

尚硅谷大模型技术之数学基础

（作者：尚硅谷研究院）

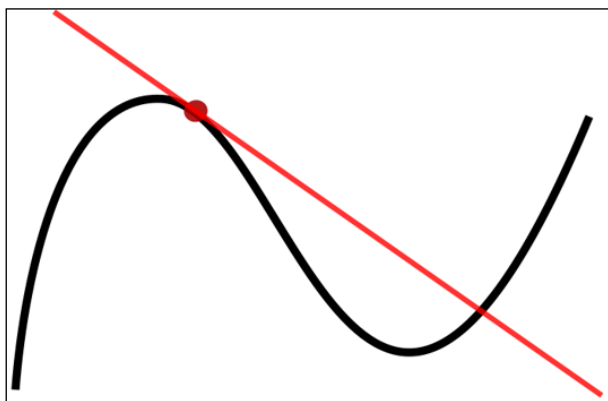
版本：V0.9.2

第 1 章 高等数学

1.1 导数

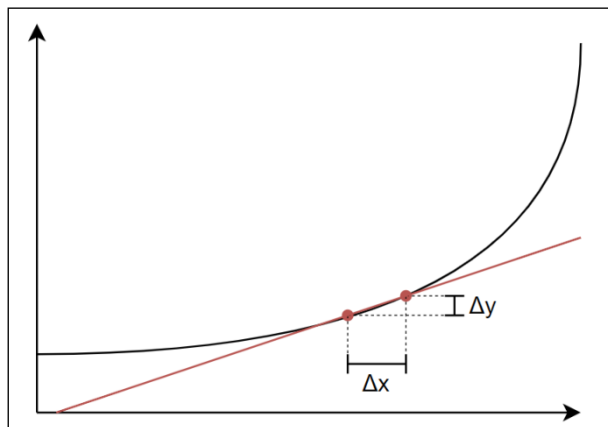
1.1.1 导数的概念

导数（derivative）是微积分中的一个概念。函数在某一点的导数是指这个函数在这一点附近的变化率（即函数在这一点处的切线斜率）。导数的本质是通过极限的概念对函数进行局部的线性逼近。



当函数 f 的自变量在一点 x_0 上产生一个增量 h 时，函数输出值的增量 Δy 与自变量增量 Δx 的比值在 Δx 趋于 0 时的极限如果存在，即为 f 在 x_0 处的导数，记作 $f'(x_0)$ 、 $\frac{df}{dx}(x_0)$ 或 $\frac{df}{dx} \Big|_{x=x_0}$ 。

$$f'(x_0) = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$



例如在运动学中，物体的位移对于时间的导数就是物体的瞬时速度： $v = \frac{dx}{dt}$ 。

1.1.2 基本函数的导数

说明	公式	例子
常数的导数	$(C)' = 0$	$(3)' = 0$
幂函数的导数	$(x^a)' = ax^{a-1}$	$(x^3)' = 3x^2$
指数函数的导数	$(a^x)' = a^x \ln a$	$(3^x)' = 3^x \ln 3$
	$(e^x)' = e^x$	—
对数函数的导数	$(\log_a x)' = \frac{1}{x \ln a}$	$(\log_3 x)' = \frac{1}{x \ln 3}$
	$(\ln x)' = \frac{1}{x}$	—
三角函数的导数	$(\sin x)' = \cos x$	—
	$(\cos x)' = -\sin x$	—
	$(\tan x)' = \sec^2 x = \frac{1}{\cos^2 x}$	—
	$(\cot x)' = -\csc^2 x = \frac{-1}{\sin^2 x}$	—

1.1.3 导数的求导法则

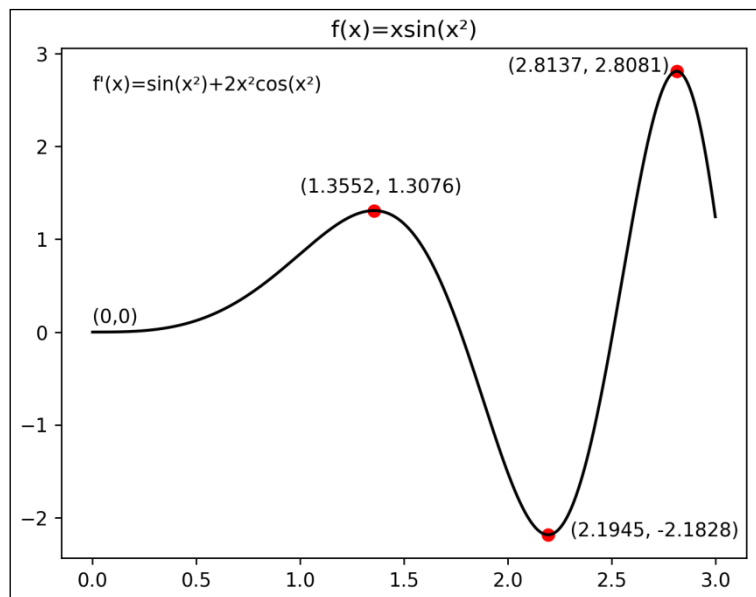
说明	公式
两函数之和求导	$(f + g)' = f' + g'$
两函数之积求导	$(fg)' = f'g + fg'$
两函数之商求导	$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$
复合函数的导数	若 $f(x) = h[g(x)]$, 则 $f'(x) = h'[g(x)] \cdot g'(x)$

例如：求函数 $f(x) = x^4 + \sin(x^2) - \ln(x)e^x + 7$ 在 $x = 3$ 处的导数。

$$\begin{aligned} f'(x) &= 4x^{4-1} + \cos(x^2) \cdot 2x - \left(\frac{e^x}{x} + \ln(x)e^x\right) + 0 \\ &= 4x^3 + 2x\cos(x^2) - \frac{e^x}{x} - \ln(x)e^x \\ f'(3) &= 108 + 6\cos(9) - \frac{e^3}{3} - \ln(3)e^3 \end{aligned}$$

1.1.4 利用导数求极值

导数等于零的点称为函数的驻点（或极值可疑点），在这类点上函数可能会取得极大值或极小值。进一步判断则需要知道导数在附近的符号。



1.1.5 二阶导数

1) 二阶导数的概念

在微积分中，函数的二阶导数是函数导数的导数。粗略来说，某个量的二阶导数描述该量变化率变化的快慢。例如物体位置对时间的二阶导数是物体的瞬时加速度，即该物体速度随时间的变化率： $a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$ 。

函数 f 的二阶导数通常记作 f'' 、 $\frac{d^2y}{dx^2}$ 或 $\frac{d}{dx}\left(\frac{dy}{dx}\right)$ 。

2) 二阶导数与函数凹凸的关系

函数的二阶导数描述了函数图像的凹凸方向和程度。若二阶导数在某区间恒为正，则函数在该区间向上弯（也称下凸函数）。反之，若二阶导数在某区间恒为负，则函数在该区间向下弯（也称上凸函数）。

若函数的二阶导数在某点左右异号，则图像由向上弯转为向下弯，或反之。这种点称之为拐点。若二阶导数连续，则在该点处二阶导数为 0。但反之二阶导数为 0 的点不一定是拐点。如 $f(x) = x^4$ ，在 $x = 0$ 处有 $f''(0) = 0$ ，但 $f(x)$ 在实数系上无拐点。

二阶导数与凹凸性的关系有助于判断函数的驻点是否为极大值点或极小值点：

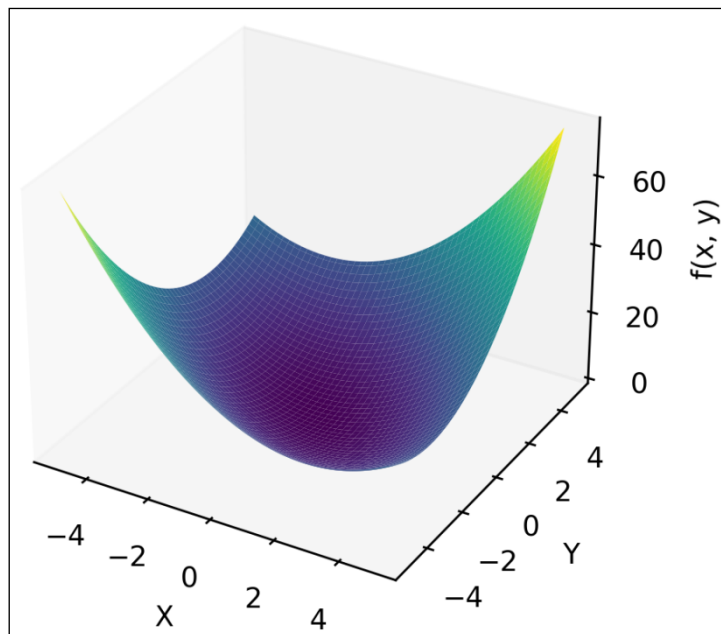
- 若 $f'(x) = 0$ ， $f''(x) < 0$ ，则 f 在 x 取得极大值。
- 若 $f'(x) = 0$ ， $f''(x) > 0$ ，则 f 在 x 取得极小值。
- 若 $f'(x) = 0$ ， $f''(x) = 0$ ，则该点可能是拐点，也可能是极大值点或极小值点。

1.2 偏导与梯度

1.2.1 偏导数

如果函数 f 的自变量并非单个元素，而是多个元素，例如：

$$f(x,y) = x^2 + xy + y^2$$



可将其中一个元素 x 看作参数，此时 f 可看作关于另一元素 y 的函数。

$$f_x(y) = x^2 + xy + y^2$$

在 $x = a$ 固定的情况下，可计算 f_x 关于 y 的导数：

$$f_{x=a}'(y) = a + 2y$$

这种导数称为偏导数，一般记作：

$$\frac{\partial f}{\partial y}(x,y) = x + 2y$$

更一般地来说，一个多元函数 $f(x_1, x_2, \dots, x_n)$ 在点 (a_1, a_2, \dots, a_n) 处对 x_i 的偏导数定义为：

$$\frac{\partial f}{\partial x_i}(a_1, a_2, \dots, a_n) = \lim_{\Delta x_i \rightarrow 0} \frac{f(a_1, \dots, a_i + \Delta x_i, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{\Delta x_i}$$

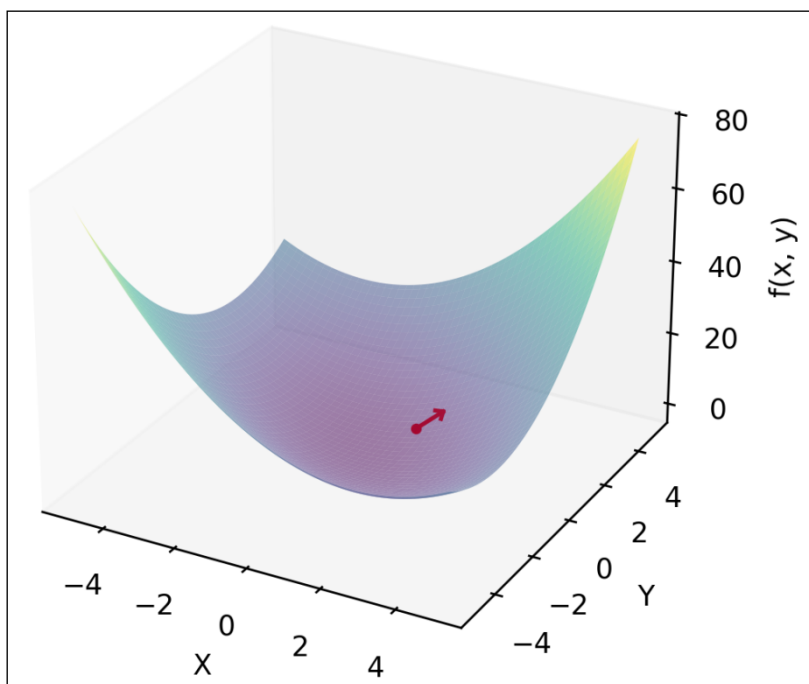
1.2.2 梯度

多元函数 $f(x_1, \dots, x_n)$ 关于每个变量 x_i 都有偏导数 $\frac{\partial f}{\partial x_i}$ ，在点 a 处，这些偏导数定义了一个向量。

$$\nabla f(a) = \left[\frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), \dots, \frac{\partial f}{\partial x_n}(a) \right]$$

这个向量称为 f 在点 a 的梯度。

例如： $f(x, y) = x^2 + xy + y^2$ 在 $(1, 1)$ 处的梯度为 $[3, 3]$ 。



第2章 线性代数

2.1 标量与向量

2.1.1 标量与向量的概念

1) 标量 (scalar)

标量是一个单独的数，只有大小。

2) 向量 (vector)

向量由标量组成，有大小有方向。

行向量： $(2 \ 5 \ 8)$

列向量： $\begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}$

2.1.2 向量运算

1) 向量转置：列向量转置结果为行向量

$$\begin{aligned} x &= \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} \\ x^T &= (2 \quad 5 \quad 8) \end{aligned}$$

2) 向量相加：对应元素相加

$$\begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} + \begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} = \begin{pmatrix} 3 \\ 8 \\ 15 \end{pmatrix}$$

3) 向量与标量相乘：标量与向量每个元素相乘

$$3 \times \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} = \begin{pmatrix} 6 \\ 15 \\ 24 \end{pmatrix}$$

4) 向量内积：又称向量点乘，两向量对应元素乘积之和，结果为标量

$$\langle x, y \rangle = \left\langle \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} \right\rangle = 2 + 15 + 56 = 73$$

两向量之间夹角表示为

$$\cos \theta = \frac{\langle x, y \rangle}{\sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle}}$$

2.1.3 向量范数

范数（norm）是具有“长度”概念的函数。

1) L_0 范数（也称 0 范数）

$$\|x\|_0 = \text{非零元素的个数}$$

例如：

$$x = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|x\|_0 = 2$$

2) L_1 范数（也称和范数或 1 范数）

$$\|x\|_1 = \sum_{i=1}^m |x_i| = |x_1| + \dots + |x_m|$$

例如：

$$x = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|x\|_1 = 0 + 2 + 1 = 3$$

3) L_2 范数（也称欧几里得范数或 2 范数）

$$\|x\|_2 = \left(\sum_{i=1}^m |x_i|^2 \right)^{\frac{1}{2}} = \sqrt{|x_1|^2 + \dots + |x_m|^2}$$

例如：

$$x = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|x\|_2 = \sqrt{0 + 4 + 1} = \sqrt{5}$$

4) L_p 范数

$$\|x\|_p = \left(\sum_{i=1}^m |x_i|^p \right)^{\frac{1}{p}} = (|x_1|^p + \dots + |x_m|^p)^{\frac{1}{p}}$$

在 numpy 中，可以利用 `linalg.norm` 函数方便地计算向量的范数。

2.2 矩阵与张量

2.2.1 矩阵的概念

一个 $m \times n$ 的矩阵 (matrix) 是一个有 m 行 n 列元素的矩形阵列。用 $\mathbb{R}^{m \times n}$ 表示所有 $m \times n$ 实数矩阵的向量空间。

$$\begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in \mathbb{R}^{3 \times 2}$$

1) 方阵：行数等于列数的矩阵

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

2) 对角矩阵：主对角线以外元素全为 0 的方阵

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

3) 单位矩阵：主对角线元素全为 1 的对角矩阵

$$I_{3 \times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2.2.2 矩阵乘法

1) 矩阵乘法运算

两个矩阵的乘法仅当矩阵A的列数和矩阵B的行数相等时才能定义。如 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, 它们的乘积 $AB \in \mathbb{R}^{m \times p}$

$$[AB]_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{in}b_{nj} = \sum_{r=1}^n a_{ir}b_{rj}$$

例如:

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{bmatrix} \times \begin{bmatrix} 3 & 1 \\ 2 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 \times 3 + 0 \times 2 + 2 \times 1 & 1 \times 1 + 0 \times 1 + 2 \times 0 \\ (-1) \times 3 + 3 \times 2 + 1 \times 1 & (-1) \times 1 + 3 \times 1 + 1 \times 0 \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix}$$

特别地, 矩阵与单位矩阵相乘等于矩阵本身:

$$AI = A (A \in \mathbb{R}^{m \times n}, I \in \mathbb{R}^{n \times n}) \text{ 或 } IA = A (I \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{n \times m})$$

例如:

$$AI = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 2 \times 0 & 1 \times 0 + 2 \times 1 \\ 3 \times 1 + 5 \times 0 & 3 \times 0 + 5 \times 1 \\ 4 \times 1 + 8 \times 0 & 4 \times 0 + 8 \times 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} = A$$

2) 矩阵乘法的性质

矩阵乘法满足结合律、左分配律和右分配律。不满足交换律即 $AB \neq BA$ 。

结合律: 若 $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}, C \in \mathbb{R}^{p \times q}$, 则 $(AB)C = A(BC)$

左分配律: 若 $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^{n \times p}$, 则 $(A + B)C = AC + BC$

右分配律: 若 $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^{n \times p}$, 则 $A(B + C) = AB + AC$

2.2.3 矩阵转置

1) 矩阵转置运算

矩阵 $A \in \mathbb{R}^{m \times n}$ 的转置是一个 $n \times m$ 的矩阵, 记为 A^T 。其中的第i个行向量是原矩阵的第i个列向量; 或者说, 转置矩阵 A^T 第i行第j列的元素是原矩阵A第j行第i列的元素。

$$[A^T]_{ij} = a_{ji}$$

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in \mathbb{R}^{3 \times 2}$$

$$A^T = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 5 & 8 \end{bmatrix} \in \mathbb{R}^{2 \times 3}$$

2) 矩阵转置的性质

$$(A^T)^T = A$$

$$(A + B)^T = A^T + B^T$$

$$(kA)^T = kA^T$$

$$(AB)^T = B^T A^T$$

2.2.4 矩阵的逆

对于方阵 A ，如果存在另一个方阵 A^{-1} ，使得 $AA^{-1} = I$ 成立，此时 $A^{-1}A = I$ 也同样成立。

称 A^{-1} 为 A 的逆矩阵。例如：

$$AA^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \times \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 \times (-5) + 2 \times 3 & 1 \times 2 + 2 \times (-1) \\ 3 \times (-5) + 5 \times 3 & 3 \times 2 + 5 \times (-1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

2.2.5 其他矩阵运算

1) 矩阵的向量化

矩阵 $A \in R^{m \times n}$ 的向量化 $\text{vec}(A)$ 将矩阵 A 的元素按列排列成一个 $mn \times 1$ 的向量。

$$\text{vec}(A) = [a_{11}, \dots, a_{m1}, \dots, a_{1n}, \dots, a_{mn}]^T$$

矩阵也可以转化为行向量 $\text{rvec}(A)$ ，称为矩阵的行向量化。

$$\text{rvec}(A) = [a_{11}, \dots, a_{m1}, \dots, a_{1n}, \dots, a_{mn}]$$

$$\text{例如：} A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \text{vec}(A) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \text{rvec}(A) = (1 \quad 2 \quad 3 \quad 4)。$$

2) 矩阵的内积

矩阵 $A \in R^{m \times n}$ 和矩阵 $B \in R^{m \times n}$ 的内积记作 $\langle A, B \rangle$ ，它是两个矩阵对应元素乘积之和，是一个标量。

$$\langle A, B \rangle = \langle \text{vec}(A), \text{vec}(B) \rangle = \sum a_{ij} b_{ij}$$

3) 矩阵的 Hadamard 积

矩阵 $A \in R^{m \times n}$ 和矩阵 $B \in R^{m \times n}$ 的Hadamard积记作 $A \odot B$ ，它是两个矩阵对应元素的乘积，是一个 $m \times n$ 的矩阵。

$$(A \odot B)_{ij} = a_{ij} b_{ij}$$

4) 矩阵的 Kronecker 积

矩阵 $A \in R^{m \times n}$ 和矩阵 $B \in R^{p \times q}$ 的Kronecker积记作 $A \otimes B$ ，它是矩阵 A 中每个元素与矩阵 B 的乘积，是一个 $mp \times nq$ 的矩阵。

$$\begin{aligned} (A \otimes B)_{ij} &= [a_1 B \quad a_2 B \quad \cdots \quad a_n B] \\ &= [a_{ij} B]_{i=1, j=1}^{m, n} \\ &= \begin{bmatrix} a_{11} B & a_{12} B & \cdots & a_{1n} B \\ a_{21} B & a_{22} B & \cdots & a_{2n} B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} B & a_{m2} B & \cdots & a_{mn} B \end{bmatrix} \end{aligned}$$

Kronecker 积也称为直积或张量积。

2.2.6 张量

张量（tensor）可视为多维数组，是标量，1 维向量和 2 维矩阵的 n 维推广。

例如：3 维张量

$$\begin{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} & \begin{bmatrix} 3 & 2 \\ 1 & 6 \\ 7 & 3 \end{bmatrix} & \begin{bmatrix} 5 & 6 \\ 9 & 1 \\ 2 & 4 \end{bmatrix} \end{bmatrix}$$

2.3 矩阵求导

矩阵求导的本质就是函数对变元的每个元素逐个求导，只是写成了向量、矩阵的形式。

为方便理解，首先对变元和函数作统一的符号规定：

$\mathbf{x} = [x_1, x_2, \dots, x_m]^T \in R^m$ 为实向量变元。

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]^T \in R^{m \times n}$ 为实矩阵变元。

$f(\mathbf{x}) \in R$ 为实标量函数，其变元 \mathbf{x} 为实向量。

$f(\mathbf{X}) \in R$ 为实标量函数，其变元 \mathbf{X} 为实矩阵。

$\mathbf{f}(\mathbf{x}) \in R^p$ 为实向量函数，其变元 \mathbf{x} 为实向量。

$\mathbf{f}(\mathbf{X}) \in R^p$ 为实向量函数，其变元 \mathbf{X} 为实矩阵。

$\mathbf{F}(\mathbf{x}) \in R^{p \times q}$ 为实矩阵函数，其变元 \mathbf{x} 为实向量。

$\mathbf{F}(\mathbf{X}) \in R^{p \times q}$ 为实矩阵函数，其变元 \mathbf{X} 为实矩阵。

2.3.1 常用求导结果

$$\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}^T + \mathbf{A})\mathbf{x}$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{b}}{\partial \mathbf{X}} = \frac{\partial (\mathbf{a}^T \mathbf{X}^T \mathbf{b})^T}{\partial \mathbf{X}} = \frac{\partial \mathbf{b}^T \mathbf{X} \mathbf{a}}{\partial \mathbf{X}} = \mathbf{b} \mathbf{a}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{X}^T \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^T \mathbf{X} + \mathbf{b} \mathbf{a}^T \mathbf{X}$$

$$\frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{X} \mathbf{b} \mathbf{a}^T + \mathbf{X} \mathbf{a} \mathbf{b}^T$$

2.3.2 Jacobian 矩阵与梯度矩阵

1) Jacobian 矩阵

实标量函数 $f(\mathbf{x})$ 在 \mathbf{x} 的偏导向量为 $1 \times m$ 的行向量：

$$D_{\mathbf{x}}f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_m} \right]$$

当实标量函数 $f(\mathbf{X})$ 的变元为实矩阵 \mathbf{X} 时，存在两种可能的定义：

$$D_{\mathbf{X}}f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}^T} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f(\mathbf{X})}{\partial x_{m1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{1n}} & \dots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in R^{n \times m}$$

和

$$D_{\text{vec} \mathbf{X}}f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \text{vec}^T(\mathbf{X})} = \left[\frac{\partial f(\mathbf{X})}{\partial x_{11}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{m1}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{1n}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \right]$$

其中 $D_{\mathbf{X}}f(\mathbf{X})$ 与 $D_{\text{vec} \mathbf{X}}f(\mathbf{X})$ 分别称为实标量函数 $f(\mathbf{X})$ 关于实矩阵变元 \mathbf{X} 的 Jacobian 矩阵和行偏导矩阵。两者之间的关系为 $D_{\text{vec} \mathbf{X}}f(\mathbf{X}) = \text{rvec}(D_{\mathbf{X}}f(\mathbf{X})) = (\text{vec}(D_{\mathbf{X}}^T f(\mathbf{X})))^T$ 。

实矩阵函数 $\mathbf{F}(\mathbf{X}) = [f_{kl}]_{k=1,l=1}^{p,q} \in R^{p \times q}$ ，其实矩阵变元 $\mathbf{X} \in R^{m \times n}$ 。为得到该函数的 Jacobian 矩阵，先通过列向量化，将 $p \times q$ 的矩阵函数 $\mathbf{F}(\mathbf{X})$ 转换为 $pq \times 1$ 的列向量：

$$\text{vec}(\mathbf{F}(\mathbf{X})) = [f_{11}(\mathbf{X}), \dots, f_{p1}(\mathbf{X}), \dots, f_{1q}(\mathbf{X}), \dots, f_{pq}(\mathbf{X})]^T \in R^{pq}$$

然后，该列向量对列向量化的 \mathbf{X} 求偏导，得到 $pq \times mn$ 的 Jacobian 矩阵：

$$D_{\mathbf{X}}f(\mathbf{X}) = \frac{\partial \text{vec}(\mathbf{F}(\mathbf{X}))}{\partial (\text{vec} \mathbf{X})^T} \in R^{pq \times mn}$$

其具体表达式为：

$$D_{\mathbf{X}}f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f_{11}(\mathbf{X})}{\partial (\text{vec} \mathbf{X})^T} \\ \vdots \\ \frac{\partial f_{p1}(\mathbf{X})}{\partial (\text{vec} \mathbf{X})^T} \\ \vdots \\ \frac{\partial f_{1q}(\mathbf{X})}{\partial (\text{vec} \mathbf{X})^T} \\ \vdots \\ \frac{\partial f_{pq}(\mathbf{X})}{\partial (\text{vec} \mathbf{X})^T} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_{11}(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f_{11}(\mathbf{X})}{\partial x_{m1}} & \dots & \frac{\partial f_{11}(\mathbf{X})}{\partial x_{1n}} & \dots & \frac{\partial f_{11}(\mathbf{X})}{\partial x_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial f_{p1}(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f_{p1}(\mathbf{X})}{\partial x_{m1}} & \dots & \frac{\partial f_{p1}(\mathbf{X})}{\partial x_{1n}} & \dots & \frac{\partial f_{p1}(\mathbf{X})}{\partial x_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial f_{1q}(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f_{1q}(\mathbf{X})}{\partial x_{m1}} & \dots & \frac{\partial f_{1q}(\mathbf{X})}{\partial x_{1n}} & \dots & \frac{\partial f_{1q}(\mathbf{X})}{\partial x_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial f_{pq}(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f_{pq}(\mathbf{X})}{\partial x_{m1}} & \dots & \frac{\partial f_{pq}(\mathbf{X})}{\partial x_{1n}} & \dots & \frac{\partial f_{pq}(\mathbf{X})}{\partial x_{mn}} \end{bmatrix}$$

2) 梯度矩阵

采用列向量形式定义的偏导算子称为列向量偏导算子，习惯称为梯度算子。

实标量函数 $f(\mathbf{x})$ 的梯度向量为 $m \times 1$ 的列向量：

$$\nabla_{\mathbf{x}}f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_m} \right]^T$$

将矩阵变元 \mathbf{X} 列向量化，得到 $f(\mathbf{X})$ 的梯度算子：

$$\nabla_{\text{vec} \mathbf{X}} f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \text{vec}(\mathbf{X})} = \left[\frac{\partial f(\mathbf{X})}{\partial x_{11}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{m1}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{1n}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \right]^T$$

另外，可直接得到其梯度矩阵：

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \dots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \dots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in R^{m \times n}$$

可以发现，实标量函数 $f(\mathbf{X})$ 的梯度矩阵等于 Jacobian 矩阵的转置， $\nabla_{\mathbf{X}} f(\mathbf{X}) = (\mathbf{D}_{\mathbf{X}} f(\mathbf{X}))^T$ 。另外，实矩阵函数 $\mathbf{F}(\mathbf{X})$ 的梯度矩阵也是其 Jacobian 矩阵的转置， $\nabla_{\mathbf{X}} \mathbf{F}(\mathbf{X}) = (\mathbf{D}_{\mathbf{X}} \mathbf{F}(\mathbf{X}))^T$ 。

类似地， $f(\mathbf{x})$ 的二阶偏导构成的矩阵被称为“黑塞矩阵”（Hessian Matrix）：

$$\mathbf{H}(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{n \times n}$$

2.3.3 一阶实矩阵微分

1) 矩阵的迹

方阵 $\mathbf{A} \in R^{n \times n}$ 的主对角线元素之和为矩阵 \mathbf{A} 的迹（trace），记作 $\text{tr}(\mathbf{A})$ ：

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

$$\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \dots + a_{nn} = \sum_{i=1}^n a_{ii}$$

标量的迹是其本身： $x = \text{tr}(x)$

转置： $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T)$

线性： $\text{tr}(c_1 \mathbf{A} + c_2 \mathbf{B}) = c_1 \text{tr}(\mathbf{A}) + c_2 \text{tr}(\mathbf{B})$

交换律： $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$

2) 矩阵微分的性质

矩阵微分用符号 $d\mathbf{X}$ 表示，定义为 $d\mathbf{X} = [dx_{ij}]_{i=1, j=1}^{m, n}$ 。

考虑标量函数的微分，有 $d(\text{tr} \mathbf{U}) = \text{tr}(d\mathbf{U})$ ：

$$d(\text{tr} \mathbf{U}) = d\left(\sum_{i=1}^n u_{ii}\right) = \sum_{i=1}^n du_{ii} = \text{tr}(d\mathbf{U})$$

考虑矩阵乘积 \mathbf{UV} 的微分矩阵，有 $d(\mathbf{UV}) = (d\mathbf{U})\mathbf{V} + \mathbf{U}d\mathbf{V}$ ：

$$\begin{aligned}
 [d(UV)]_{ij} &= d([UV]_{ij}) \\
 &= d\left(\sum_k u_{ik}v_{kj}\right) = \sum_k d(u_{ik}v_{kj}) \\
 &= \sum_k [(du_{ik})v_{kj} + u_{ik}dv_{kj}] \\
 &= \sum_k (du_{ik})v_{kj} + \sum_k u_{ik}dv_{kj} \\
 &= [(dU)V]_{ij} + [UdV]_{ij}
 \end{aligned}$$

上两例表明，实矩阵微分具有以下两个基本性质：

- 转置：矩阵转置的微分等于矩阵微分的转置，即 $d(X^T) = (dX)^T$ 。
- 线性： $d(\alpha X + \beta Y) = \alpha dX + \beta dY$ 。

下面汇总矩阵微分常用的计算公式：

- $dA = O$; A 为常数矩阵, O 为零矩阵
- $d(\alpha X) = \alpha d(X)$
- $d(X^T) = (dX)^T$
- $d(U \pm V) = dU \pm dV$
- $d(AXB) = A(dX)B$
- $d(UVW) = d(U)VW + U(dV)W + UVd(W)$

$U = F(X), V = G(X), W = H(X)$ 为矩阵函数

- $d(\text{tr}X) = \text{tr}(dX)$, 并且 $d(\text{tr}(F(X))) = \text{tr}(d(F(X)))$
- $d|X| = |X|\text{tr}(X^{-1}dX)$; $|X|$ 为行列式, 并且 $d|F(X)| = |F(X)|\text{tr}(F^{-1}(X)d(F(X)))$
- $d(U \otimes V) = (dU) \otimes V + U \otimes d(V)$; $U \otimes V$ 为矩阵Kronecker积
- $d(U \odot V) = (dU) \odot V + U \odot d(V)$; $U \odot V$ 为矩阵Hadamard积
- $d(\text{vec}(X)) = \text{vec}(dX)$
- $d\log X = X^{-1}dX$, 并且 $d\log(F(X)) = F^{-1}(X)d(F(X))$
- $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

3) 标量函数的矩阵微分

考虑标量函数 $f(x)$ 的全微分：

$$\begin{aligned} df(\mathbf{x}) &= \frac{\partial f(\mathbf{x})}{\partial x_1} dx_1 + \dots + \frac{\partial f(\mathbf{x})}{\partial x_m} dx_m \\ &= \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_m} \right] \begin{bmatrix} dx_1 \\ \vdots \\ dx_m \end{bmatrix} \\ &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} d\mathbf{x} \end{aligned}$$

令 $\mathbf{A} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T}$ ，则有：

$$df(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} d\mathbf{x} = \mathbf{A} d\mathbf{x} = \text{tr}(\mathbf{A} d\mathbf{x})$$

这表明，标量函数 $f(\mathbf{x})$ 的 Jacobian 矩阵与微分矩阵之间存在等价关系：

$$df(\mathbf{x}) = \text{tr}(\mathbf{A} d\mathbf{x}) \Leftrightarrow D_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \mathbf{A}$$

换言之，若函数 $f(\mathbf{x})$ 的微分可以写作 $df(\mathbf{x}) = \text{tr}(\mathbf{A} d\mathbf{x})$ ，则矩阵 \mathbf{A} 就是函数 $f(\mathbf{x})$ 关于其变元向量 \mathbf{x} 的 Jacobian 矩阵。

考虑标量函数 $f(\mathbf{X})$ ，其变元为 $\mathbf{X} \in R^{m \times n}$ 的全微分：

$$\begin{aligned} df(\mathbf{X}) &= \frac{\partial f(\mathbf{X})}{\partial x_{11}} dx_{11} + \dots + \frac{\partial f(\mathbf{X})}{\partial x_{mn}} dx_{mn} \\ &= \left[\frac{\partial f(\mathbf{X})}{\partial x_{11}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{m1}} \right] \begin{bmatrix} dx_{11} \\ \vdots \\ dx_{m1} \end{bmatrix} + \dots + \left[\frac{\partial f(\mathbf{X})}{\partial x_{1n}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \right] \begin{bmatrix} dx_{1n} \\ \vdots \\ dx_{mn} \end{bmatrix} \\ &= \left[\frac{\partial f(\mathbf{X})}{\partial x_{11}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{m1}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{1n}}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \right] \begin{bmatrix} dx_{11} \\ \vdots \\ dx_{m1} \\ \vdots \\ dx_{1n} \\ \vdots \\ dx_{mn} \end{bmatrix} \\ &= \frac{\partial f(\mathbf{x})}{\partial \text{vec}^T(\mathbf{X})} d(\text{vec} \mathbf{X}) = D_{\text{vec} \mathbf{X}} f(\mathbf{X}) d(\text{vec} \mathbf{X}) \end{aligned}$$

利用行向量偏导与 Jacobian 矩阵的关系 $D_{\text{vec} \mathbf{X}} f(\mathbf{X}) = (\text{vec}(D_{\mathbf{X}}^T f(\mathbf{X})))^T$ ，上式可以改写为：

$$df(\mathbf{X}) = (\text{vec}(\mathbf{A}^T))^T d(\text{vec} \mathbf{X})$$

式中 $\mathbf{A} = D_{\mathbf{X}} f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}^T}$ 是标量函数 $f(\mathbf{X})$ 的 Jacobian 矩阵。

利用向量化算子和迹函数之间的关系 $\text{tr}(\mathbf{B}^T \mathbf{C}) = (\text{vec}(\mathbf{B}))^T \text{vec}(\mathbf{C})$ ，令 $\mathbf{B} = \mathbf{A}^T$ ， $\mathbf{C} = d\mathbf{X}$ ，则有：

$$df(\mathbf{X}) = (\text{vec}(\mathbf{A}^T))^T d(\text{vec} \mathbf{X}) = \text{tr}(\mathbf{A} d\mathbf{X})$$

综上：

$$df(\mathbf{x}) = \text{tr}(\mathbf{A} d\mathbf{x}) \Leftrightarrow D_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \mathbf{A}, \nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}^T$$

$$df(\mathbf{X}) = \text{tr}(\mathbf{A}d\mathbf{X}) \Leftrightarrow D_{\mathbf{X}}f(\mathbf{X}) = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}^T} = \mathbf{A}, \nabla_{\mathbf{X}}f(\mathbf{X}) = \mathbf{A}^T$$

因此可以利用矩阵微分直接求标量函数 $f(\mathbf{X})$ 的 Jacobian 矩阵 $D_{\mathbf{X}}f(\mathbf{X})$ 和梯度矩阵 $\nabla_{\mathbf{X}}f(\mathbf{X})$ 。

(1) 二次型函数 $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$

标量函数可直接写成迹的形式：

$$\begin{aligned} f(\mathbf{x}) &= \text{tr}(f(\mathbf{x})) \\ d(f(\mathbf{x})) &= d(\text{tr}(\mathbf{x}^T \mathbf{A} \mathbf{x})) \\ &= \text{tr}[(d\mathbf{x})^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{A} d\mathbf{x}] \\ &= \text{tr}[(d\mathbf{x}^T \mathbf{A} \mathbf{x})^T + \mathbf{x}^T \mathbf{A} d\mathbf{x}] \\ &= \text{tr}(\mathbf{x}^T \mathbf{A}^T d\mathbf{x} + \mathbf{x}^T \mathbf{A} d\mathbf{x}) \\ &= \text{tr}(\mathbf{x}^T (\mathbf{A} + \mathbf{A}^T) d\mathbf{x}) \end{aligned}$$

根据 $df(\mathbf{x}) = \text{tr}(\mathbf{A}d\mathbf{x}) \Leftrightarrow D_{\mathbf{x}}f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \mathbf{A}, \nabla_{\mathbf{x}}f(\mathbf{x}) = \mathbf{A}^T$ 可得：

$$\nabla_{\mathbf{x}}f(\mathbf{x}) = \mathbf{A}^T = [\mathbf{x}^T (\mathbf{A} + \mathbf{A}^T)]^T = (\mathbf{A}^T + \mathbf{A})\mathbf{x}$$

(2) 矩阵的标量函数：迹 $f(\mathbf{X}) = \text{tr}(\mathbf{X}^T \mathbf{X})$

$$\begin{aligned} d\text{tr}(\mathbf{X}^T \mathbf{X}) &= \text{tr}(d(\mathbf{X}^T \mathbf{X})) \\ &= \text{tr}(((d\mathbf{X})^T \mathbf{X}) + \mathbf{X}^T d\mathbf{X}) \\ &= \text{tr}((d\mathbf{X})^T \mathbf{X}) + \text{tr}(\mathbf{X}^T d\mathbf{X}) \end{aligned}$$

因为 $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T)$ ，所以有：

$$\begin{aligned} d\text{tr}(\mathbf{X}^T \mathbf{X}) &= \text{tr}((d\mathbf{X})^T \mathbf{X}) + \text{tr}(\mathbf{X}^T d\mathbf{X}) \\ &= \text{tr}(((d\mathbf{X})^T \mathbf{X})^T) + \text{tr}(\mathbf{X}^T d\mathbf{X}) \\ &= \text{tr}(2\mathbf{X}^T d\mathbf{X}) \\ \nabla_{\mathbf{X}}f(\mathbf{X}) &= \frac{\partial \text{tr}(\mathbf{X}^T \mathbf{X})}{\partial \mathbf{X}} = (2\mathbf{X}^T)^T = 2\mathbf{X} \end{aligned}$$

(3) 三矩阵乘积的迹函数 $f(\mathbf{X}) = \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})$

$$\begin{aligned} d\text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X}) &= \text{tr}(d(\mathbf{X}^T \mathbf{A} \mathbf{X})) \\ &= \text{tr}(((d\mathbf{X})^T \mathbf{A} \mathbf{X}) + \mathbf{X}^T \mathbf{A} d\mathbf{X}) \\ &= \text{tr}((d\mathbf{X})^T \mathbf{A} \mathbf{X}) + \text{tr}(\mathbf{X}^T \mathbf{A} d\mathbf{X}) \\ &= \text{tr}(((d\mathbf{X})^T \mathbf{A} \mathbf{X})^T) + \text{tr}(\mathbf{X}^T \mathbf{A} d\mathbf{X}) \\ &= \text{tr}((\mathbf{A} \mathbf{X})^T d\mathbf{X}) + \text{tr}(\mathbf{X}^T \mathbf{A} d\mathbf{X}) \\ &= \text{tr}(\mathbf{X}^T (\mathbf{A}^T + \mathbf{A}) d\mathbf{X}) \end{aligned}$$

$$\nabla_{\mathbf{X}}f(\mathbf{X}) = \frac{\partial \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})}{\partial \mathbf{X}} = (\mathbf{X}^T (\mathbf{A}^T + \mathbf{A}))^T = (\mathbf{A} + \mathbf{A}^T)\mathbf{X}$$

(4) 包含逆矩阵的迹函数 $f(\mathbf{X}) = \text{tr}(\mathbf{A} \mathbf{X}^{-1})$

$$\begin{aligned}
 d\text{tr}(\mathbf{A}\mathbf{X}^{-1}) &= \text{tr}(d(\mathbf{A}\mathbf{X}^{-1})) \\
 &= \text{tr}(\mathbf{A}d(\mathbf{X}^{-1})) \\
 &= -\text{tr}(\mathbf{A}\mathbf{X}^{-1}d(\mathbf{X})\mathbf{X}^{-1}) \\
 &= -\text{tr}(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}d(\mathbf{X})) \\
 \nabla_{\mathbf{X}}f(\mathbf{X}) &= \frac{\partial \text{tr}(\mathbf{A}\mathbf{X}^{-1})}{\partial \mathbf{X}} = (\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1})^T
 \end{aligned}$$

(5) 四矩阵乘积的迹函数 $f(\mathbf{X}) = \text{tr}(\mathbf{X}\mathbf{A}\mathbf{X}\mathbf{B})$

$$\begin{aligned}
 d\text{tr}(\mathbf{X}\mathbf{A}\mathbf{X}\mathbf{B}) &= \text{tr}(d(\mathbf{X}\mathbf{A}\mathbf{X}\mathbf{B})) \\
 &= \text{tr}(d(\mathbf{X})\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{X}d(\mathbf{A})\mathbf{X}\mathbf{B} + \mathbf{X}\mathbf{A}d(\mathbf{X})\mathbf{B}) \\
 &= \text{tr}((\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{B}\mathbf{X}\mathbf{A})d(\mathbf{X})) \\
 \nabla_{\mathbf{X}}f(\mathbf{X}) &= \frac{\partial \text{tr}(\mathbf{X}\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = (\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{B}\mathbf{X}\mathbf{A})^T
 \end{aligned}$$

(6) 总结

标量函数总可以写为迹函数的形式：

$$f(\mathbf{X}) = \text{tr}(f(\mathbf{X}))$$

无论 $d\mathbf{X}$ 出现在迹函数内的任何位置，总可以通过迹的交换律将其写到最右边：

$$\text{tr}[\mathbf{A}(d\mathbf{X})\mathbf{B}] = \text{tr}[\mathbf{B}\mathbf{A}d\mathbf{X}]$$

对于 $(d\mathbf{X})^T$ ，总可以通过迹函数的性质转换为 $d\mathbf{X}$ ：

$$\text{tr}[\mathbf{A}(d\mathbf{X})^T\mathbf{B}] = \text{tr}[\mathbf{A}^T\mathbf{B}^Td\mathbf{X}]$$

第 3 章 概率论

3.1 概率

3.1.1 概率的概念

概率是对事件发生的可能性的度量。通常将事件 A 的概率写作 $P(A)$ 。

3.1.2 概率的计算

事件	概率
A	$P(A) \in [0,1]$
非 A	$P(\bar{A}) = 1 - P(A)$
A 和 B (联合概率)	$P(A \cap B) = P(A B)P(B) = P(B A)P(A)$ 当 A 、 B 相互独立时： $P(A \cap B) = P(A)P(B)$
A 或 B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 当 A 、 B 互斥时： $P(A \cup B) = P(A) + P(B)$

B 的情况下 A 的概率 (条件概率)	$P(A B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B A)P(A)}{P(B)}$
-------------------------------	---------------------------------------------------------------

例如：现有一个装有 10 个球的袋子，其中有 6 个红球和 4 个蓝球。从中随机抽取两个球。我们定义以下事件：

事件 A：第一个抽到的是红球。

事件 B：两个抽到的球都是红球。

1) 计算联合概率 $P(A \cap B)$

第一个球是红球的概率：

$$P(A) = \frac{6}{10}$$

在第一个球是红球的情况下，两个球都是红球的概率：

$$P(B|A) = \frac{5}{9}$$

联合概率：

$$P(A \cap B) = P(B|A)P(A) = \frac{5}{9} \times \frac{6}{10} = \frac{1}{3}$$

2) 计算条件概率 $P(A|B)$

条件概率 $P(A|B)$ 表示在已知两个球都是红球的情况下，第一个球是红球的概率。

两个球都是红球的概率：

$$P(B) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}$$

在两个球都是红球的情况下，第一个球是红球的概率：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{1}{3}} = 1$$

3.2 概率分布

概率分布，是指用于表述随机变量取值的概率规律。事件的概率表示了一次试验中某一个结果发生的可能性大小。如果试验结果用变量 X 的取值来表示，则随机试验的概率分布就是随机变量的概率分布，即随机变量的可能取值及取得对应值的概率。

3.2.1 均匀分布

均匀分布也叫矩形分布，它表示在相同长度间隔的分布概率是等可能的。均匀分布由两个参数 a 和 b 定义，它们是数轴上的最小值和最大值，通常缩写为 U (a, b)。

均匀分布的概率密度函数可写为：

$$P(x) = \frac{1}{b-a}, \quad a < x < b$$

$$P(x) = 0, \quad \text{else}$$

3.2.2 伯努利分布

伯努利分布（Bernoulli Distribution）又称两点分布或 0-1 分布，是一个离散概率分布，适用于只有两种可能结果的随机试验。

伯努利分布描述了一个随机变量 x ，它只有两个可能的取值：

- $x = 1$ ，发生概率为 p 。
- $x = 0$ ，发生概率为 $1 - p$ 。

伯努利分布的概率密度函数可写为：

$$P(x) = p^x(1-p)^{1-x}$$

3.2.3 二项分布

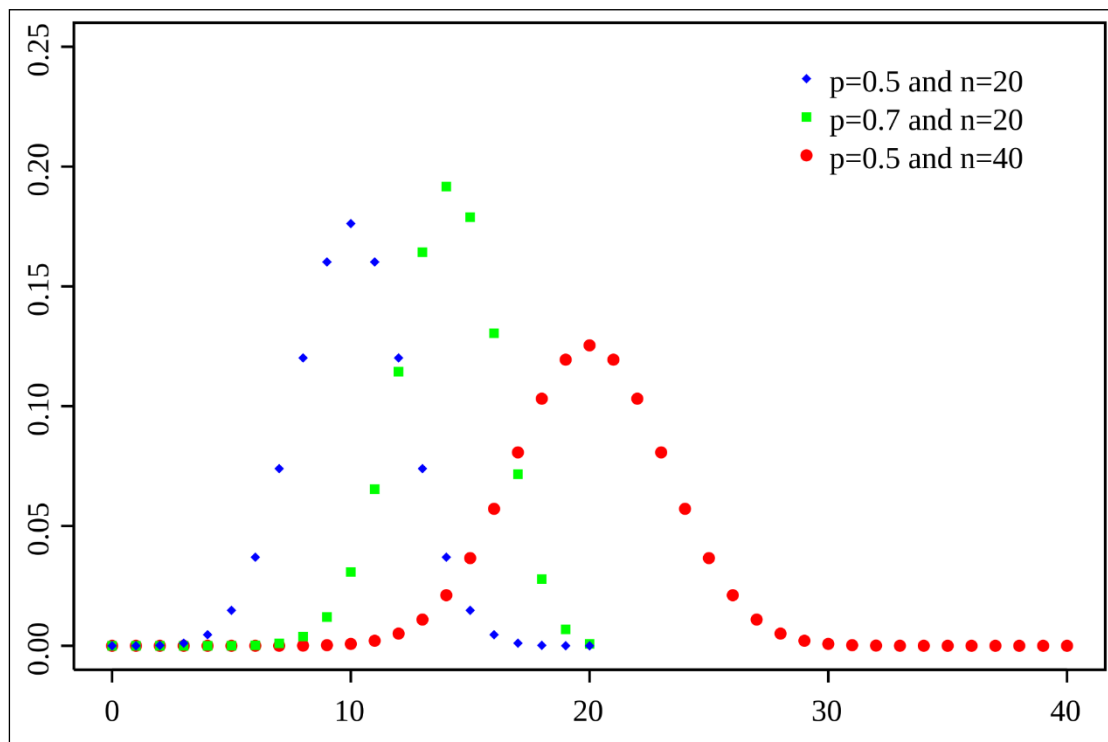
二项分布（Binomial Distribution）是一种离散概率分布，它描述了在固定次数的独立伯努利试验（每次试验只有成功和失败两种结果）中，成功次数的概率。二项分布可以看作是伯努利分布的扩展，从单次试验推广到多次试验。

对于随机变量 X ，有概率密度函数：

$$P(X = k) = \binom{n}{k} p^k (1-p)^{1-k} \quad (k = 0, 1, \dots, n)$$

- n 为试验的总次数
- k 为成功的次数
- p 为单次试验成功的概率
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ 表示从 n 次试验中选择 k 次成功的组合数

二项分布概率密度函数图：



3.2.4 正态分布

正态分布（Normal Distribution）也称高斯分布，是常见的连续概率分布。正态分布在统计学上十分重要，经常用在自然和社会科学来代表一个不明的随机变量。

若随机变量 X 服从一个平均数为 μ 、标准差为 σ （ $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$ ）的正态分布，则记为 $X \sim N(\mu, \sigma^2)$ ，其概率密度函数

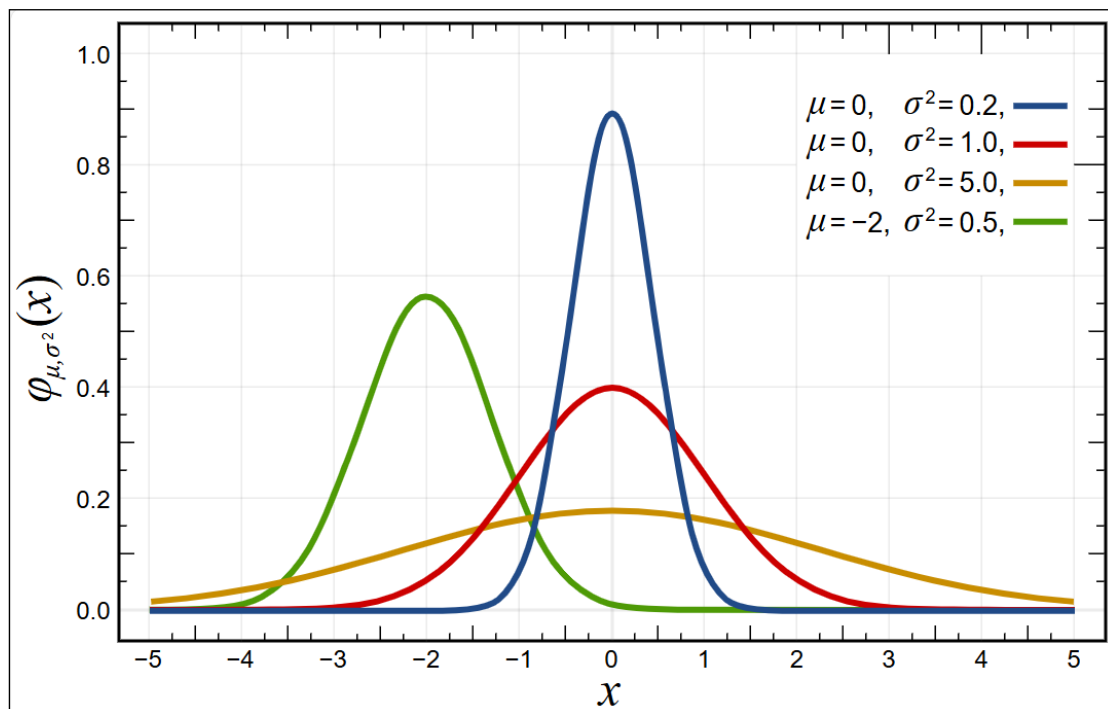
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正态分布的期望 μ 可解释为位置参数，决定了分布的位置；其方差 σ^2 可解释为尺度参数，决定了分布的幅度。

中心极限定理指出，在特定条件下，一个具有有限均值和方差的随机变量的多个样本的平均值本身就是一个随机变量，其分布随着样本数量的增加而收敛于正态分布。因此，许多与独立过程总和有关的物理量，例如测量误差，通常可被近似为正态分布。

正态分布的概率密度函数曲线呈钟形，因此人们又经常称之为钟形曲线。我们通常所说的标准正态分布是位置参数 $\mu = 0$ ，尺度参数 $\sigma^2 = 1$ 的正态分布。

正态分布概率密度函数图：



在 numpy 中，提供了各种随机函数，可以用来生成服从特定分布的数据。

3.3 贝叶斯定理

贝叶斯定理（Bayes' Theorem）是概率论中的一个核心定理，用于描述在已有条件概率信息的基础上，如何更新或计算事件的概率。它以英国数学家托马斯·贝叶斯的名字命名。贝叶斯定理特别适合处理“逆向概率”问题，即从结果反推原因的概率。

贝叶斯定理建立在条件概率的基础上，假设有两事件 A, B ，贝叶斯定理描述了在已知 B 发生的情况下， A 发生的概率：

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- $P(A|B)$ ：后验概率， B 发生的情况下 A 发生的概率。
- $P(B|A)$ ：似然概率， A 发生的情况下 B 发生的概率。
- $P(A)$ ：先验概率， A 发生的概率。
- $P(B)$ ： B 发生的概率，通常通过全概率公式计算。

在实际问题中 $P(B)$ 通常不是直接给出，而是需要通过全概率公式计算。假设样本空间被一组互斥且完备的事件 A_1, A_2, \dots, A_n 划分，则：

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$$

例如：某疾病发病率为 1%，如果一个人有疾病，检测呈阳性的概率为 95%；如果一个人没有疾病，检测呈阳性的概率为 5%，现有一人检测结果呈阳性，问他真正患病的概率是多少？

$$\begin{aligned}
 P(\text{疾病}|\text{阳性}) &= \frac{P(\text{阳性}|\text{疾病}) \cdot P(\text{疾病})}{P(\text{阳性})} \\
 P(\text{阳性}) &= P(\text{阳性}|\text{疾病}) \cdot P(\text{疾病}) + P(\text{阳性}|\text{无疾病}) \cdot P(\text{无疾病}) \\
 &= 0.95 \times 0.01 + 0.05 \times 0.99 \\
 &= 0.0095 + 0.0495 \\
 &= 0.059 \\
 P(\text{疾病}|\text{阳性}) &= \frac{0.0095}{0.059} \approx 0.161
 \end{aligned}$$

检测呈阳性的人真正患病的概率为 16.1%。

3.4 似然函数

3.4.1 似然函数的概念

概率用于在已知一些参数的情况下，预测接下来在观测上所得到的结果。而似然性则是用于在已知某些观测所得到的结果时，对有关事物之性质的参数进行估值。

似然函数是对参数的函数，其定义为在给定参数值的条件下，观察到某个特定数据的概率。换句话说，似然函数是一个关于参数的函数，而不是关于数据的函数。

如果我们有一个参数化的概率模型 $P(X|\theta)$ ，其中 X 是观测数据， θ 是模型参数，似然函数 $L(\theta|X)$ 定义为：

$$L(\theta|X) = P(X|\theta)$$

这里， $P(X|\theta)$ 表示在参数为 θ 的情况下，观察到数据 X 的概率。

设有一组独立同分布的观测数据 $X = (x_1, x_2, \dots, x_n)$ ，并且这些数据服从某个分布（例如正态分布、二项分布等），比如服从参数为 θ 的某个分布，那么似然函数可以写作：

$$L(\theta|X) = P(X|\theta) = \prod_{i=1}^n P(x_i|\theta)$$

针对其中存在的乘法，可以使对数函数将其转化为加法：

$$\log L(\theta|X) = \log \prod_{i=1}^n P(x_i|\theta) = \sum_{i=1}^n \log P(x_i|\theta)$$

3.4.2 极大似然估计

似然函数常用于极大似然估计。我们希望找到使似然函数最大化的参数 θ 。这意味着在给定观测数据的情况下，选择最可能生成这些数据的参数值。

例如，掷硬币 3 次，2 次正面 1 次背面，能否依据此结果逆推出正面的概率；正面概率为 0.5 的概率为多少、正面概率为 0.6 的概率为多少；最有可能的正面概率是多少？

我们用 θ 代表硬币正面朝上的概率，用 X 代表 2 次正面 1 次背面的结果

$$L(\theta | X) = P(X | \theta) = C_3^2 \theta^2 (1 - \theta)$$

$$\text{当正面概率为 } 0.5 \text{ 时: } P(X | \theta = 0.5) = C_3^2 0.5^2 (1 - 0.5) = 0.375$$

$$\text{当正面概率为 } 0.6 \text{ 时: } P(X | \theta = 0.6) = C_3^2 0.6^2 (1 - 0.6) = 0.432$$

为了找出极大似然估计，对似然函数取对数并求导，使其等于 0

$$\begin{aligned} \log L(\theta | X) &= \log C_3^2 \cdot \theta^2 (1 - \theta) \\ &= \log 3 + 2 \log \theta + \log(1 - \theta) \\ \frac{d \log L(\theta | X)}{d \theta} &= \frac{2}{\theta} - \frac{1}{1 - \theta} = 0 \end{aligned}$$

解得 $\theta = \frac{2}{3}$ ，意味着当掷硬币 3 次，出现 2 次正面 1 次背面的结果时，硬币正面朝上的概率最有可能为 $\frac{2}{3}$ 。