

## 17

## Derivatives of Multivariable Functions

## 多元函数微分

将偏微分延伸到高维和任意方向



数学的终极目标是人类精神的荣誉。

*The object of mathematics is the honor of the human spirit.*

—— 卡尔·雅可比 (Carl Jacobi) | 普鲁士数学家 | 1804 ~ 1851



- ◀ `numpy.meshgrid()` 获得网格数据
- ◀ `numpy.multiply()` 向量或矩阵逐项乘积
- ◀ `numpy.roots()` 多项式求根
- ◀ `numpy.sqrt()` 平方根
- ◀ `sympy.abc import x` 定义符号变量 `x`
- ◀ `sympy.diff()` 求解符号导数和偏导解析式
- ◀ `sympy.Eq()` 定义符号等式
- ◀ `sympy.evalf()` 将符号解析式中未知量替换为具体数值
- ◀ `sympy.plot_implicit()` 绘制隐函数方程
- ◀ `sympy.symbols()` 定义符号变量

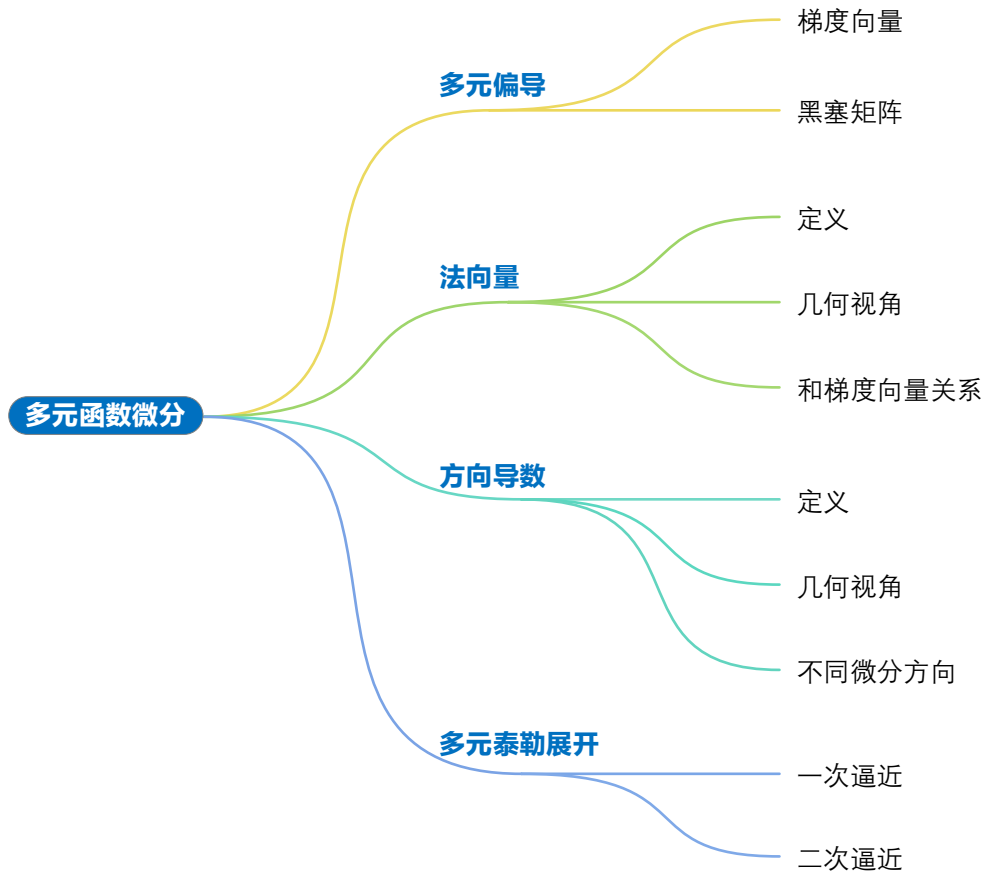
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)



## 17.1 偏导：特定方向的变化率

### 回顾偏导



本系列丛书《数学要素》第 16 章讲过偏导数 (partial derivative) 内容。

一个多变量的函数的偏导数是函数关于其中一个变量的导数，而保持其他变量恒定。白话说，偏导数关注曲面某个特定方向上的变化率。换个角度，一元函数导数这个工具改造成偏导数后，可以用在多元函数上。

下面以二元函数为例回顾偏导数定义。设  $f(x_1, x_2)$  是定义在平面  $\mathbb{R}^2$  上的二元函数， $f(x_1, x_2)$  在点  $(a, b)$  的某一邻域内有定义。

图 1 (a) 网格面为  $f(x_1, x_2)$  函数曲面，平行  $x_1$  轴在  $x_2 = b$  切一刀得到浅蓝色剖面，偏导  $f_{x_1}(a, b)$  就是浅蓝色剖面在  $(a, b)$  一点的切线斜率。该切线平行  $x_1y$  平面。

同理，如图 1 (b) 所示，平行于  $x_2$  轴在  $x_1 = a$  切一刀，偏导  $f_{x_2}(a, b)$  就是浅蓝色剖面在  $(a, b)$  一点的切线斜率。该切线平行  $x_2y$  平面。

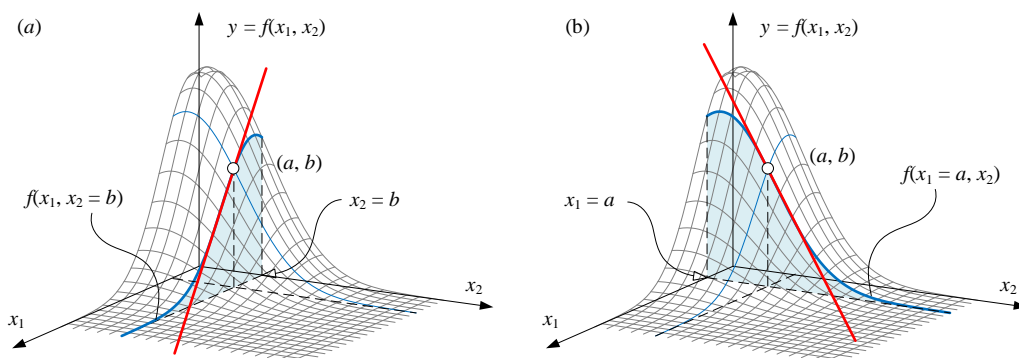


图 1.  $f(x_1, x_2)$  偏导定义，图片来自《数学要素》

### 向量形式

为了方便表达和运算，我们可以把上述二元函数在  $x_1$  和  $x_2$  方向上的偏导写成列向量形式：

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} \quad (1)$$

其中， $\mathbf{x}$  为列向量， $\mathbf{x} = [x_1, x_2]^T$ 。

## 一次函数

给定如下多元一次函数  $f(\mathbf{x})$ :

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \mathbf{x}^T \mathbf{w} + b \quad (2)$$

其中,  $\mathbf{x}$  和  $\mathbf{w}$  均为列向量:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \quad (3)$$

(2) 展开得到大家熟悉的一次函数形式:

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_D x_D + b \quad (4)$$

从空间角度来看, 当  $b = 0$  时, 上式代表的超平面通过原点, 可以看做是向量空间; 当  $b \neq 0$  时, 超平面不过原点, 上式可以视作仿射空间。

(2) 多元一次函数  $f(\mathbf{x})$  对  $\mathbf{x}$  求一阶导, 并写成列向量形式:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_D} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} = \mathbf{w} \quad (5)$$

本章后文会给上式一个新的名字——**梯度向量** (gradient vector)。另外, 请大家注意以下等价关系:

$$\frac{\partial(\mathbf{w}^T \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T \mathbf{w})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{w} \cdot \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x} \cdot \mathbf{w})}{\partial \mathbf{x}} = \frac{\partial \langle \mathbf{w}, \mathbf{x} \rangle}{\partial \mathbf{x}} = \mathbf{w} \quad (6)$$

## 二次函数

给定如下二次函数:

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{x} = x_1^2 + x_2^2 + \cdots + x_D^2 \quad (7)$$

从几何角度来看, 上式是多元空间的正圆抛物面。特别地,  $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x} = c$  ( $c > 0$ ) 时, 上式代表  $D$  维正球体。

(7) 对向量  $\mathbf{x}$  求一阶导：

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_D} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_D \end{bmatrix} = 2\mathbf{x} \quad (8)$$

要是类比的话， $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x}$  相当于  $f(x) = x^2$ 。而上式相当于  $f(x)$  的一阶导数  $f'(x) = 2x$ 。

(8) 等价于：

$$\frac{\partial (\mathbf{x}^T \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x} \cdot \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \langle \mathbf{x}, \mathbf{x} \rangle}{\partial \mathbf{x}} = \frac{\partial (\|\mathbf{x}\|_2^2)}{\partial \mathbf{x}} = 2\mathbf{x} \quad (9)$$

## 二次型

给定如下二次型：

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{Q} \mathbf{x} \quad (10)$$

(10) 对  $\mathbf{x}$  求一阶导：

$$\frac{\partial (\mathbf{x}^T \mathbf{Q} \mathbf{x})}{\partial \mathbf{x}} = (\mathbf{Q} + \mathbf{Q}^T) \mathbf{x} \quad (11)$$

⚠ 注意， $\mathbf{Q}$  为常数方阵。

如果  $\mathbf{Q}$  为对称矩阵，(10) 对  $\mathbf{x}$  一阶导数可以写成：

$$\frac{\partial (\mathbf{x}^T \mathbf{Q} \mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{Q} \mathbf{x} \quad (12)$$

假设  $\mathbf{Q}$  为对称矩阵，给定如下二次函数：

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{w}^T \mathbf{x} + b \quad (13)$$

(13) 对  $\mathbf{x}$  求一阶导：

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{Q} \mathbf{x} + \mathbf{w} \quad (14)$$

举个形似 (13) 的例子：

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}}_{\mathbf{Q}} \mathbf{x} + \underbrace{\begin{bmatrix} 4 \\ 5 \end{bmatrix}}_{\mathbf{w}}^T \mathbf{x} + 6 \quad (15)$$

(15) 向量  $\mathbf{x}$  求一阶导：

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 4 \\ 5 \end{bmatrix} \quad (16)$$

如下形式函数对向量  $\mathbf{x}$  求一阶导：

$$\frac{\partial \left( (\mathbf{x} - \mathbf{c})^T \mathbf{Q} (\mathbf{x} - \mathbf{c}) \right)}{\partial \mathbf{x}} = 2\mathbf{Q}(\mathbf{x} - \mathbf{c}) \quad (17)$$

其中， $\mathbf{Q}$  为对称矩阵。

## 二阶偏导：黑塞矩阵

**黑塞矩阵** (Hessian matrix) 是一个多元函数的二阶偏导数构成的方阵，黑塞矩阵描述了函数的局部曲率。黑塞矩阵由德国数学家**奥托·黑塞** (Otto Hesse) 引入并以其名字命名。

假设有一实值函数  $f(\mathbf{x})$ ，如果它的所有二阶偏导数都存在、并在定义域内连续，那么  $f(\mathbf{x})$  的黑塞矩阵  $\mathbf{H}$  为：

$$\mathbf{H} = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad (18)$$

注意， $\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_2} \left( \frac{\partial f}{\partial x_1} \right)$  代表先对  $x_1$ 、后对  $x_2$  二阶混合偏导。  
 $x_1 \rightarrow x_2$

(10) 中给定二次函数对向量  $\mathbf{x}$  求二阶导，获得黑塞矩阵：

$$\mathbf{H} = \frac{\partial^2 (\mathbf{x}^T \mathbf{Q} \mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \mathbf{Q} + \mathbf{Q}^T \quad (19)$$

如果  $\mathbf{Q}$  为对称，(19) 中黑塞矩阵为：

$$\mathbf{H} = \frac{\partial^2 (\mathbf{x}^T \mathbf{Q} \mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{Q} \quad (20)$$

以 (15) 为例，这个二元函数的黑塞矩阵为：

$$\mathbf{H} = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \quad (21)$$

本书后续会在优化问题中用到黑塞矩阵判断极值点。本节的内容可能会显得单调。本章后续将依托几何视角帮助大家理解本节内容。

## 17.2 梯度向量：上山方向

我们给上节讨论的一阶导数新名字——**梯度向量** (gradient vector)。函数  $f(\mathbf{x})$  的梯度向量定义如下：

$$\text{grad } f(\mathbf{x}) = \nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \quad (22)$$

梯度向量可以使用 `grad()` 作为运算符，也常使用**倒三角微分算子**  $\nabla$ ， $\nabla$  也叫 **Nabla 算子** (Nabla symbol)。

### 几何视角

从几何视角来看梯度向量，如图 2 所示，在坡面  $P$  点处放置一个小球，轻轻松手一瞬间，小球沿着坡面最陡峭方向 (绿色箭头) 滚下。瞬间滚动方向在平面上的投影方向便是**梯度下降方向** (direction of gradient descent)，也称“下山”方向。

数学中，下山方向的反方向即梯度向量方向，也称作“上山”方向。

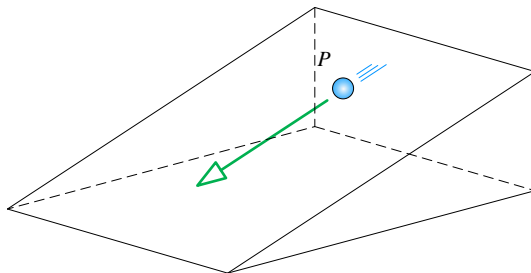


图 2. 梯度方向原理

## 二元函数

以二元函数为例， $f(x_1, x_2)$  某一点  $P$  处梯度向量为：

$$\nabla f(\mathbf{x}_p) = \text{grad } f(\mathbf{x}_p) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix}_{\mathbf{x}_p} \quad (23)$$

$P$  处于不同点时，可以得到**梯度向量场** (gradient vector field)。图 3 所示为某个函数梯度向量的分布。大家容易发现，梯度向量垂直所在位置等高线。某点梯度向量长度越长，即向量模越大，这说明该处越陡峭。相反，如果梯度向量模越小，这说明该点越平坦。特殊情况是，梯度向量为  $\mathbf{0}$  向量时，这一点便是驻点，该点切平面平行于水平面。

白话讲，把图 3 看成一幅地图的话，某点梯度向量指向的方向就是该点最陡峭的上山方向。梯度向量的垂直方向就是该点等高线切线。沿着等高线规划的路径运动，高度不变。

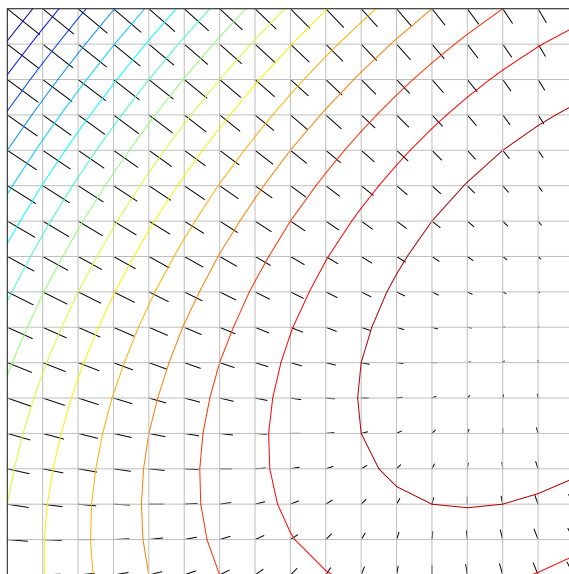


图 3. 梯度向量场

下面我们来看三个例子。

### 第一个例子：一次函数

给定二元一次  $f(x_1, x_2)$  函数如下：

$$f(x_1, x_2) = x_1 + x_2 \quad (24)$$



如图 2 (a) 所示, 这个函数在三维空间的形状是个平面。这个平面通过原点, 可以视作向量空间。

(24) 函数  $f(\mathbf{x})$  的梯度向量为:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (25)$$

观察上式, 容易发现二元一次函数的梯度向量的方向和大小不随位置改变, 具体如图 2 (b) 所示。不存在任何约束条件的话, 这个平面不存在任何极值点。沿着梯度向量方向运动, 函数值增大, 即上山。



本书第 19 章会专门讲解直线、平面和超平面。

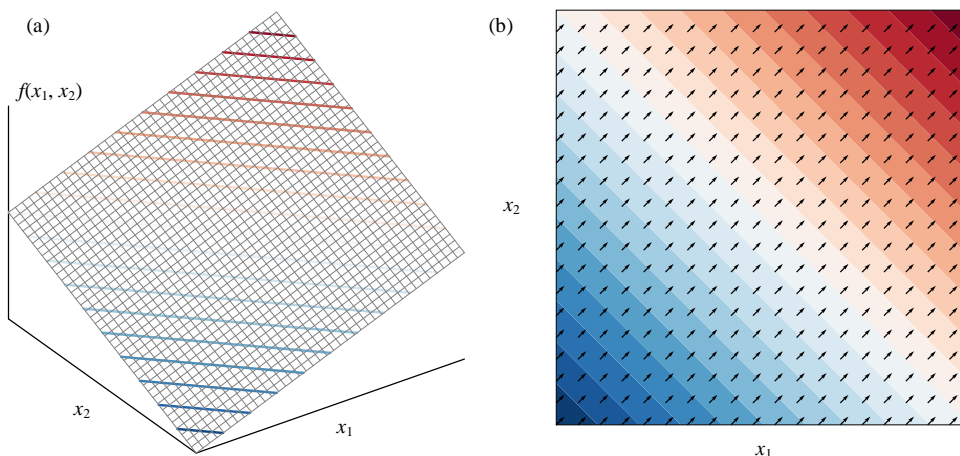


图 4. 平面的梯度向量场

## 第二个例子：二次函数

$f(x_1, x_2)$  为二元二次函数, 具体如下:

$$f(x_1, x_2) = x_1^2 + x_2^2 \quad (26)$$

图 5 (a) 告诉我们这个二元二次函数图像是个开口朝上的正圆抛物面, 曲面显然存在最小值点, 位于  $(0, 0)$ 。

(26) 函数  $f(\mathbf{x})$  的梯度向量定义如下:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \quad (27)$$

观察图 5 (b)，容易发现越靠近  $(0, 0)$ ，也就是最小值点附近，曲面梯度向量的模越小。在  $(0, 0)$  处，梯度向量为  $\mathbf{0}$ 。也就是说，该点处  $f(x_1, x_2)$  对  $x_1$  和  $x_2$  偏导数都为 0。显然  $\mathbf{0}$  是函数的最小值点。图 5 (b) 中不同点处梯度向量均垂直于等高线，指向背离最小值点，即上山方向。离  $\mathbf{0}$  越远，梯度向量模越大，曲面坡度越陡峭。

➡ 如果我们现在处于曲面上某一点，沿着下山方向一步步行走，最终我们会到达最小值点处。这个思路就是基于梯度的优化方法。当然，我们需要制定一个下山的策略。比如，下山的步伐怎么确定？路径怎么规划？怎么判定是否到达极值点？不同的基于梯度的优化方法在具体下山策略上有差别。这些内容，我们会在本系列丛书后续分册中讨论。

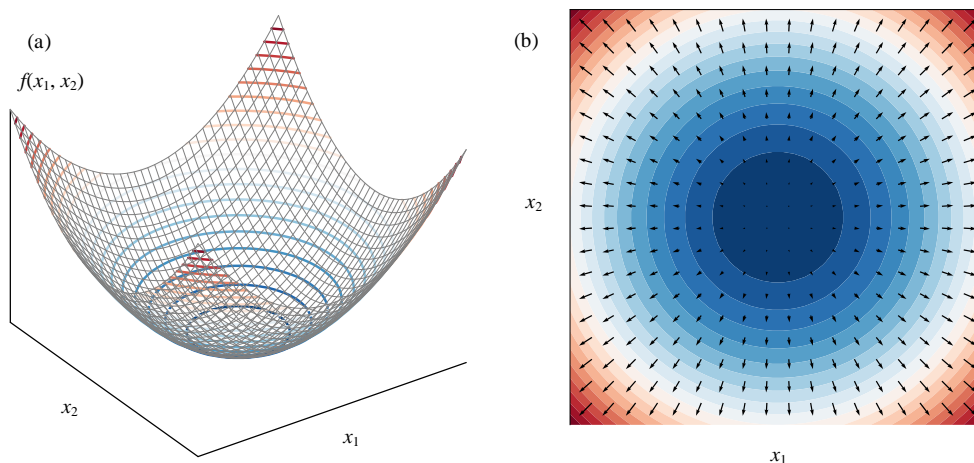


图 5. 正圆抛物面的向量场

### 第三个例子：复合函数

给定  $f(x_1, x_2)$  函数如下：

$$f(x_1, x_2) = x_1 \exp\left(-(x_1^2 + x_2^2)\right) \quad (28)$$

图 6 (a) 所示为函数曲面，它存在一个最大值点和一个最小值点。

函数  $f(\mathbf{x})$  的梯度向量定义如下：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2x_1^2 \exp(-(x_1^2 + x_2^2)) + \exp(-(x_1^2 + x_2^2)) \\ -2x_1 x_2 \exp(-(x_1^2 + x_2^2)) \end{bmatrix} \quad (29)$$

图 6 (b) 中，最大值点附近，梯度向量均指向最大值点。最小值点附近，梯度向量均背离最小值点。

在最大值和最小值点处，梯度向量都是  $\mathbf{0}$  向量。

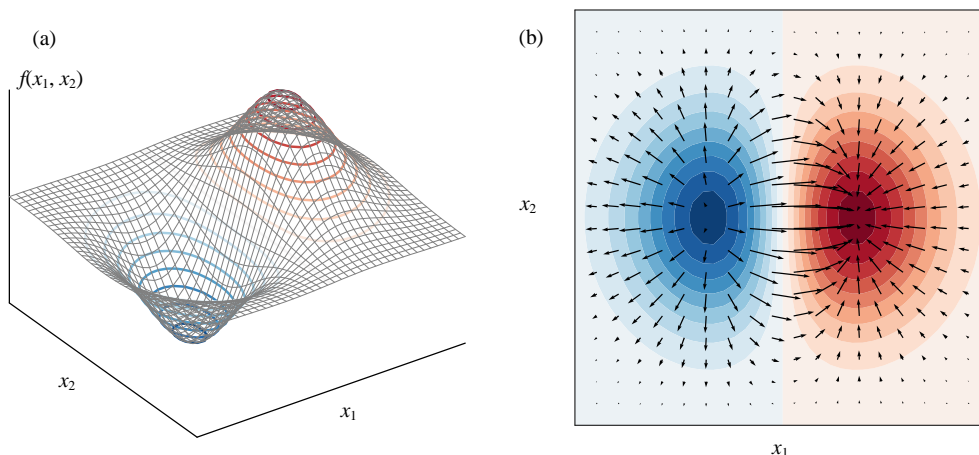


图 6.  $x_1 \exp(-(x_1^2 + x_2^2))$  的梯度向量场



请大家修改 `Bk4_Ch17_01.py` 并绘制图 4、图 5、图 6。



在 `Bk4_Ch17_01.py` 基础上，我们用 Streamlit 和 Plotly 制作了一个 App，用来交互可视化图 6 两幅图像。请大家参考 `Streamlit_Bk4_Ch17_01.py`。

## 17.3 法向量：垂直于切平面

对于  $y = f(\mathbf{x})$  函数，我们可以把它看做是等式  $f(\mathbf{x}) - y = 0$ 。定义  $F(\mathbf{x}, y)$  如下：

$$F(\mathbf{x}, y) = f(\mathbf{x}) - y \quad (30)$$

函数  $F(\mathbf{x}, y)$  梯度向量为：

$$\nabla F(\mathbf{x}, y) = \begin{bmatrix} \nabla f(\mathbf{x}) \\ -1 \end{bmatrix} \quad (31)$$

这个梯度向量就是  $f(\mathbf{x})$  点  $\mathbf{x}$  处曲面的法向量  $\mathbf{n}$ ：

$$\mathbf{n} = \begin{bmatrix} \nabla f(\mathbf{x}) \\ -1 \end{bmatrix} \quad (32)$$

如图 7 所示，以二元函数  $f(\mathbf{x})$  为例， $\mathbf{n}$  向水平面投影得到梯度向量  $\nabla f(\mathbf{x})$ 。

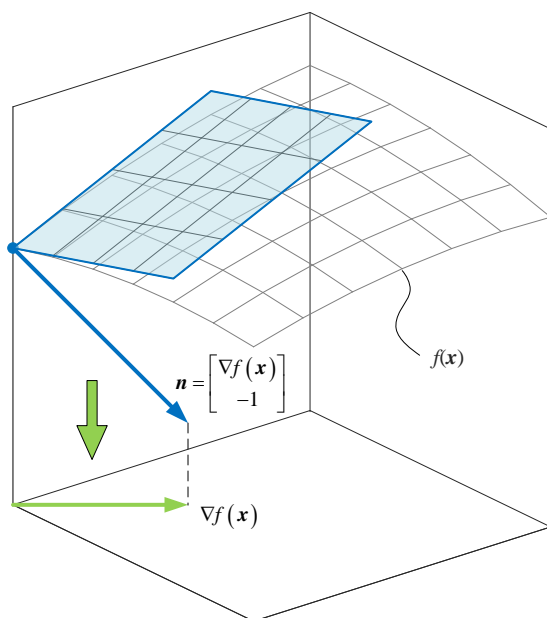


图 7.  $\mathbf{n}$  向水平面投影得到梯度向量

图 8 左图所示为某个二元函数  $f(\mathbf{x})$  曲面上不同点处的法向量，这些法向量向  $x_1x_2$  平面投影便得到  $f(\mathbf{x})$  的梯度向量，具体如图 8 右图所示。这个视角非常重要，本书第 21 章还会继续用到。

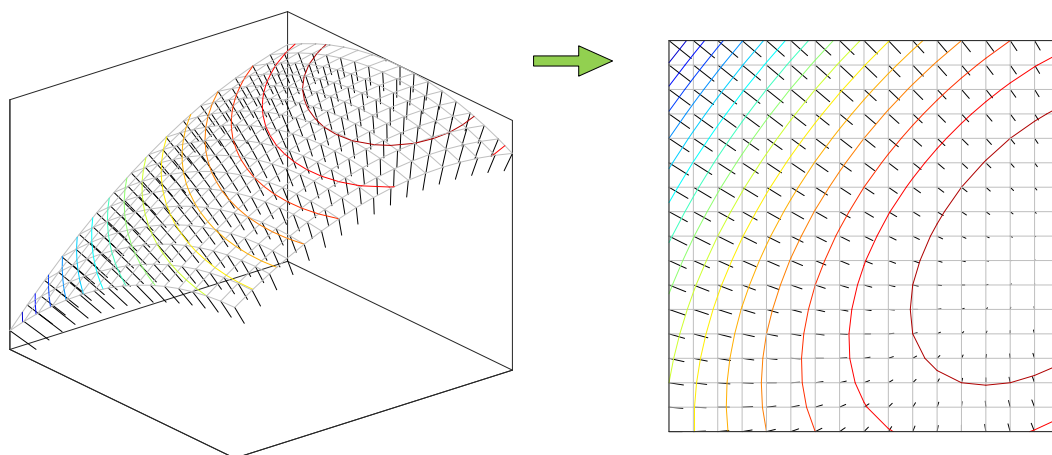
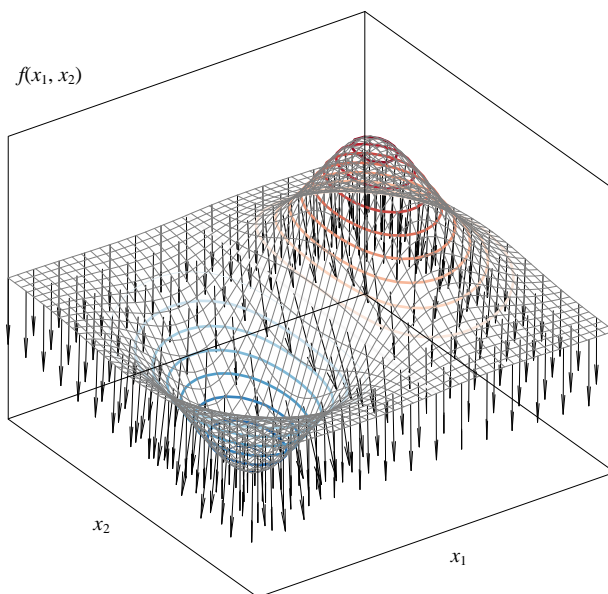


图 8. 曲面法向量场投影得到梯度向量场

图 9 给出的是 (28) 中函数在不同点处的法向量，这些向量朝水平面投影便得到图 6 (b)。曲面越陡峭，法向量在水平面投影的分量越多。举个极端例子，曲面某点处切面垂直于水平面，即坡度为 90 度，它的法线则平行于水平面。特别地，在极值点处，曲面的法向量垂直于水平面，因此在水平面的投影为零向量  $\mathbf{0}$ 。觉得图 9 不容易看的话，请大家参考图 10。

图 9.  $x_1 \exp(-(x_1^2 + x_2^2))$  的法向量场

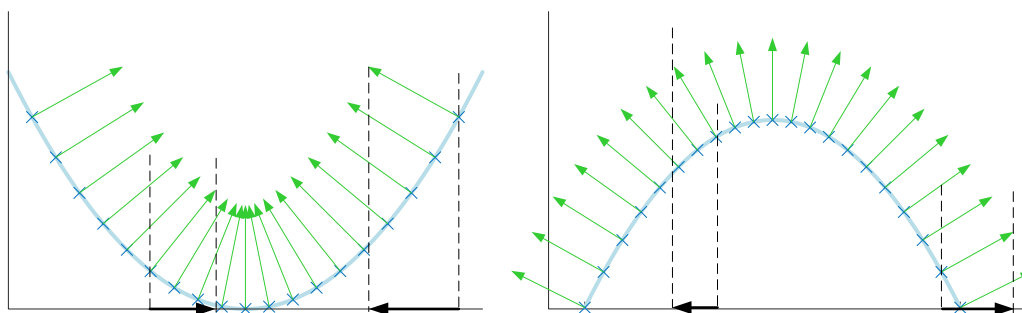


图 10. 曲线法向量在水平面上投影

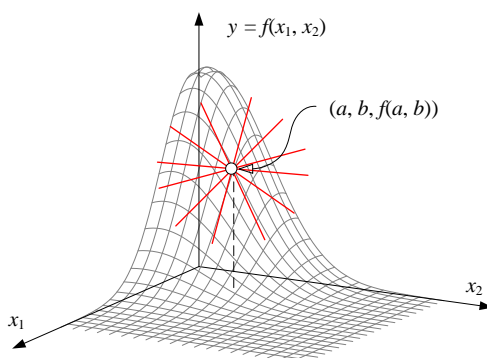


Bk4\_Ch17\_02.py 绘制图 9。

## 17.4 方向性微分：函数任意方向的变化率

《数学要素》一册提到过，光滑曲面  $f(x_1, x_2)$  某点的切线有无数条，如图 11 所示。而偏导数仅分析了其中两条切线的变化率，它们分别沿着  $x_1$  和  $x_2$  轴方向。

本节将介绍一个全新的数学工具——**方向性微分** (directional derivative)，它可以分析光滑曲面某点处不同方向切线的变化率。

图 11. 光滑曲面  $f(x_1, x_2)$  某点的切线有无数条

### 以二元函数为例

二元函数  $f(x_1, x_2)$  写作  $f(\mathbf{x})$ ：

$$f(\mathbf{x}) = f(x_1, x_2) \quad (33)$$

$P(x_1, x_2)$  点处，任意偏离  $P$  点微小移动  $(\Delta x_1, \Delta x_2)$  可能导致  $f(\mathbf{x})$  大小发生变化，函数值变化具体为：

$$\Delta f = f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) = f(x_1 + \Delta x_1, x_2 + \Delta x_2) - f(x_1, x_2) \quad (34)$$

如图 12 所示，曲面从  $P$  点移动到  $Q$  点高度变化就是上式中的  $\Delta f$ 。

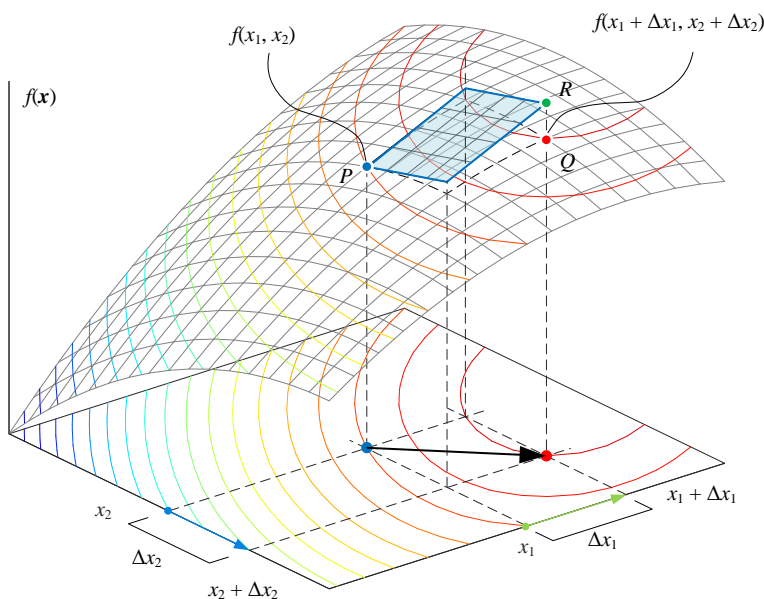


图 12. 曲面从  $P$  点移动到  $Q$  点对应位置变化

用一阶偏微分近似求解  $\Delta f$ :

$$\begin{aligned} \Delta f &= f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) \\ &= \underbrace{f(x_1 + \Delta x_1, x_2 + \Delta x_2)}_Q - \underbrace{f(x_1, x_2)}_R \approx \frac{\partial f(\mathbf{x})}{\partial x_1} \Delta x_1 + \frac{\partial f(\mathbf{x})}{\partial x_2} \Delta x_2 \end{aligned} \quad (35)$$

上式便是本系列丛书《数学要素》讲过的二元函数泰勒一阶展开。如图 12 所示，上式相当于用二元一次函数斜面（浅蓝色背景）近似函数曲面，即：

$$\underbrace{f(x_1 + \Delta x_1, x_2 + \Delta x_2)}_Q \approx \underbrace{f(x_1, x_2) + \frac{\partial f(\mathbf{x})}{\partial x_1} \Delta x_1 + \frac{\partial f(\mathbf{x})}{\partial x_2} \Delta x_2}_R \quad (36)$$

上式左侧代表  $Q$  点高度，右侧代表  $R$  点高度。两者之差就是估算误差。

## 几何视角

图 13 为图 12 局部放大图，这张图更清晰地展示估算过程。

在  $P(x_1, x_2)$  点处，二元函数曲面的高度为  $f(x_1, x_2)$ 。沿着蓝色斜面从  $P$  点运动到  $R$  点，我们把高度变化分成两步阶梯来看。沿着  $x_1$  方向上移动  $\Delta x_1$  带来的高度变化为  $\left. \frac{\partial f(\mathbf{x})}{\partial x_1} \right|_P \Delta x_1$ 。类似地，在  $x_2$  方向上移动  $\Delta x_2$  带来的高度变化为  $\left. \frac{\partial f(\mathbf{x})}{\partial x_2} \right|_P \Delta x_2$ 。两个高度变化之和便是对的  $\Delta f$  一阶逼近。

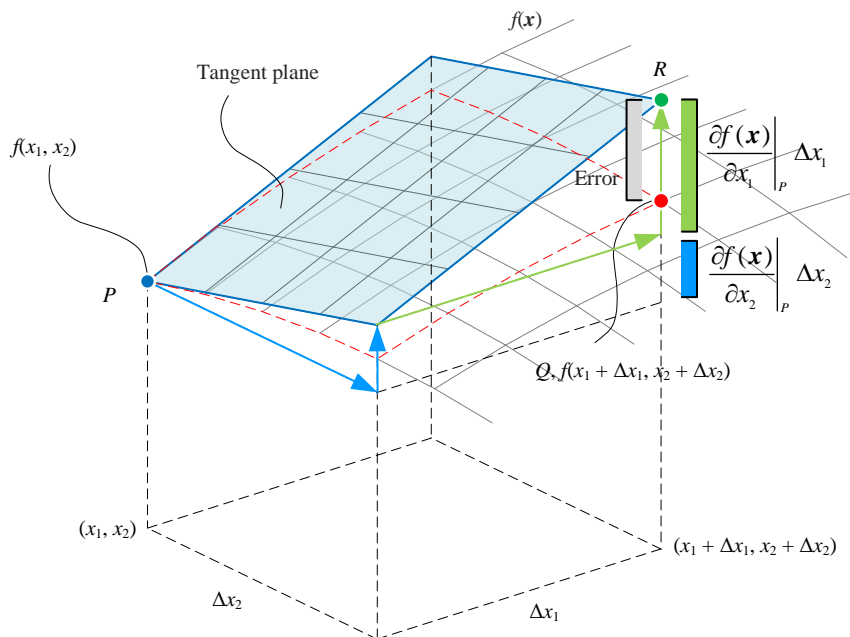


图 13. 二元函数一阶泰勒展开估算

(35) 可以写成两个向量内积关系：

$$\Delta f \approx \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix}^T \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} \quad (37)$$

换个角度，向量  $[\Delta x_1, \Delta x_2]^T$  决定了  $P$  点方向微分方向，如图 14 所示。

也就是说，有了向量  $[\Delta x_1, \Delta x_2]^T$ ，我们可以量化二元函数  $f(x_1, x_2)$  在任意方向的函数变化，以及变化率。

## 单位向量

$x_1 x_2$  平面上，给定一个方向，用单位向量  $\mathbf{v}$  表示：



$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (38)$$

令单位向量  $\mathbf{v}$  为：

$$\mathbf{v} = \begin{bmatrix} \cos \theta_1 \\ \cos \theta_2 \end{bmatrix} \quad (39)$$

图 14 给出  $\theta_1$  和  $\theta_2$  角度定义。可以这样理解单位向量  $\mathbf{v}$ ，模为 1 代表“一步”， $\mathbf{v}$  的方向代表运动方向。也就是说，单位向量  $\mathbf{v}$  确定了朝哪个方向运动一步。

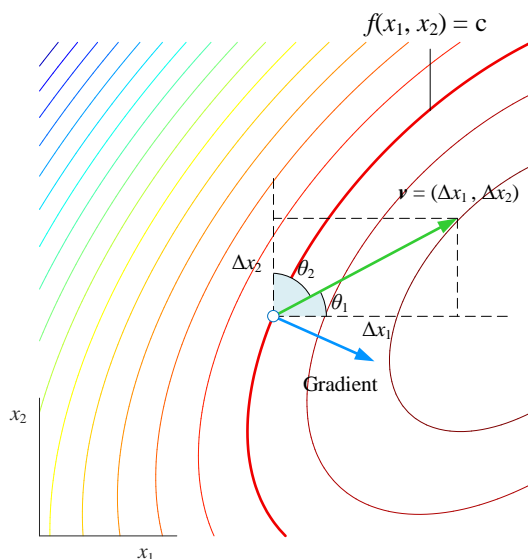


图 14.  $x_1x_2$  平面上方向微分

对于上述二元函数，定义方向性微分为：

$$\nabla_{\mathbf{v}} f(\mathbf{x}) = \mathbf{v} \cdot \nabla f(\mathbf{x}) = \mathbf{v}^T \nabla f(\mathbf{x}) = \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle \quad (40)$$

展开得到方向导数和偏导之间关系为：

$$\nabla_{\mathbf{v}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_1} \cos \theta_1 + \frac{\partial f(\mathbf{x})}{\partial x_2} \cos \theta_2 = \begin{bmatrix} \cos \theta_1 \\ \cos \theta_2 \end{bmatrix}^T \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} \quad (41)$$

(40) 也适用于多元函数。

## 不同方向

根据向量内积法则，(40) 可以写成：

$$\begin{aligned}\nabla_{\mathbf{v}} f(\mathbf{x}) &= \nabla f(\mathbf{x}) \cdot \mathbf{v} \\ &= \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{v}\| \cos(\theta) \\ &= \|\nabla f(\mathbf{x})\| \cos(\theta)\end{aligned}\tag{42}$$

其中， $\mathbf{v}$  为单位向量， $\theta$  为  $\nabla f(\mathbf{x})$  和  $\mathbf{v}$  之间相对夹角。

图 15 所示为  $x_1x_2$  平面上六种不同方向导数情况。

如图 15 (a) 和 (b) 所示，若  $\theta = 90^\circ$ ，方向导数垂直于梯度向量，(42) 为 0。这说明沿着等高线运动，函数值不会有任何变化。

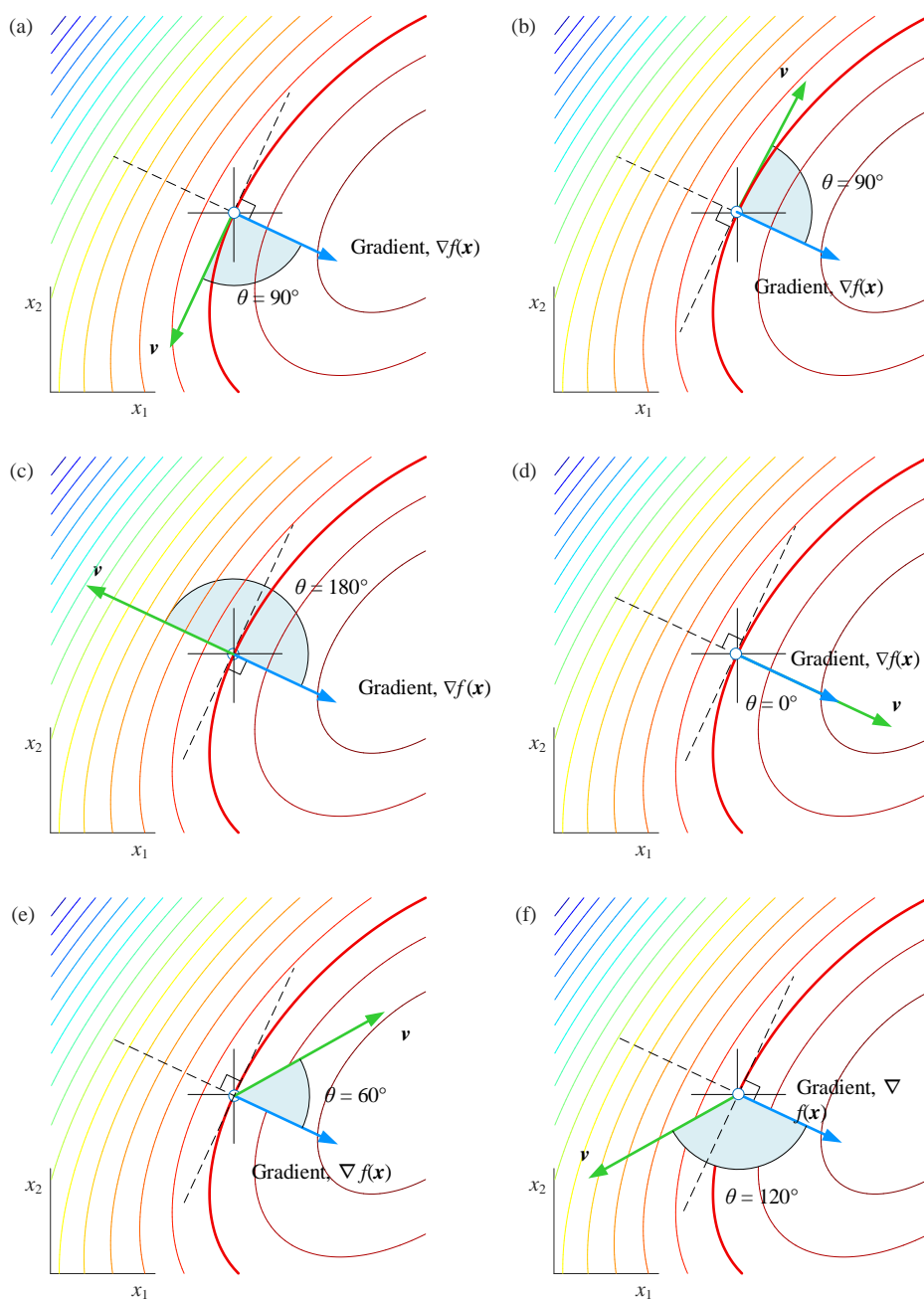
如图 15 (c)，若  $\theta = 180^\circ$ ，(42) 取得最小值。此时， $\mathbf{v}$  方向为梯度向量反方向，即下山方向。沿着  $\mathbf{v}$  运动瞬间，函数值减小最快。

如图 15 (d)， $\theta = 0^\circ$ ，(42) 取得最大值。方向导数和梯度向量同向，对应该点处函数值增大最快方向，即上山方向。

当  $\theta$  为锐角，(42) 大于 0。沿着  $\mathbf{v}$  运动瞬间，函数变化值大于 0，如图 15 (e)。当  $\theta$  为钝角，(42) 小于 0。沿着  $\mathbf{v}$  运动瞬间，函数变化值小于 0，如图 15 (f)。

特别地， $\mathbf{v} = [1, 0]^T$  对应  $f(x_1, x_2)$  对  $x_1$  偏导。 $\mathbf{v} = [0, 1]^T$  对应  $f(x_1, x_2)$  对  $x_2$  偏导。可见，方向性微分比偏导更灵活。

方向导数可以用来研究多元函数在某一特定方向的功能变化率，机器学习和深度学习很多算法在求解优化问题时都会用到方向导数这个重要的数学工具。

图 15.  $x_1x_2$  平面上六种方向导数情况

## 17.5 泰勒展开：一元到多元

丛书《数学要素》第 17 章介绍**泰勒展开** (Taylor series expansion)。本节将一元泰勒展开扩展到多元函数。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

## 一元函数泰勒展开

一元函数  $f(x)$  在展开点  $x = a$  处泰勒展开形式为：

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n \\ &= \underbrace{f(a)}_{\text{Constant}} + \underbrace{\frac{f'(a)}{1!}(x-a)}_{\text{Linear}} + \underbrace{\frac{f''(a)}{2!}(x-a)^2}_{\text{Quadratic}} + \underbrace{\frac{f'''(a)}{3!}(x-a)^3}_{\text{Cubic}} + \dots \end{aligned} \quad (43)$$

上式保留“常数 + 一阶导数”两个成分就是线性逼近：

$$f(x) \approx \underbrace{f(a)}_{\text{Constant}} + \underbrace{\frac{f'(a)}{1!}(x-a)}_{\text{Linear}} \quad (44)$$

我们在《数学要素》第 17 章中讲过，如图 16 所示，从几何角度，二元函数泰勒展开相当于，水平面、斜面、二次曲面、三次曲面等多项式曲面叠加。

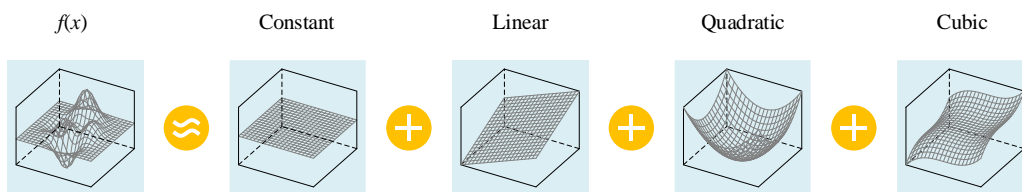


图 16. 二元函数泰勒展开原理，来自《数学要素》

## 线性逼近

更一般情况，对于多元函数  $f(\mathbf{x})$ ，当  $\mathbf{x}$  足够靠近展开点  $\mathbf{x}_p$  时， $f(\mathbf{x})$  函数值可以用泰勒一阶展开逼近，如下式：

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^T (\mathbf{x} - \mathbf{x}_p) \\ &= f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^T \Delta \mathbf{x} \end{aligned} \quad (45)$$

$\mathbf{x}_p$  为**泰勒级数展开点** (expansion point of Taylor series)， $\nabla f(\mathbf{x}_p)$  为多元函数  $f(\mathbf{x})$  在  $\mathbf{x}_p$  处梯度向量。

图 17 比较一元函数和二元函数线性逼近。一元线性逼近是用切线逼近曲线，二元线性逼近是用切面逼近曲面。

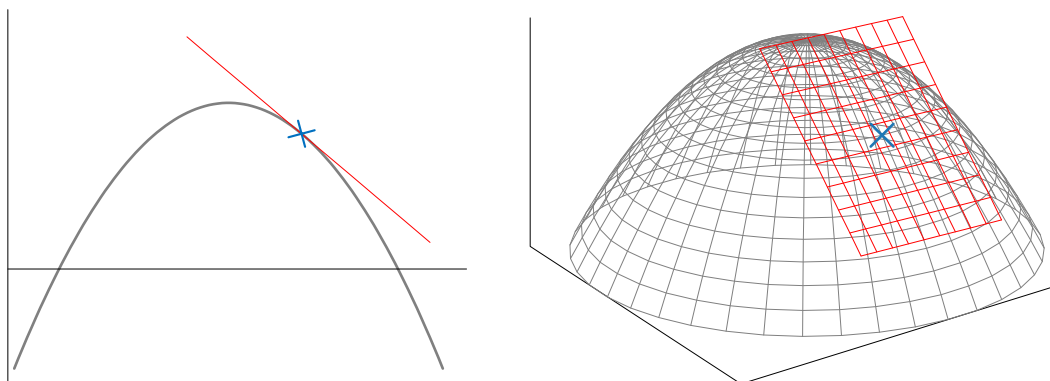


图 17. 一元到二元线性逼近

## 二次逼近

多元函数  $f(\mathbf{x})$  泰勒二阶级数展开式对应的矩阵运算如下：

$$\begin{aligned}
 f(\mathbf{x}) &\approx f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^T (\mathbf{x} - \mathbf{x}_p) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_p)^T \nabla^2 f(\mathbf{x}_p) (\mathbf{x} - \mathbf{x}_p) \\
 &= f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \nabla^2 f(\mathbf{x}_p) \Delta \mathbf{x} \\
 &= f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{H} \Delta \mathbf{x}
 \end{aligned} \tag{46}$$

上式就是二次逼近。其中， $\mathbf{H}$  为黑塞矩阵。

## 二次曲面

本章最后讨论二次曲面在某点切面，即一次逼近。采用圆锥曲线一般式，令  $y = f(x_1, x_2)$ ：

$$y = f(x_1, x_2) = Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F \tag{47}$$

$y = f(x_1, x_2)$  写成矩阵运算式：

$$y = f(x_1, x_2) = \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 2A & B \\ B & 2C \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} D \\ E \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + F \tag{48}$$

构造函数  $F(x_1, x_2, y)$ ：

$$F(x_1, x_2, y) = Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F - y \tag{49}$$

在三维空间中一点  $P(p_1, p_2, p_y)$ ， $F(x_1, x_2, y)$  曲面法向量  $\mathbf{n}_p$  通过下式得到：

$$\mathbf{n}_P = \left[ \begin{array}{c} \frac{\partial F}{\partial x_1} \\ \frac{\partial F}{\partial x_2} \\ \frac{\partial F}{\partial y} \end{array} \right]_{(p_1, p_2, p_y)} = \left[ \begin{array}{c} 2Ap_1 + Bp_2 + D \\ Bp_1 + 2Cp_2 + E \\ -1 \end{array} \right] \quad (50)$$

切面上任意一点  $(x_1, x_2, y)$  和切点  $P$  构成向量  $\mathbf{p}$ :

$$\mathbf{p} = \left[ \begin{array}{c} x_1 - p_1 \\ x_2 - p_2 \\ y - p_y \end{array} \right] \quad (51)$$

$\mathbf{p}$  垂直于  $\mathbf{n}_P$ , 因此两者向量内积为 0, 得到如下等式:

$$(2Ap_1 + Bp_2 + D)(x_1 - p_1) + (Bp_1 + 2Cp_2 + E)(x_2 - p_2) - y + p_y = 0 \quad (52)$$

整理得到切面解析式  $t(x_1, x_2)$ :

$$t(x_1, x_2) = (2Ap_1 + Bp_2 + D)(x_1 - p_1) + (Bp_1 + 2Cp_2 + E)(x_2 - p_2) + p_y \quad (53)$$

另外, 以上切面解析式就是  $P$  点泰勒一次逼近:

$$t(x_1, x_2) = f(p_1, p_2) + \nabla f(p_1, p_2)^T \left[ \begin{array}{c} x_1 - p_1 \\ x_2 - p_2 \end{array} \right] \quad (54)$$

$y = f(x_1, x_2)$  在  $P$  点梯度向量:

$$\nabla f(p_1, p_2) = \begin{bmatrix} 2A & B \\ B & 2C \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} + \begin{bmatrix} D \\ E \end{bmatrix} = \begin{bmatrix} 2Ap_1 + Bp_2 + D \\ Bp_1 + 2Cp_2 + E \end{bmatrix} \quad (55)$$

将 (55) 代入 (54), 同样可以得到 (53) 结果。

## 举个例子

给定二元函数  $y = f(x_1, x_2)$ ,

$$y = f(x_1, x_2) = -4x_1^2 - 4x_2^2 \quad (56)$$

将  $A$  点坐标  $(0, -1.5, -9)$  带入 (53), 得到曲面  $A$  点处切面解析式, 具体如下:

$$t(x_1, x_2) = 12x_2 + 9 \quad (57)$$

图 18 (a) 所示为二次曲面和曲面上  $A$  点  $(0, -1.5, -9)$  切面。图 18 (b) 所示为  $B$  点  $(-1.5, 0, -9)$  曲面切面。请大家自行计算曲面  $B$  点处切面解析式。

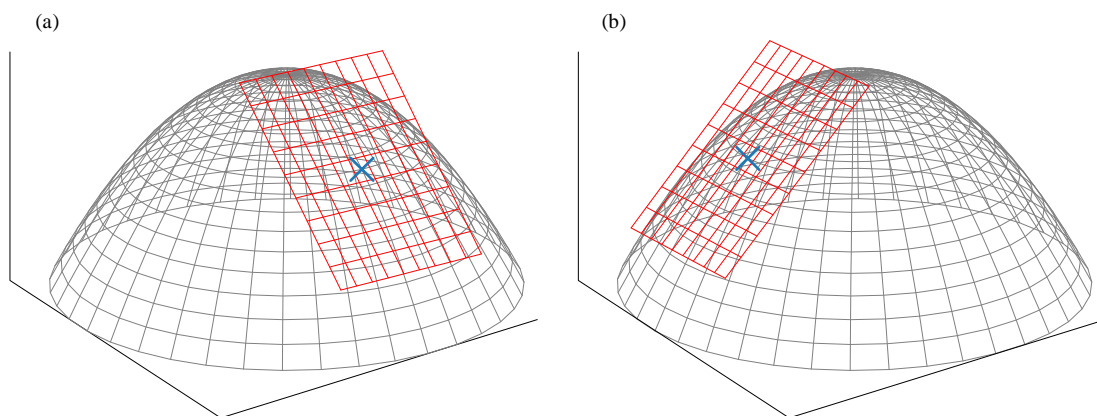


图 18. 二次凹曲面 A 点处切面



Bk4\_Ch17\_03.py 绘制图 18。



本章将一元函数导数和微分工具推广到多元函数，并介绍了几个重要数学工具——梯度向量、黑塞矩阵、法向量、方向导数、一次泰勒逼近、二次泰勒逼近。本书后续将利用这些数学工具分析解决各种数学问题。

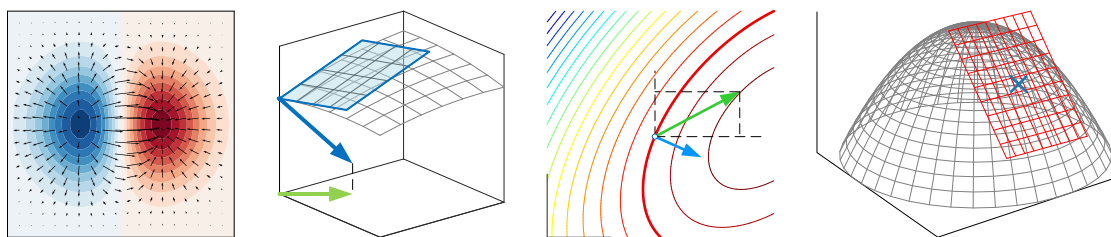


图 19. 总结本章重要内容的四副图



本章仅仅讨论了本书后续将会用到的矩阵微分法则。大家如果对这个话题感兴趣的话，推荐大家参考 *The Matrix Cookbook*。下载地址为：

<https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>