsymbol{x}+\nabla f(\boldsymbol{0})^\top\boldsymbol{x}+f(\boldsymbol{0})$, 则 $\nabla f(\boldsymbol{x})=L\boldsymbol{x}+\nabla f(\boldsymbol{0})$, 我们自然有

   $$\begin{aligned}\langle\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle&=\langle L(\boldsymbol{y}-\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle\\&=\frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2+\frac{1}{2L}\|L(\boldsymbol{y}-\boldsymbol{x})\|^2\\&=\frac{\mu L}{\mu+L}\|\boldsymbol{y}-\boldsymbol{x}\|^2+\frac{1}{\mu+L}\|\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\|^2.\end{aligned}$$

2. **情形二、** $\mu<L$, 此时 $g(\boldsymbol{x}):=f(\boldsymbol{x})-\frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x}$ 是一个凸函数，我们进一步可以得到

Yuan and Song

它拥有 $L-\mu$ 的二次上界，这是因为，$\forall \boldsymbol{x},\boldsymbol{y}\in X$, 有

$$\begin{aligned} g(\boldsymbol{y}) &= f(\boldsymbol{y}) - \frac{\mu}{2}\boldsymbol{y}^\top\boldsymbol{y} \\ &\le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 - \frac{\mu}{2}\boldsymbol{y}^\top\boldsymbol{y} \\ &= f(\boldsymbol{x}) - \frac{\mu}{2}\boldsymbol{x}^\top\boldsymbol{x} + \big(\nabla f(\boldsymbol{x}) - \mu\boldsymbol{x}\big)^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{L-\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \\ &= g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{L-\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \end{aligned}$$

因此，我们可以使用光滑凸函数的余强制性（性质4.2.4），得到：

$$\langle \boldsymbol{y}-\boldsymbol{x}, \nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\rangle \ge \frac{1}{L-\mu}\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\|^2,$$

即

$$\langle \boldsymbol{y}-\boldsymbol{x}, \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \mu(\boldsymbol{y}-\boldsymbol{x})\rangle \ge \frac{1}{L-\mu}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \mu(\boldsymbol{y}-\boldsymbol{x})\|^2,$$

从而

$$\begin{aligned} &\Big(1+\frac{2}{L-\mu}\Big)\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \\ &\ge \frac{1}{L-\mu}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2 + \Big(\mu + \frac{\mu^2}{L-\mu}\Big)\|\boldsymbol{y}-\boldsymbol{x}\|^2, \end{aligned}$$

整理可得最后结果.

■

我们下面来考虑光滑且强凸函数作为目标函数时，梯度下降算法的收敛性质. 当拥有强凸性这一非常有利的条件后，我们的收敛指标可以直接选取最能反映收敛性的值，即迭代点距最优点的距离 $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$.

事实上，强凸函数的最优解一定存在且有限，这是因为根据其二次下界（性质4.2.7），任取一个点 $\boldsymbol{x}\in X$, 有：

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

我们注意到，集合 $\{\boldsymbol{y}\in\mathbb{R}^n : \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) \le 0\}$ 是空间 $\mathbb{R}^n$ 中的一个椭球，是一个有限区域，且 $\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ 的最小值一定是一个有限数. 因此函数 $f(\boldsymbol{x})$ 的最优解一定在该有限区域之内，且最优值有限，我们将最优解记为 $\boldsymbol{x}^*$.

光滑且强凸函数作为目标函数时的收敛性结论如下：

定理 4.2.9 (强凸情形梯度下降法收敛性) 设目标函数 $f(\boldsymbol{x})$ L-光滑且 μ-强凸, 设其最优解为 $\boldsymbol{x}^*$, 使用梯度下降法迭代 K 步，若取步长 $\gamma = \dfrac{2}{\mu+L}$, 则

$$\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \le \Big(\frac{L-\mu}{L+\mu}\Big)^K \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|.$$

Qian and Song

证明. 对于强凸性情形下的证明，我们不使用叠加法，而是直接使用迭代法，寻找 $\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|$ 与 $\|\boldsymbol{x}_k-\boldsymbol{x}^*\|$ 之间的直接关系. 我们仍然考虑拆分 $\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|$, 且由于 $\nabla f(\boldsymbol{x}^*)=\boldsymbol{0}$, 我们有

$$\begin{aligned}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_k-\gamma\nabla f(\boldsymbol{x}_k)-\boldsymbol{x}^*\|^2 = \|\boldsymbol{x}_k-\boldsymbol{x}^*-\gamma\nabla f(\boldsymbol{x}_k)+\gamma\nabla f(\boldsymbol{x}^*)\|^2\\ &= \|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+\gamma^2\|\nabla f(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}^*)\|^2-2\gamma\langle\boldsymbol{x}_k-\boldsymbol{x}^*,\nabla f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\rangle\\ &\le \Big(1-\frac{2\gamma\mu L}{\mu+L}\Big)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+\Big(\gamma^2-\frac{2\gamma}{\mu+L}\Big)\|\nabla f(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}^*)\|^2,\end{aligned}$$

其中最后一个不等式即使用到了光滑强凸函数的扩展的余强制性（性质4.2.8）. 进一步，当 $\gamma\le 2/(\mu+L)$ 时，$\nabla f(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}^*)$ 前的系数小于 0，而对于前面 $\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2$ 项的系数，γ 越大这里越小，因此我们就取最大的可能的值 $\gamma=2/(\mu+L)$, 此时有：

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\le\Big(1-\frac{4\mu L}{(\mu+L)^2}\Big)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2=\Big(\frac{L-\mu}{L+\mu}\Big)^2\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2,$$

即

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|\le\frac{L-\mu}{L+\mu}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|.$$

使用上述式子连续迭代 K 次即得最终结论. ■

观察此时光滑强凸函数的收敛性定理，我们发现此时收敛速率有很大改善，变为了指数函数的速率，即 $\mathcal{O}(\alpha^K)$, α 为一个小于 1 的常数. 在最优化理论中，我们一般称与指数函数同阶收敛速率的过程为**线性收敛**（因为在对数坐标轴上收敛曲线是一条直线），而将快于或慢于线性收敛的收敛过程分别称为超线性收敛和次线性收敛.

条件数的讨论.

总结三种下降结论

4.3　随机梯度下降及其收敛性

4.3.1　什么是随机梯度下降

4.3.2　收敛性分析

4.3.3　批量随机梯度下降

Yuan and Song

4.3.4 采样策略

4.4 动量类算法与自适应算法

4.4.1 （波利亚克）动量算法

4.4.2 涅斯捷罗夫动量算法

4.4.3 自适应算法

Yuan and Song

附录 A　数学基础

A.1　线性代数基础

A.1.1 线性空间及其子空间

考虑实数域 $\mathbb{R}$, 设 n 为给定正整数，记 n 元有序数组集合

$$\mathbb{R}^n := \{(x_1, x_2, \ldots, x_n) | x_i \in \mathbb{R}, i = 1, 2, \ldots, n\}.$$

$\mathbb{R}^n$ 中两元素 $(x_1, \ldots, x_n), (y_1, \ldots, y_n)$ 相等，当且仅当 $x_i = y_i, \forall i = 1, 2, \ldots$. $\mathbb{R}^n$ 中的加法定义为：

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n);$$

数量乘法定义为 $k(x_1, \ldots, x_n) := (kx_1, \ldots, kx_n)$. 不难验证 $\mathbb{R}^n$ 中的加法与数量乘法满足如下 8 条运算法则：

1. $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ （加法交换律）.
2. $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}), \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ （加法结合律）.
3. 记零元素为 $\mathbf{0} := (0, 0, \ldots, 0) \in \mathbb{R}^n$, 其满足 $\mathbf{0} + \boldsymbol{x} = \boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$.
4. 对任意 $\boldsymbol{x} \in \mathbb{R}^n$, 存在其负元素 $-\boldsymbol{x}$ 使得 $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$.
5. 对任意 $\boldsymbol{x} \in \mathbb{R}^n$, 有 $1\boldsymbol{x} = \boldsymbol{x}$.
6. 对任意 $\boldsymbol{x} \in \mathbb{R}^n, k, l \in \mathbb{R}$, 有 $(kl)\boldsymbol{x} = k(l\boldsymbol{x})$.
7. 对任意 $\boldsymbol{x} \in \mathbb{R}^n, k, l \in \mathbb{R}$, 有 $(k + l)\boldsymbol{x} = k\boldsymbol{x} + l\boldsymbol{x}$.
8. 对任意 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, k \in \mathbb{R}$, 有 $k(\boldsymbol{x} + \boldsymbol{y}) = k\boldsymbol{x} + k\boldsymbol{y}$.

定义 A.1.1 (n 维实数向量空间) 实数域上的 n 元有序数组集合 $\mathbb{R}^n$, 连同定义在其上的加法与数量乘法运算，及其满足的 8 条运算法则一起，共同构成n 维实数向量空间.

注 A.1.2 n 维实数向量空间是线性空间最基础的一个特例，另有定义在复数域 $\mathbb{C}$ 上、素域 $\mathbb{Z}/p\mathbb{Z}$ 上的线性空间，*[** 函数空间 **]* 等. 判断一个集合及定义在其上的加法与数量

Yuan and Song

乘法是否构成一个线性空间的标准即它们是否满足上述的 8 条运算法则. 本书绝大多数情形下仅考虑 n 维实数向量空间，在不引起歧义的前提下，可直接称其为线性空间.

在线性代数中，由于需要考虑矩阵间的运算，因此需要明确向量为行向量或是列向量. 如果未加说明，向量 $\boldsymbol{x}$ 通常指**列向量**，而使用 $\boldsymbol{x}^\top$ 来特指其对应的行向量.

定义 A.1.3 (线性子空间) $\mathbb{R}^n$ 的一个非空子集 U 如果满足：

1. $\boldsymbol{x}, \boldsymbol{y} \in U \Longrightarrow \boldsymbol{x}+\boldsymbol{y} \in U$（加法封闭）；
2. $\boldsymbol{x} \in U, k \in \mathbb{R} \Longrightarrow k\boldsymbol{x} \in \mathbb{R}$（数量乘法封闭），

则称 U 为 $\mathbb{R}^n$ 的一个（线性）子空间.

例 A.1.4 设 $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s \in \mathbb{R}^n$, 定义集合

$$\operatorname{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s) := \{k_1\boldsymbol{a}_1 + \cdots + k_s\boldsymbol{a}_s | k_i \in \mathbb{R}, \forall i = 1, 2, \ldots, s\}$$

为 $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s$ 张成的子空间，可以验证其确实为 R^n 上的一个子空间.

A.1.2 矩阵及其基本运算

记 m 行 n 列的有序数组集合为

$$\mathbb{R}^{m\times n} := \left\{ \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \middle| a_{ij} \in \mathbb{R}, \forall i = 1, 2, \ldots, m; j = 1, 2, \ldots, n \right\}.$$

该集合中的每个元素称为一个 $m \times n$ 维矩阵.

仿照 n 维实数向量空间中的运算，我们同样可以定义矩阵的加法与数量乘法：

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix} := \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{pmatrix};$$

$$k \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} := \begin{pmatrix} ka_{11} & \cdots & ka_{1n} \\ \vdots & & \vdots \\ ka_{m1} & \cdots & ka_{mn} \end{pmatrix}, \quad k \in \mathbb{R}.$$

自此，我们可以得出，所有 $m \times n$ 维矩阵构成的集合同样满足线性空间的定义，我们可以证明其同构于 mn 维实数向量空间，在某些操作中（如复合矩阵函数的求导），我们需要将矩阵展开为 mn 维向量才可使用链式法则.

然而，矩阵在线性空间的基础上，还拥有额外的结构，矩阵乘法即为其最基本的体

Yuan and Song

现. 设矩阵 $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}$, 矩阵乘法的定义如下：

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} a_{1i}b_{i1} & \cdots & \sum_{i=1}^{n} a_{1i}b_{ip} \\ \vdots & & \vdots \\ \sum_{i=1}^{n} a_{mi}b_{i1} & \cdots & \sum_{i=1}^{n} a_{mi}b_{ip} \end{pmatrix} \in \mathbb{R}^{m\times p}.$$

矩阵的乘法不满足交换律（$AB = BA$）, 也不满足无零因子性（$AB = 0 \Longrightarrow A = 0$或$B = 0$），但满足结合律与左右分配律，即

$$(AB)C = A(BC); \quad A(B + C) = AB + AC; \quad (A + B)C = AC + BC$$

除矩阵乘法外,我们可能还会在某些特定的场景中使用到矩阵的哈达玛积（Hadamard product）与矩阵的克罗内克积（Kronecker product），它们的定义如下：

定义 A.1.5 (哈达玛积) 设矩阵 $A, B \in \mathbb{R}^{m\times n}$, 则二者的哈达玛积为

$$A \odot B := \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1n}b_{1n} \\ \vdots & & \vdots \\ a_{m1}b_{m1} & \cdots & a_{mn}b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

定义 A.1.6 (克罗内克积) 设矩阵 $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$, 则二者的克罗内克积为

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \in \mathbb{R}^{(mp)\times(nq)}.$$

A.1.3 线性相关、线性无关及矩阵的秩

定义 A.1.7 (线性组合与线性表出) 给定 $\mathbb{R}^n$ 中的一组向量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$, 给定一组数 $k_1, \ldots, k_s \in \mathbb{R}$, 称 $k_1\boldsymbol{x}_1 + \cdots + k_s\boldsymbol{x}_s$ 为该向量组的一个线性组合. 对于另一向量 $\boldsymbol{y} \in \mathbb{R}^n$, 如果存在一组 $k_1, \ldots, k_s \in \mathbb{R}$ 使得

$$\boldsymbol{y} = k_1\boldsymbol{x}_1 + \cdots + k_s\boldsymbol{x}_s,$$

则称 $\boldsymbol{y}$ 可以由 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ 线性表出.

定义 A.1.8 (线性相关) 称 $\mathbb{R}^n$ 中的一组向量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ 线性相关，如果存在不全为 0 的实数 $k_1, \ldots, k_s \in \mathbb{R}$, 使得 $k_1\boldsymbol{x}_1 + \cdots + k_s\boldsymbol{x}_s = \boldsymbol{0}$.

定义 A.1.9 (线性无关) 称 $\mathbb{R}^n$ 中的一组向量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ 线性无关，如果

$$k_1\boldsymbol{x}_1 + \cdots + k_s\boldsymbol{x}_s = \boldsymbol{0} \Longrightarrow k_i = 1, \forall i = 1, \ldots, s.$$

Yuan and Song

定义 A.1.10 (极大线性无关组) 对于 $\mathbb{R}^n$ 中的向量组 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$, 如果其中的一部分向量本身是线性无关的，但加入该向量组中的其他任何一个向量（如果还有的话）都是线性相关的，则称该部分向量为这个向量组的极大线性无关组.

基于线性代数的推导结果，我们不加证明地引出如下推论：

推论 A.1.11 向量组的任意两个极大线性无关组所含向量的数目相等.

基于此，我们可以定义向量组的秩：

定义 A.1.12 (向量组的秩) 向量组 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s \in \mathbb{R}^n$ 的极大线性无关组所含的向量数目称为向量组的秩，记为 $\text{rank}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\}$. 全由零向量 $\mathbf{0}$ 构成的向量组的秩规定为 0.

对于矩阵 $A \in \mathbb{R}^{m\times n}$, 其可视为由 n 个 m 维列向量构成的向量组，也可以视为由 m 个 n 维行向量构成的向量组，根据线性代数的结论，两个向量组拥有相同的秩，我们将其统称为矩阵 A 的秩，记为 $\text{rank}(A)$.

A.1.4 线性空间的基与坐标表示

A.2 微积分基础

多元函数连续