

## 21

## Surfaces and Positive Definiteness

## 曲面和正定性

代数、微积分、几何、线性代数的结合体



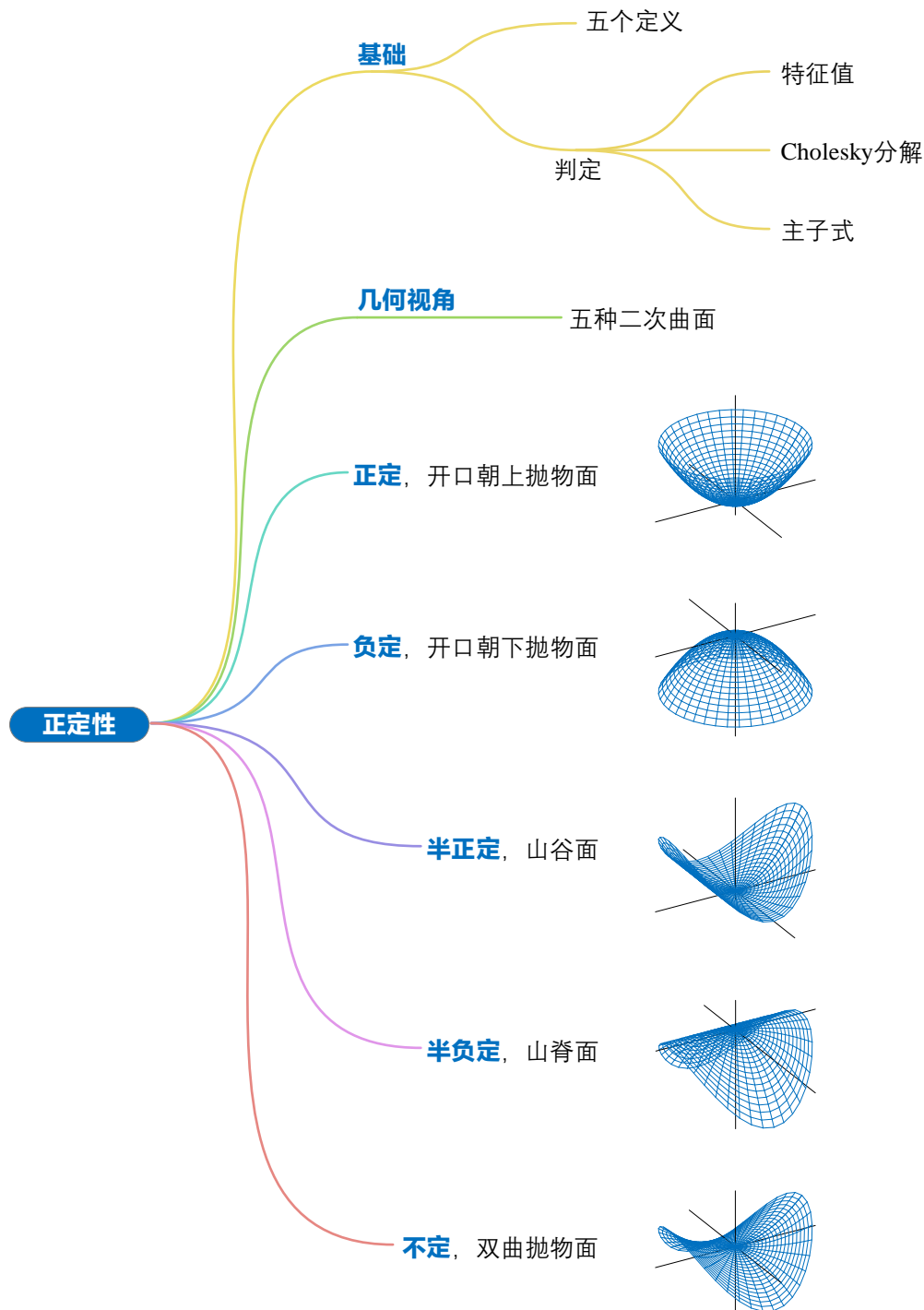
神，几何化一切。

*God ever geometrizes.*

—— 柏拉图 (Plato) | 古希腊哲学家 | 424/423 ~ 348/347 BC



- ◀ matplotlib.pyplot.contour() 绘制等高线图
- ◀ matplotlib.pyplot.contourf() 绘制填充等高线图
- ◀ matplotlib.pyplot.scatter() 绘制散点图
- ◀ numpy.arange() 在指定区间内返回均匀间隔数组
- ◀ numpy.array() 创建 array 数据类型
- ◀ numpy.cos() 余弦函数
- ◀ numpy.linalg.cholesky() Cholesky 分解函数
- ◀ numpy.linspace() 产生连续均匀间隔数组
- ◀ numpy.meshgrid() 生成网格化数据
- ◀ numpy.multiply() 向量或矩阵逐项乘积
- ◀ numpy.roots() 多项式求根
- ◀ numpy.sin() 正弦函数
- ◀ numpy.sqrt() 计算平方根
- ◀ 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)

## 21.1 正定性

**正定性** (positive definiteness) 是优化问题经常出现线性代数概念。本章结合**二次曲面** (quadratic surface)，和大家聊一聊正定性及其应用。

### 五个定义

矩阵正定性分为如下五种情况。

当  $\mathbf{x} \neq \mathbf{0}$  ( $\mathbf{x}$  为非零列向量) 时，如果满足：

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad (1)$$

矩阵  $\mathbf{A}$  为**正定矩阵** (positive definite matrix)。

当  $\mathbf{x} \neq \mathbf{0}$  时，

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad (2)$$

矩阵  $\mathbf{A}$  为**半正定矩阵** (positive semi-definite matrix)。

当  $\mathbf{x} \neq \mathbf{0}$  时，

$$\mathbf{x}^T \mathbf{A} \mathbf{x} < 0 \quad (3)$$

矩阵  $\mathbf{A}$  为**负定矩阵** (negative definite matrix)。

当  $\mathbf{x} \neq \mathbf{0}$  时，

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \leq 0 \quad (4)$$

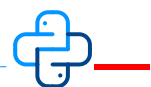
矩阵  $\mathbf{A}$  为**半负定矩阵** (negative semi-definite matrix)。

矩阵  $\mathbf{A}$  不属于以上任何一种情况， $\mathbf{A}$  为**不定矩阵** (indefinite matrix)。

### 判定正定矩阵

判断矩阵是否为正定矩阵，本书主要采用如下两种方法：

- ◀ 若矩阵为对称矩阵，并且所有特征值为正，则矩阵为正定矩阵；
- ◀ 若矩阵可以进行 Cholesky 分解，则矩阵为正定矩阵。



Bk4\_Ch21\_01.py 介绍如何使用 Cholesky 分解判定矩阵是否为正定矩阵。

### Cholesky 分解

如果矩阵  $\mathbf{A}$  为正定矩阵，对  $\mathbf{A}$  进行 Cholesky 分解，得到：

$$\mathbf{A} = \mathbf{R}^T \mathbf{R} \quad (5)$$

利用 (5)，将  $\mathbf{x}^T \mathbf{A} \mathbf{x}$  写成如下形式：

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T \mathbf{R}^T \mathbf{R} \mathbf{x} = (\mathbf{R} \mathbf{x})^T \mathbf{R} \mathbf{x} = \|\mathbf{R} \mathbf{x}\|^2 \quad (6)$$

$\mathbf{R}$  中列向量线性无关，若  $\mathbf{x}$  为非零向量，则  $\mathbf{R} \mathbf{x} \neq \mathbf{0}$ ，因此  $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ 。

### 特征值分解

对称矩阵  $\mathbf{A}$  进行特征值分解得到：

$$\mathbf{A} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \quad (7)$$

将 (7) 代入  $\mathbf{x}^T \mathbf{A} \mathbf{x}$ ，得到：

$$\begin{aligned} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \mathbf{x}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{x} \\ &= \left( \mathbf{V}^T \mathbf{x} \right)^T \mathbf{\Lambda} \left( \mathbf{V}^T \mathbf{x} \right) \end{aligned} \quad (8)$$

令：

$$\mathbf{z} = \mathbf{V}^T \mathbf{x} \quad (9)$$

(8) 可以写成：

$$\begin{aligned} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \mathbf{z}^T \mathbf{\Lambda} \mathbf{z} \\ &= \lambda_1 z_1^2 + \lambda_2 z_2^2 + \cdots + \lambda_D z_D^2 = \sum_{j=1}^D \lambda_j z_j^2 \end{aligned} \quad (10)$$

当上式中特征值均为正数，除非  $z_1, z_2, \dots, z_D$  均为 0 (即  $\mathbf{z}$  为零向量)，否则上式大于 0。

若  $\mathbf{A}$  的特征值均为负值，则矩阵  $\mathbf{A}$  为负定矩阵。若矩阵  $\mathbf{A}$  特征值为正值或 0， $\mathbf{A}$  为半正定矩阵。若矩阵特征值为负值或 0，则矩阵  $\mathbf{A}$  为半负定矩阵。

### 格拉姆矩阵

给定数据矩阵  $\mathbf{X}$ ，它的格拉姆矩阵为  $\mathbf{G} = \mathbf{X}^T \mathbf{X}$ 。格拉姆矩阵至少都是半正定矩阵。

将  $\mathbf{x}^T \mathbf{G} \mathbf{x}$  写成如下形式：

$$\mathbf{x}^T \mathbf{G} \mathbf{x} = \mathbf{x}^T \mathbf{X}^T \mathbf{X} \mathbf{x} = \|\mathbf{X} \mathbf{x}\|^2 \geq 0 \quad (11)$$

特别地，当  $\mathbf{X}$  满秩时， $\mathbf{x}$  为非零向量，则  $\mathbf{X} \mathbf{x} \neq \mathbf{0}$ ，因此  $\mathbf{x}^T \mathbf{G} \mathbf{x} > 0$ 。也就是说，当  $\mathbf{X}$  满秩，格拉姆矩阵  $\mathbf{G} = \mathbf{X}^T \mathbf{X}$  为正定矩阵。

这一节介绍了正定性相关性质，但是想要直观理解这个概念，还需要借助几何视角。

## 21.2 几何视角看正定性

给定如下  $2 \times 2$  对称矩阵  $\mathbf{A}$ ：

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \quad (12)$$

构造如下二元函数  $y = f(x_1, x_2)$ ：

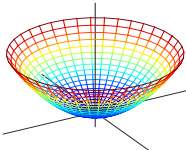
$$\begin{aligned} y = f(x_1, x_2) &= \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= ax_1^2 + 2bx_1x_2 + cx_2^2 \end{aligned} \quad (13)$$

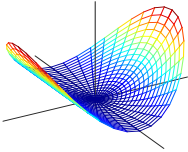
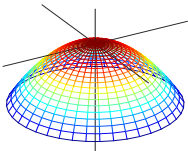
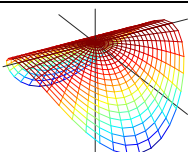
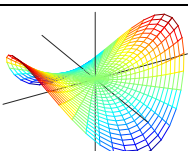
在三维正交空间中，当矩阵  $\mathbf{A}_{2 \times 2}$  正定性不同时， $y = f(x_1, x_2)$  对应曲面展现出不同的形状：

- ◀ 当  $\mathbf{A}_{2 \times 2}$  为正定矩阵时， $y = f(x_1, x_2)$  为开口向上抛物面；
- ◀ 当  $\mathbf{A}_{2 \times 2}$  为半正定矩阵时， $y = f(x_1, x_2)$  为山谷面；
- ◀ 当  $\mathbf{A}_{2 \times 2}$  为负定矩阵时， $y = f(x_1, x_2)$  为开口向下抛物面；
- ◀ 当  $\mathbf{A}_{2 \times 2}$  为半负定矩阵时， $y = f(x_1, x_2)$  为山脊面；
- ◀ 当  $\mathbf{A}_{2 \times 2}$  不定时， $y = f(x_1, x_2)$  为马鞍面，也叫做双曲抛物面。

表 1 总结了矩阵  $\mathbf{A}$  不同正定性条件下对应的曲面形状。本章以下六节就按表中形状顺序展开。

表 1. 正定性的几何意义

$\mathbf{A}_{D \times D}$	特征值	形状
$\mathbf{A}_{D \times D}$ 为正定矩阵 $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0, \mathbf{x} \neq \mathbf{0}$	$D$ 个特征值均为正值	

$A_{D \times D}$ 为半正定矩阵, 秩为 $r$ $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0, \mathbf{x} \neq \mathbf{0}$	$r$ 个正特征值, $D-r$ 个特征值为 0	
$A_{D \times D}$ 为负定矩阵 $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$	$D$ 个特征值均为负值	
$A_{D \times D}$ 为半负定矩阵, 秩为 $r$ $\mathbf{x}^T \mathbf{A} \mathbf{x} \leq 0$	$r$ 个负特征值, $D-r$ 个特征值为 0	
$A_{D \times D}$ 为不定矩阵	特征值符号正负不定	

## 21.3 开口朝上抛物面：正定

### 正圆

先来看一个单位矩阵的例子。若矩阵  $A$  为  $2 \times 2$  单位矩阵：

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (14)$$

单位矩阵显然是正定矩阵。构造如下二元函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y &= f(x_1, x_2) = \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 + x_2^2 \end{aligned} \quad (15)$$

观察上式，容易发现只有当  $x_1 = 0$  且  $x_2 = 0$  时，即  $\mathbf{x} = \mathbf{0}$ ， $y = f(x_1, x_2) = 0$ 。

容易求得  $A$  特征值分别为  $\lambda_1 = 1$  和  $\lambda_2 = 1$ ，对应特征向量分别为：

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (16)$$

计算矩阵  $A$  的秩,  $\text{rank}(A) = 2$ 。

图 1 (a) 所示为  $y = f(x_1, x_2)$  曲面。在该曲面边缘  $A$ 、 $B$  和  $C$  放置小球, 小球都会朝着曲面最低点滚动。

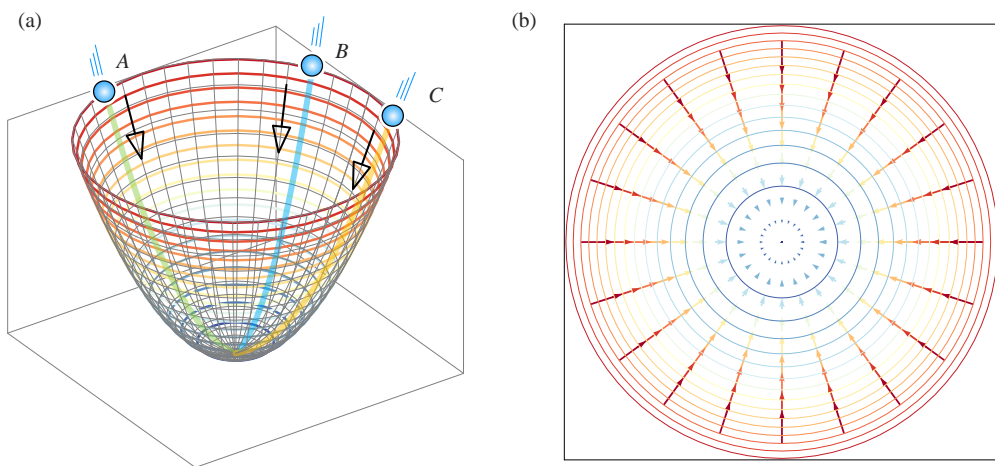


图 1. 正定矩阵曲面和梯度下降, 正圆抛物面

(15) 的梯度向量为:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \quad (17)$$

而 (15) 的梯度下降向量就是上式中梯度向量反向:

$$-\nabla f(\mathbf{x}) = \begin{bmatrix} -2x_1 \\ -2x_2 \end{bmatrix} \quad (18)$$

图 1 (b) 展示  $f(x_1, x_2)$  平面等高线为正圆。图 1 (b) 还给出不同位置的梯度下降向量, 即指向下山方向, 梯度向量的反方向。在本章中, 除了最后一节外, 平面等高线中的向量场都是梯度下降向量。

如图 1 (b) 所示, 梯度下降向量均指向最小值点。此外, 梯度下降向量方向垂直所在等高线。梯度下降向量的长度代表坡度的陡峭程度。向量长度越大, 坡度越陡, 该方向上函数值变化率越大。当梯度下降向量的长度为 0 时, 对应驻点。

梯度下降向量为零向量  $\mathbf{0}$  的点, 就是  $y = f(x_1, x_2)$  两个偏导均为 0 的点。本系列丛书《数学要素》介绍过,  $(0, 0)$  这个点被称作驻点。通过图 1, 很容易判断  $(0, 0)$  就是二元函数最小值点。

**▲** 再次强调, 图 1 给出的是梯度下降向量 (下山方向), 方向和梯度向量 (上山方向) 正好相反。沿着梯度下降向量方向移动, 函数值减小; 沿着梯度向量方向移动, 函数值增大。

## 正椭圆

再看一个  $2 \times 2$  正定矩阵例子。矩阵  $\mathbf{A}$  具体值如下：

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad (19)$$

同样，构造二元函数  $y = f(x_1, x_2)$ ，具体如下：

$$\begin{aligned} y &= f(x_1, x_2) = \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 + 2x_2^2 \end{aligned} \quad (20)$$

同样，只有  $x_1 = 0$  且  $x_2 = 0$  时， $y = f(x_1, x_2) = 0$ 。图 2 所示为 (20) 对应开口向上正椭圆抛物面，函数等高线为一系列正椭圆。

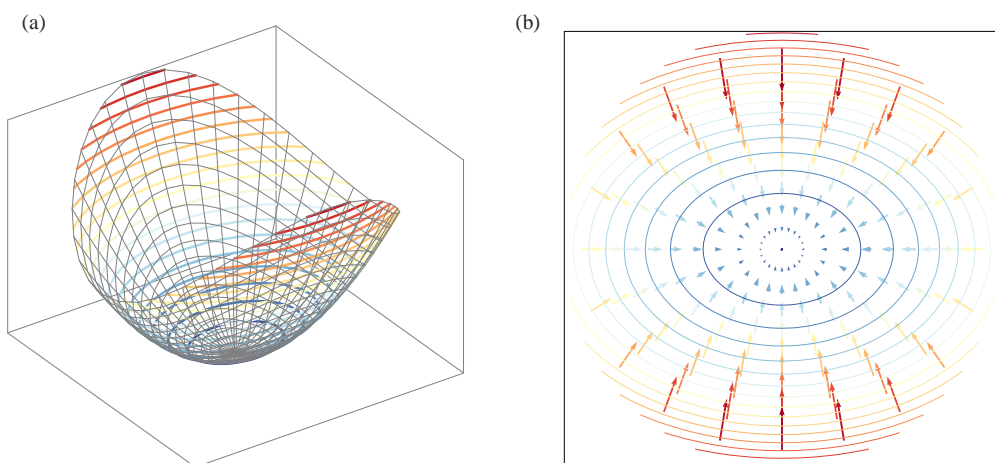


图 2. 正定矩阵曲面和梯度下降，正椭圆抛物面

容易求得  $\mathbf{A}$  特征值分别为  $\lambda_1 = 1$  和  $\lambda_2 = 2$ ，对应特征向量分别为：

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (21)$$

(15) 的梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 4x_2 \end{bmatrix} \quad (22)$$

梯度向量为  $\mathbf{0}$  的点  $(0, 0)$  是 (20) 函数的最小值点。



## 旋转椭圆

本节前两个例子对应的曲面的等高线分别是正圆和正椭圆，下面再看一个旋转椭圆情况。 $A$  矩阵具体如下：

$$A = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \quad (23)$$

构造函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y = f(x_1, x_2) &= \mathbf{x}^T A \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 1.5x_1^2 + x_1x_2 + 1.5x_2^2 \end{aligned} \quad (24)$$

同样，只有当  $x_1 = 0$  且  $x_2 = 0$  时， $y = f(x_1, x_2) = 0$ 。

经过计算得到  $A$  特征值也是  $\lambda_1 = 1$  和  $\lambda_2 = 2$ ；这两个特征值对应特征向量分别为：

$$\mathbf{v}_1 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \quad (25)$$

(24) 梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 3x_1 + x_2 \\ x_1 + 3x_2 \end{bmatrix} \quad (26)$$

$y = f(x_1, x_2)$  曲面对应图像如图 3。图 2 和图 3 两个椭圆唯一的差别就是旋转角度。根据前文所学，我们知道这两组椭圆的半长轴和半短轴的比例关系为  $\sqrt{\lambda_2}/\sqrt{\lambda_1}$ ，即  $\sqrt{2}/1$ 。

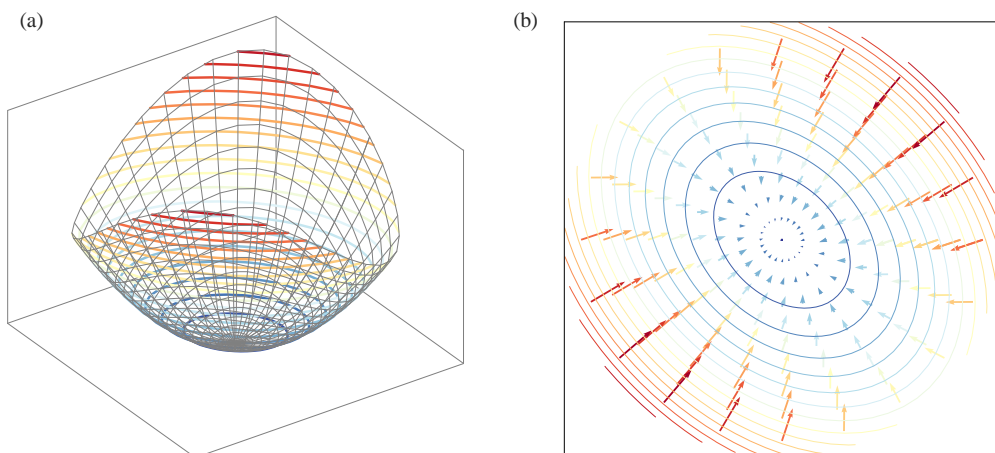


图 3. 正定矩阵曲面和梯度下降，开口向上旋转椭圆抛物面



Bk4\_Ch21\_02.py 绘制图 1、图 2、图 3，此外请大家修改代码并绘制本章其他图像。

## 21.4 山谷面：半正定

下面来聊一聊半正定矩阵情况。举个例子，矩阵  $A$  取值如下：

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad (27)$$

容易判定  $\text{rank}(A) = 1$ 。构造如下二元函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y &= f(x_1, x_2) = \mathbf{x}^T A \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 \end{aligned} \quad (28)$$

$x_1 = 0$  时，不管  $x_2$  取任何值，上式为 0。

图 4 展示  $y = f(x_1, x_2)$  对应曲面。观察该图容易发现，除了纵轴以外任意点处放置一个小球，小球都会滚动到谷底。

(28) 梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 0 \end{bmatrix} \quad (29)$$

谷底位置对应一条直线，这条直线上每一点处梯度向量均为  $\mathbf{0}$ ，它们都是函数  $y = f(x_1, x_2)$  极小值。

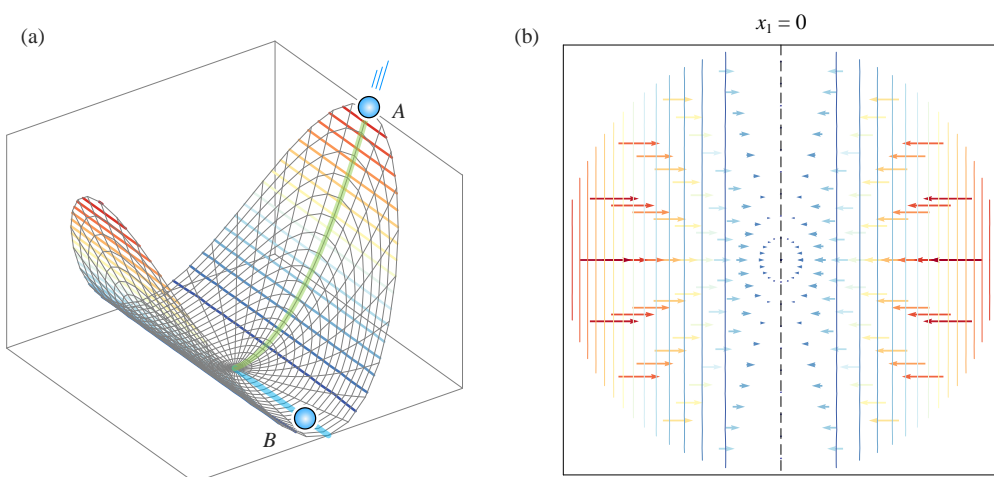


图 4. 半正定矩阵对应曲面

## 旋转山谷面

下式中矩阵  $\mathbf{A}$  也是半正定矩阵：

$$\mathbf{A} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix} \quad (30)$$

构造函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y = f(x_1, x_2) &= \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 0.5x_1^2 - x_1x_2 + 0.5x_2^2 \end{aligned} \quad (31)$$

(31) 配方得到：

$$f(x_1, x_2) = 0.5x_1^2 - x_1x_2 + 0.5x_2^2 = \frac{1}{2}(x_1 - x_2)^2 \quad (32)$$

容易发现，任何满足  $x_1 = x_2$  的点，都会使得  $y = f(x_1, x_2)$  为 0。

(31) 中矩阵  $\mathbf{A}$  特征值为  $\lambda_1 = 0$  和  $\lambda_2 = 1$ ，对应特征向量如下：

$$\mathbf{v}_1 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \quad (33)$$

图 5 展示 (31) 对应的旋转山谷面。同样，小球沿图 5 中  $\mathbf{v}_1$  (特征值为 0 对应特征向量) 方向运动，函数值没有任何变化。这条直线上的点都是 (32) 二元函数极小值点。

(32) 梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} x_1 - x_2 \\ -x_1 + x_2 \end{bmatrix} \quad (34)$$

观察图 5 (b)，容易发现梯度下降向量长度各有不同，但是它们相互平行，且都垂直于等高线，指向函数减小方向，即下山方向。

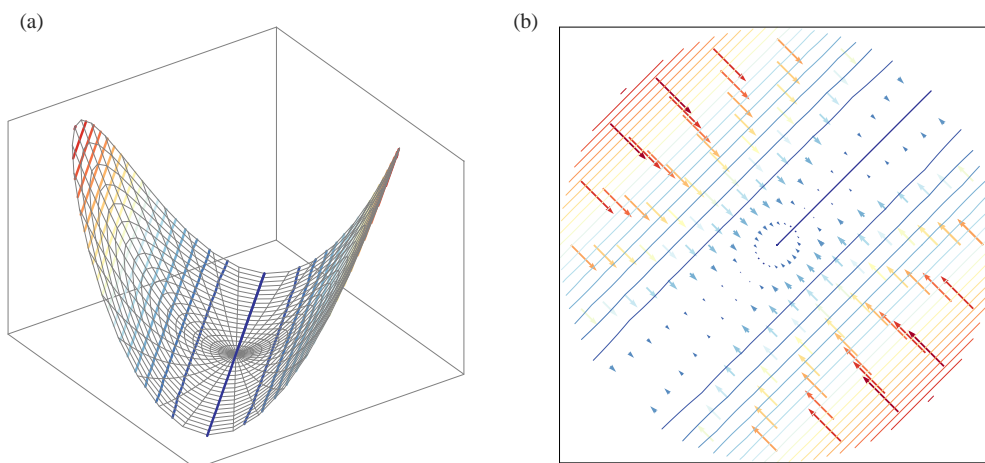


图 5. 旋转山谷面

## 21.5 开口朝下抛物面：负定

最简单的负定矩阵是单位矩阵取负，即  $-\mathbf{I}$ 。 $-\mathbf{I}$  的特征值都为  $-1$ 。

下面也用  $2 \times 2$  矩阵讨论负定。如下  $\mathbf{A}$  为负定矩阵：

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \quad (35)$$

构造如下二元函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned}
 y &= f(x_1, x_2) = \mathbf{x}^T \mathbf{A} \mathbf{x} \\
 &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
 &= -x_1^2 - 2x_2^2
 \end{aligned} \tag{36}$$

观察上式，容易发现只有当  $x_1 = 0$  且  $x_2 = 0$  时， $y = f(x_1, x_2) = 0$ 。

很容易求得  $\mathbf{A}$  特征值分别为  $\lambda_1 = -2$  和  $\lambda_2 = -1$ ，对应特征向量分别为：

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{37}$$

图 6 展示负定矩阵对应曲面，容易发现  $y = f(x_1, x_2)$  对应曲面为凹面。在曲面最大值处放置一个小球，小球处于不稳定平衡状态。受到轻微扰动后，小球沿着任意方向运动，都会下落。

(36) 中  $y = f(x_1, x_2)$  梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2x_1 \\ -4x_2 \end{bmatrix} \tag{38}$$

如图 6 所示，梯度下降向量指向均背离最大值点。

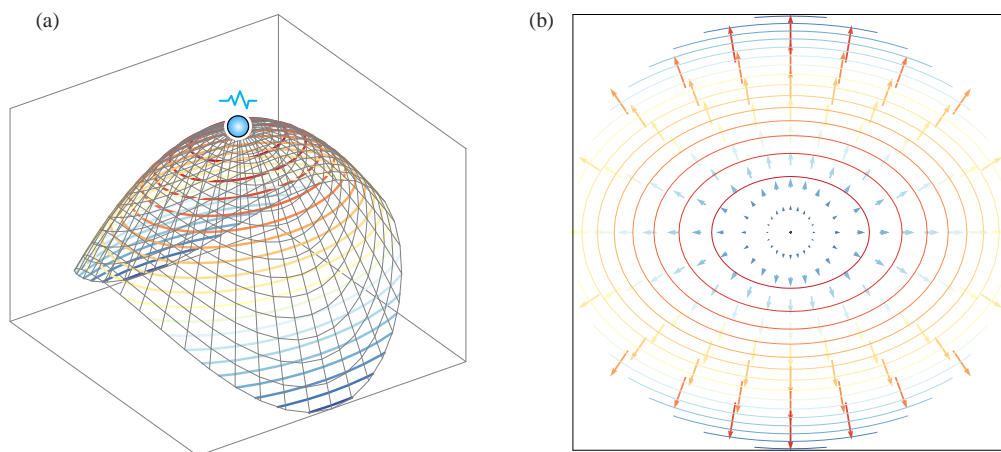


图 6. 负定矩阵对应曲面

## 21.6 山脊面：半负定

下面看一个半负定矩阵例子，矩阵  $\mathbf{A}$  取值如下：

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \quad (39)$$

构造  $y = f(x_1, x_2)$ :

$$\begin{aligned} y = f(x_1, x_2) &= \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= -x_2^2 \end{aligned} \quad (40)$$

$x_2 = 0$ ,  $x_1$  为任意值, 上式为 0。矩阵  $\mathbf{A}$  的秩为 1,  $\text{rank}(\mathbf{A}) = 1$ 。

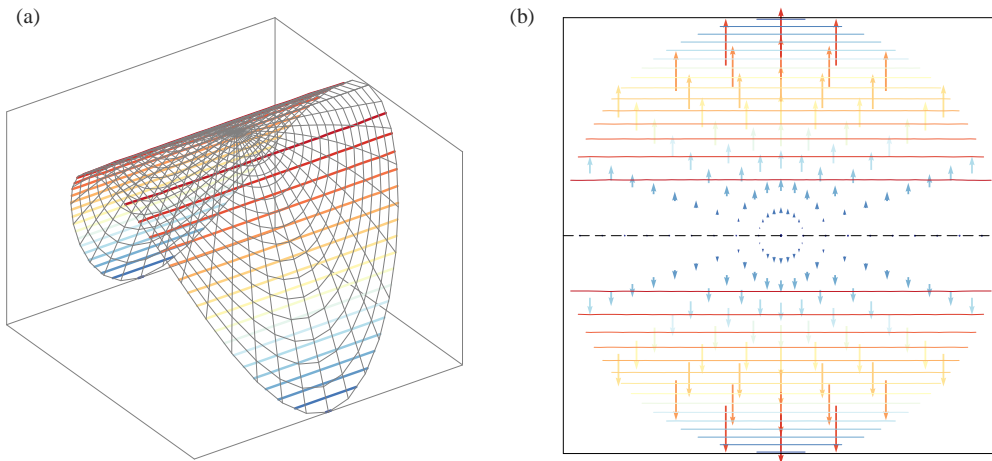


图 7. 半负定矩阵对应山脊面

图 7 展示半负定矩阵对应山脊面, 发现曲面有无数个极大值。在任意极大值 (山脊) 处放置一个小球, 受到扰动后, 小球会沿着曲面滚下。然而, 沿着山脊方向运动, 函数值没有任何变化。

(40) 梯度向量为:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ -2x_2 \end{bmatrix} \quad (41)$$

图 7 (b) 中梯度下降方向平行于纵轴, 指向函数值减小方向。

## 21.7 双曲抛物面：不定

本节最后聊一下不定矩阵情况。举个例子， $A$  为：

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (42)$$

构造函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y = f(x_1, x_2) &= \mathbf{x}^T A \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 - x_2^2 \end{aligned} \quad (43)$$

求得矩阵  $A$  对应特征值为  $\lambda_1 = -1$  和  $\lambda_2 = 1$ ，对应特征向量如下：

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (44)$$

图 8 展示  $y = f(x_1, x_2)$  对应曲面。

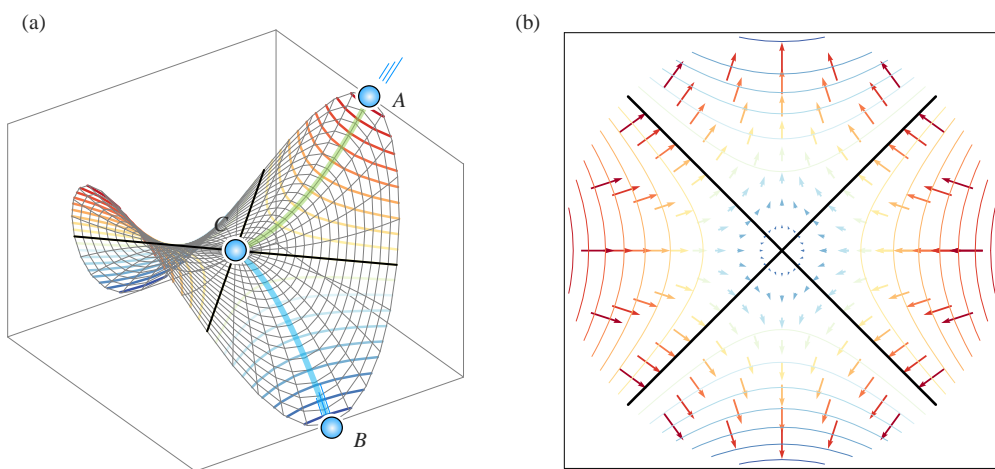


图 8. 不定矩阵对应曲面，马鞍面

当  $y \neq 0$ ，曲面对应等高线为双曲线。当  $y = 0$ ，曲面对应等高线是两条在  $x_1x_2$  平面内直线 (图 8 (a) 中黑色直线)，它们是双曲线渐近线。

图 8 告诉我们，曲面边缘不同位置放置小球会有完全不同运动方向。A 点处松手小球会向向着中心方向滚动，B 点处小球会朝远离中心方向滚动。

$y = f(x_1, x_2)$  梯度向量为：

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix} \quad (45)$$

图 8 所示马鞍面中心  $C$  既不是极小值点，也不是极大值点；图 8 中马鞍面中心点被称之为**鞍点** (saddle point)。另外，沿着图 8 中黑色轨道运动，小球高度没有任何变化。

### 旋转双曲抛物面

图 8 中马鞍面顺时针旋转  $45^\circ$  得到图 9 曲面。图 9 曲面对应矩阵  $A$  如下：

$$A = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \quad (46)$$

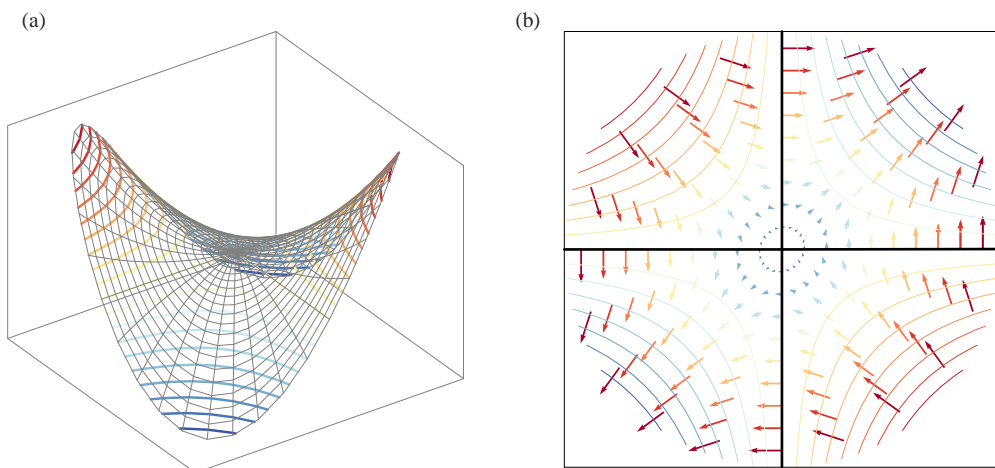


图 9. 不定矩阵对应曲面，旋转马鞍面

构造如下二元函数  $y = f(x_1, x_2)$ ：

$$\begin{aligned} y &= f(x_1, x_2) = \mathbf{x}^T A \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= -2x_1x_2 \end{aligned} \quad (47)$$

在  $y = f(x_1, x_2)$  为非零定值时，上式相当于反比例函数。

(47) 的梯度向量为：



$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2x_2 \\ -2x_1 \end{bmatrix} \quad (48)$$

请大家自行分析图 8 两幅图。



在 `Bk4_Ch21_02.py` 基础上，我们用 Streamlit 和 Plotly 制作了一个 App，可以调节参数  $a$ 、 $b$ 、 $c$  观察图像变化。App 还显示矩阵的特征值分解结果。请参考 `Streamlit_Bk4_Ch21_02.py`。

## 21.8 多极值曲面：局部正定性

### 判定二元函数极值点

本系列丛书在《数学要素》一册介绍过如何判定二元函数  $y = f(x_1, x_2)$  的极值。对于  $y = f(x_1, x_2)$ ，一阶偏导数  $f_{x_1}(x_1, x_2) = 0$  和  $f_{x_2}(x_1, x_2) = 0$  同时成立的点  $(x_1, x_2)$  为二元函数  $f(x_1, x_2)$  的驻点。如图 10 所示，驻点可以是极大值、极小值或鞍点。

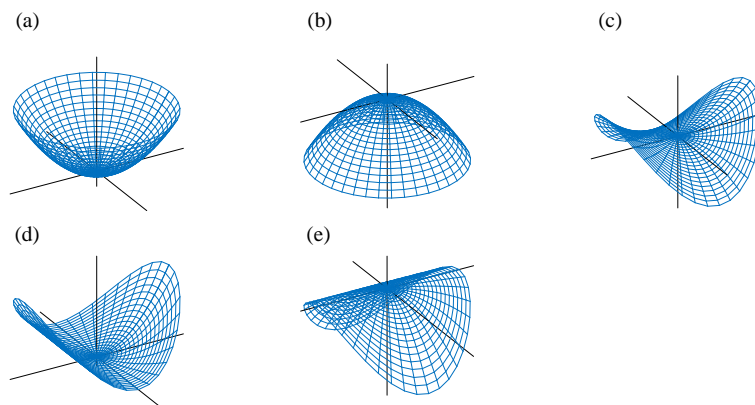


图 10. 二元函数驻点的三种情况

当时，我们聊过为了进一步判定驻点到底是极大值、极小值或是鞍点，需要知道二元函数  $f(x_1, x_2)$  二阶偏导。如果  $f(x_1, x_2)$  在  $(a, b)$  邻域内连续，且  $f(x_1, x_2)$  二阶偏导连续。令，

$$A = f_{x_1 x_1}, \quad B = f_{x_1 x_2}, \quad C = f_{x_2 x_2} \quad (49)$$

$f(a, b)$  是否为极值点可以通过如下条件判断：

- a)  $AC - B^2 > 0$  存在极值, 且当  $A < 0$  有极大值,  $A > 0$  时有极小值;
- b)  $AC - B^2 < 0$  没有极值;
- c)  $AC - B^2 = 0$ , 可能有极值, 也可能没有极值, 需要进一步讨论。

当时我们留了一个问题,  $AC - B^2$  这个表达值的含义到底是什么? 本节就来回答这个问题。

(13) 中函数的黑塞矩阵 (Hessian matrix) 为:

$$\mathbf{H} = \frac{\partial^2 (\mathbf{x}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{A} = 2 \begin{bmatrix} a & b \\ b & c \end{bmatrix} \quad (50)$$

注意上式中  $\mathbf{A}$  为对称矩阵。

$\mathbf{A}$  的行列式值为:

$$|\mathbf{A}| = ac - b^2 \quad (51)$$

相信大家已经在上式中看到和  $AC - B^2$  一样的形式。

对于二元函数,  $\mathbf{A}$  的形状为  $2 \times 2$ 。  $\mathbf{A}$  为正定或负定时,  $\mathbf{A}$  的两个特征值同号, 因此  $\mathbf{A}$  的行列式值都大于 0。而  $a$  的正负则决定了开口方向, 也就是决定了  $\mathbf{A}$  是正定还是负定, 因此决定了极大值或极小值。

再进一步,  $a$  实际上是  $\mathbf{A}$  的一阶主子式, 即矩阵  $\mathbf{A}$  的第一行、第一列元素构成矩阵的行列式值。这实际上引出了判断正定的第三个方法—— $\mathbf{A}$  正定的充分必要条件为  $\mathbf{A}$  的顺序主子式全大于零。

## 举个例子

继续采用《数学要素》一书中反复出现的多极值曲面的例子。

图 11 为曲面平面等高线。图中, 深绿色线代表  $f_{x1}(x_1, x_2) = 0$ , 深蓝色线代表  $f_{x2}(x_1, x_2) = 0$ 。两个颜色线交点标记为  $\times$ 。也就是说, 图中  $\times$  对应的位置为梯度向量为  $\mathbf{0}$ 。

观察图中等高线不难发现, I、II、III 点为极大值点, 其中 I 为最大值点。IV、V、VI 为极小值点, 其中 IV 为最小值点。VII、VIII、IX 是鞍点。

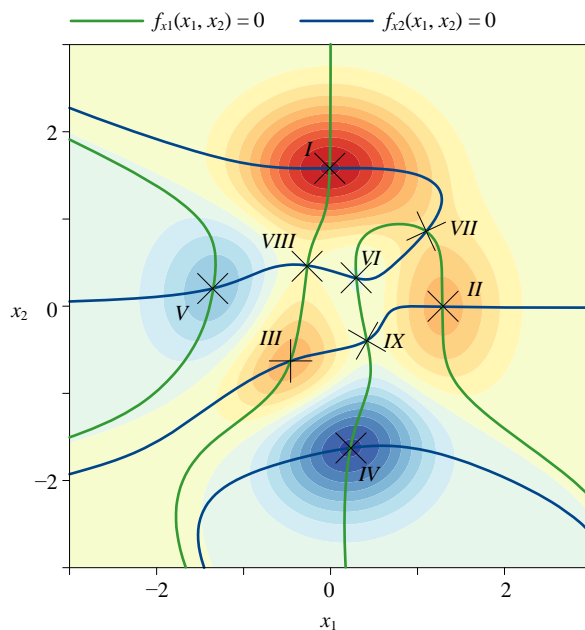


图 11.  $f_{x1}(x_1, x_2) = 0$  和  $f_{x2}(x_1, x_2) = 0$  同时投影在  $f(x_1, x_2)$  曲面填充等高线，来自本系列丛书《数学要素》

图 12 给出的是二元函数的梯度向量图 (和梯度下降向量方向相反)。极大值点处，梯度向量 (上山方向) 汇聚；极小值点处，梯度向量发散。这一点很好理解，在极大值点附近，朝着极大值走就是上山；相反，在极小值点附近，背离极小值走则对应上山，朝着极小值走则是下山。

而鞍点处，有些梯度向量指向鞍点，有些梯度向量背离鞍点。也就是说，鞍点处，既可以下山，也可以上山。

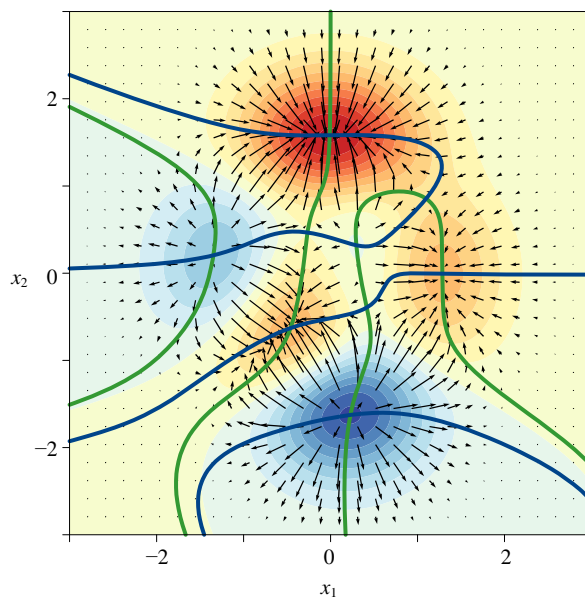


图 12.  $f(x_1, x_2)$  梯度向量图，梯度向量对应“上山”，和梯度下降 (下山) 反向

图 13 所示为二次函数黑塞矩阵行列式值对应的等高线图，阴影圈出来的六个点对应行列式值为正，因此它们是要考察的极值点。图 13 中虚线为行列式值为 0 对应位置。

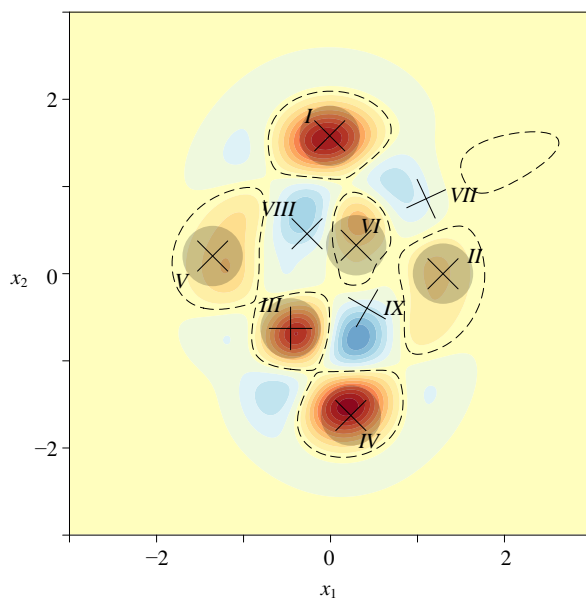


图 13. 黑塞矩阵行列式值

根据图 14 所示一阶主子式对应等高线。通过一阶主子式值的正负，即  $f_{x_1x_1}$  正负，可以进一步判定极值点为极大值或极小值点，最终得出的结论和图 11 一致。

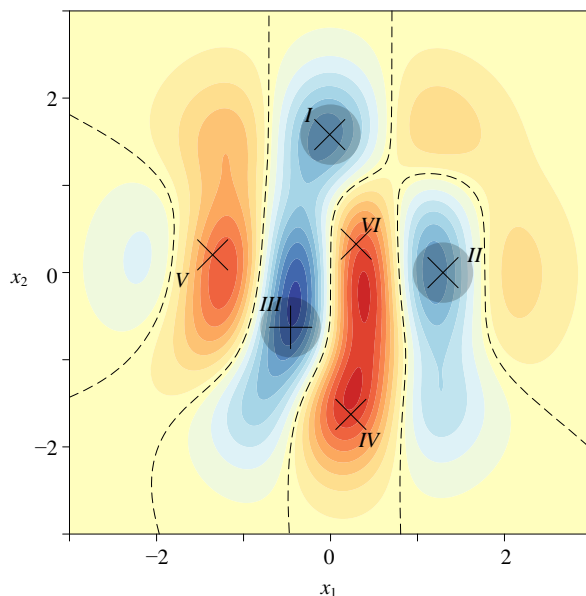


图 14. 一阶主子式正负

## 更一般情况

对于多元函数  $f(\mathbf{x})$ ，利用本书第 17 章介绍的二次逼近  $f(\mathbf{x})$  可以写成：

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^\top (\mathbf{x} - \mathbf{x}_p) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_p)^\top \nabla^2 f(\mathbf{x}_p) (\mathbf{x} - \mathbf{x}_p) \\ &= f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^\top \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^\top \nabla^2 f(\mathbf{x}_p) \Delta \mathbf{x} \\ &= f(\mathbf{x}_p) + \nabla f(\mathbf{x}_p)^\top \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^\top \mathbf{H} \Delta \mathbf{x} \end{aligned} \quad (52)$$

其中  $\mathbf{x}_p$  为展开点。

假设  $\mathbf{x}_p$  处存在梯度向量，且梯度向量为  $\mathbf{0}$ 。

当  $\mathbf{x} \rightarrow \mathbf{x}_p$  时， $\nabla f(\mathbf{x}_p)^\top \Delta \mathbf{x} \rightarrow 0$ 。但是如果在  $\mathbf{x}_p$  点处黑塞矩阵  $\mathbf{H}$  为正定， $\frac{1}{2} \Delta \mathbf{x}^\top \mathbf{H} \Delta \mathbf{x}$  为正。这意味着：

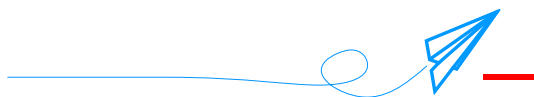
$$f(\mathbf{x}_p) + \underbrace{\nabla f(\mathbf{x}_p)^\top \Delta \mathbf{x}}_{\rightarrow 0} + \underbrace{\frac{1}{2} \Delta \mathbf{x}^\top \mathbf{H} \Delta \mathbf{x}}_{+} > f(\mathbf{x}_p) \quad (53)$$

这种情况称  $\mathbf{x}_p$  局部正定，对应  $\mathbf{x}_p$  为极小值点。这个判断也适用于半正定情况，不过要将上式的  $>$  改为  $\geq$ 。

同理，如果在  $\mathbf{x}_p$  点处黑塞矩阵  $\mathbf{H}$  为负定， $\frac{1}{2} \Delta \mathbf{x}^\top \mathbf{H} \Delta \mathbf{x}$  为负，因此：

$$f(\mathbf{x}) = f(\mathbf{x}_p) + \underbrace{\nabla f(\mathbf{x}_p)^\top \Delta \mathbf{x}}_{\rightarrow 0} + \underbrace{\frac{1}{2} \Delta \mathbf{x}^\top \mathbf{H} \Delta \mathbf{x}}_{-} < f(\mathbf{x}_p) \quad (54)$$

我们称  $\mathbf{x}_p$  局部负定，对应  $\mathbf{x}_p$  为极大值点。如上判断也适用于半负定情况，同样将上式的  $<$  改为  $\leq$ 。



本章把曲面、梯度向量、正定性、极值这几个重要的概念有机的联系起来。本章给出的各种例子告诉我们几何视角是学习线性代数的捷径。

请大家再次回顾图 15 给出的五种情况，并且将正定性、极值（最值）对号入座。相信大家学完本章之后，会觉得正定性变得极容易理解。

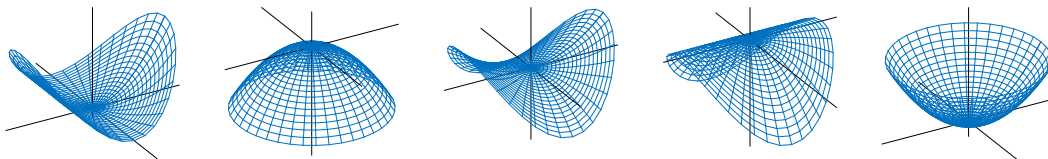


图 15. 总结本章重要内容的五副图