

# 18

## Lagrange Multiplier

# 拉格朗日乘法

把有约束优化问题转化为无约束优化问题



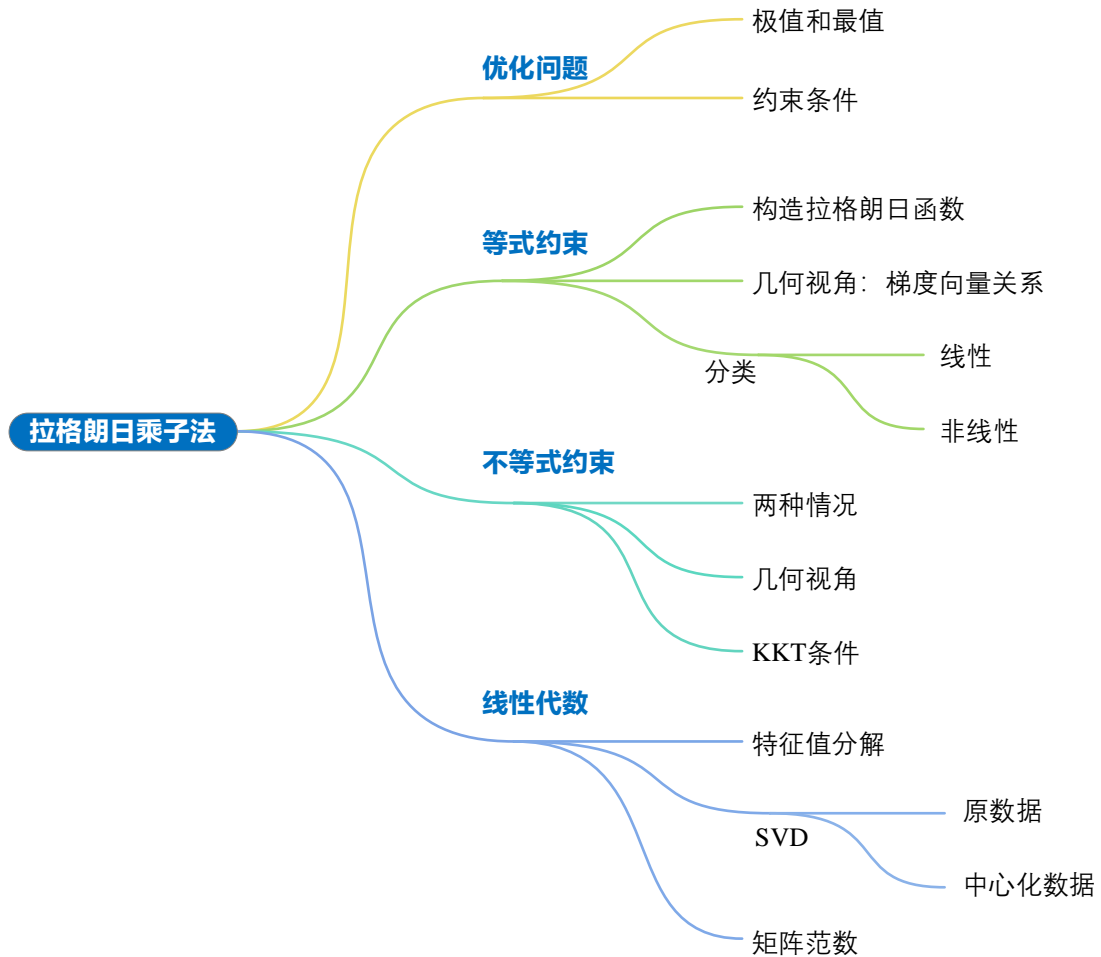
伟大的事情是由一系列小事情聚集在一起实现的。

*Great things are done by a series of small things brought together.*

—— 文森特·梵高 (Vincent van Gogh) | 荷兰后印象派画家 | 1853 ~ 1890



- ◀ `numpy.linalg.eig()` 特征值分解
- ◀ `numpy.linalg.svd()` 奇异值分解
- ◀ `sklearn.decomposition.PCA()` 主成分分析函数



# 18.1 回顾优化问题

本系列丛书《数学要素》第 19 章专门讲解过优化问题入门内容，本节稍作回顾。

## 极值、最值

优化问题好比在一定区域范围内，徒步寻找山谷或山峰。图 1 中的优化问题的目标函数  $f(x)$  就是海拔，优化变量是水平位置  $x$ 。

**极值** (extrema 或 local extrema) 是**极大值**和**极小值**的统称。白话讲，极值是搜索区域内所有的山峰和山谷，图 1 中 A、B、C、D、E 和 F 这六个点横坐标  $x$  值对应极值点。

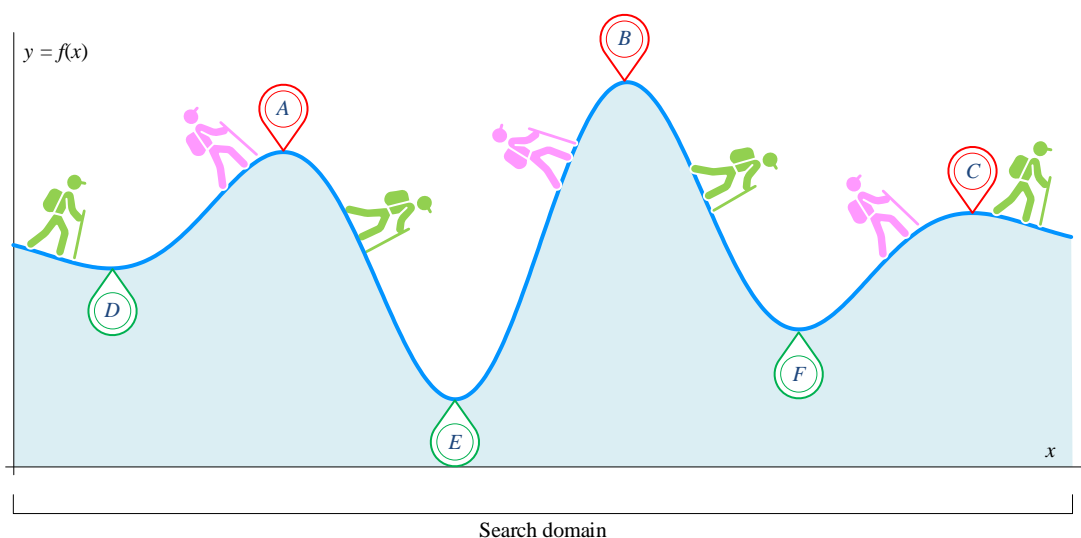


图 1. 爬上寻找山谷和山峰，图片来自《数学要素》

如果某个极值是整个指定搜索区域内的极大值或极小值，这个极值又被称作是**最大值** (maximum 或 global maximum) 或者**最小值** (minimum 或 global minimum)。最大值和最小值统称**最值** (global extrema)。

图 1 搜索域内有三座山峰 (A、B 和 C)，即搜索域极大值。而 B 是最高的山峰，因此 B 叫全局最大值，简称最大值，即站在 B 点一览众山小。E 是最深的山谷，因此 E 是全局最小值，简称最小值。

一般情况下，标准优化问题都是最小化优化问题。最大化优化问题的目标函数取个负号便转化为最小化优化问题。

## 含约束最小化优化问题

结合约束条件，完整最小化优化问题形式为：

$$\begin{aligned} \arg \min_x f(x) \\ \text{subject to: } l \leq x \leq u \\ Ax \leq b \\ A_{\text{eq}}x = b_{\text{eq}} \\ c(x) \leq 0 \\ c_{\text{eq}}(x) = 0 \end{aligned} \quad (1)$$

上式中，约束条件分为五类，按先后顺序：(a) **上下界** (lower and upper bounds)；(b) **线性不等式** (linear inequalities)；(c) **线性等式** (linear equalities)；(d) **非线性不等式** (nonlinear inequalities)；(e) **非线性等式** (nonlinear equalities)。

当约束条件存在时，如图 2 所示，最值可能出现在搜索区域内部或约束边界上。本章介绍的拉格朗日乘法就是一种能够把有约束优化问题转化成无约束优化问题的方法。

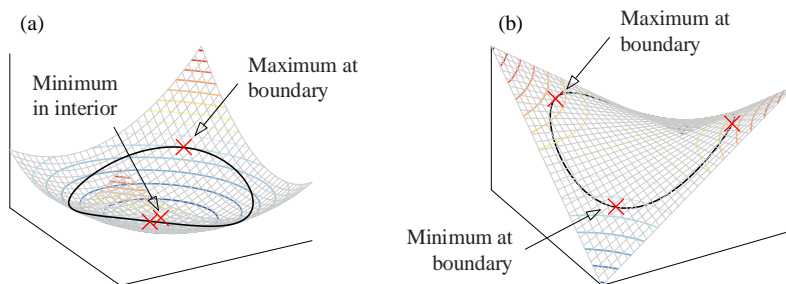


图 2. 最值和约束关系

➡ 《数学要素》还讲了如何利用导数和偏导数等数学工具求解一元和多元函数极值，本节就不再赘述。有必要的話，大家可以在学习本章之前先翻翻《数学要素》一册相关内容。

## 18.2 等式约束条件

**拉格朗日乘法** (method of Lagrange multiplier) 把有约束的优化问题转化为无约束优化问题。拉格朗日乘法是以 18 世纪法国著名数学家**约瑟夫·拉格朗日** (Joseph Lagrange) 命名。本章后续将主要从几何和数据视角来帮助大家理解拉格朗日乘法。

### 拉格朗日函数

给定含等式约束优化问题：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{subject to: } h(\mathbf{x}) = 0 \end{aligned} \quad (2)$$

其中， $f(\mathbf{x})$  和  $h(\mathbf{x})$  为连续函数。 $h(\mathbf{x}) = 0$  为等式约束条件。

构造拉格朗日函数 (Lagrangian function)  $L(\mathbf{x}, \lambda)$ ：

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda h(\mathbf{x}) \quad (3)$$

其中， $\lambda$  被称作拉格朗日乘子 (Lagrange multiplier)，或拉格朗日乘数。上式中， $\lambda$  前符号也可负号，不影响结果。本书正负号都用。

通过  $\lambda$ ，(2) 这个含等式约束优化问题便转化为一个无约束优化问题：

$$\begin{cases} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{subject to: } h(\mathbf{x}) = 0 \end{cases} \Rightarrow \arg \min_{\mathbf{x}} L(\mathbf{x}, \lambda) \quad (4)$$

$L(\mathbf{x}, \lambda)$  对  $\mathbf{x}$  和  $\lambda$  偏导都存在的情况下，最优解必要 (不是充分) 条件为一阶偏导数都零，即：

$$\begin{cases} \nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) = \frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla h(\mathbf{x}) = \mathbf{0} \\ \nabla_{\lambda} L(\mathbf{x}, \lambda) = \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = h(\mathbf{x}) = 0 \end{cases} \quad (5)$$

再次强调，(5) 存在一个重要前提，假定  $f(\mathbf{x})$  和  $h(\mathbf{x})$  在  $\mathbf{x}$  的某一邻域内均有连续一阶偏导。

(5) 中两式合并为：

$$\nabla_{\mathbf{x}, \lambda} L(\mathbf{x}, \lambda) = \mathbf{0} \quad (6)$$

求解上式得到驻点  $\mathbf{x}$ 。然后进一步判断驻点是极大值、极小值还是鞍点。

对于大部分读者来说，理解拉格朗日乘子法最大障碍在于下式：

$$\nabla f(\mathbf{x}) + \lambda \nabla h(\mathbf{x}) = \mathbf{0} \quad (7)$$

下面结合具体图形解释上式含义。

## 梯度向量方向

(7) 变形得到：

$$\nabla f(\mathbf{x}) = -\lambda \nabla h(\mathbf{x}) \quad (8)$$

(8) 等式隐含一条重要信息， $f(\mathbf{x})$  和  $h(\mathbf{x})$  在驻点  $\mathbf{x}$  处梯度同向或者反向。

图 3 中彩色等高线展示目标函数  $f(\mathbf{x})$  变化趋势，暖色系对应较大函数值，冷色系对应较小函数值。图中黑色直线对应  $h(\mathbf{x})$ ，即线性约束条件。换句话说，变量  $\mathbf{x}$  取值范围限定在图 5 黑色直线上。

图 3 中，等高线和黑色直线可以相交，甚至相切。相交意味着，交点处，沿着黑色直线稍微移动，函数值可能增大，也可能减小。这说明，交点处既不是最大值，也不是最小值。

然而，相切说明，在切点处，沿着黑色直线稍微移动，函数值有可能只朝着一个方向变动，即要么增大、要么减小。也就是说切点可能对应极值点，除非切点为驻点。

如果黑色直线和等高线相切，切点处  $f(\mathbf{x})$  和  $h(\mathbf{x})$  梯度向量平行（同向或反向）。这就是 (8) 的意义。

这种几何直觉就是理解 (8) 的“利器”。若梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  反向， $\lambda$  为正值，如图 3 (a) 所示。如果梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  正向， $\lambda$  为负值，如图 3 (b) 所示。简单来说， $h(\mathbf{x}) = 0$  约束下  $f(\mathbf{x})$  取得极值时，某点处梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  平行。

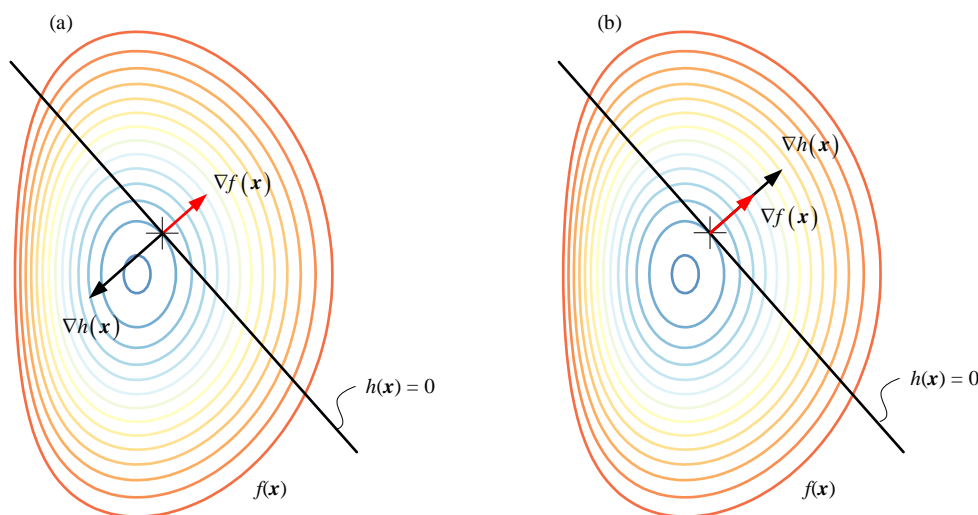


图 3. 线性等式约束条件拉格朗日算子几何意义

## 梯度平行

图 4 是图 3 (a) 局部视图，我们借助它进一步展示梯度平行的几何意义。

先看图 4 中 A 点，A 点黑色直线和某条等高线的切点。A 点处，梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  反向。梯度  $\nabla f(\mathbf{x})$  方向为函数  $f(\mathbf{x})$  上山方向，梯度下降方向  $-\nabla f(\mathbf{x})$  为函数  $f(\mathbf{x})$  下山方向。

A 点处， $f(\mathbf{x})$  在  $\mathbf{x}$  点处切线就是  $h(\mathbf{x})$ ，该切线垂直于  $\nabla h(\mathbf{x})$ ，也垂直于梯度  $\nabla f(\mathbf{x})$ 。显然，A 点处， $\nabla f(\mathbf{x})$  在  $h(\mathbf{x})$  方向标量投影为 0。

如图 4 所示, 若沿着  $h(x) = 0$  黑色直线向左或者向右偏离  $A$ ,  $f(x)$  都会增大 (对应等高线颜色从冷色系变为暖色系), 因此  $A$  点在  $h(x) = 0$  等式约束条件下为极小值点。根据目标函数曲面特征, 我们可以进一步确定该极小值点为最小值点。

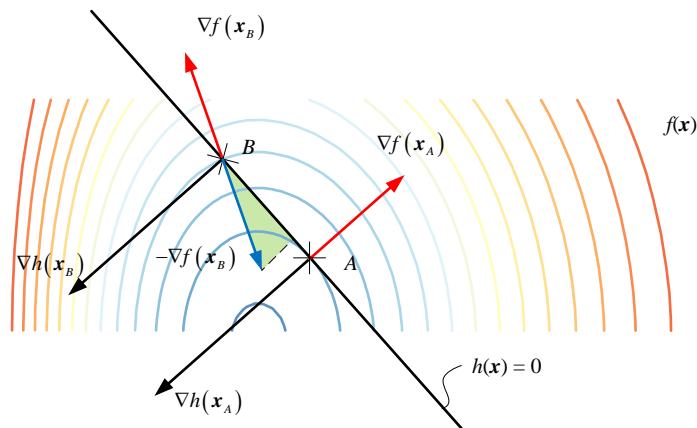


图 4. 梯度平行几何意义

再来看图 4 中  $B$  点,  $B$  点是黑色直线和某条等高线交点。同样找到  $f(x)$  梯度负方向  $-\nabla f(x)$ , 即  $f(x)$  下山方向; 容易发现  $-\nabla f(x)$  在  $h(x)$  方向, 在  $f(x)$  减小方向存在投影分量。这说明, 在  $B$  点沿着  $h(x)$  向右下方行走,  $f(x)$  进一步减小。因此,  $B$  点不是极值点。

注意, 本节没有使用“最值”这一说法, 这是因为对于多极值曲面, 曲面和线性约束条件可能存在多个“切点”, 可能对应若干“极值”。

## 非线性等式约束条件

上述分析思路也同样适用于非线性等式约束条件。请大家用“交点 + 切点”和“梯度向量投影”两个视角自行分析图 5 两幅子图。

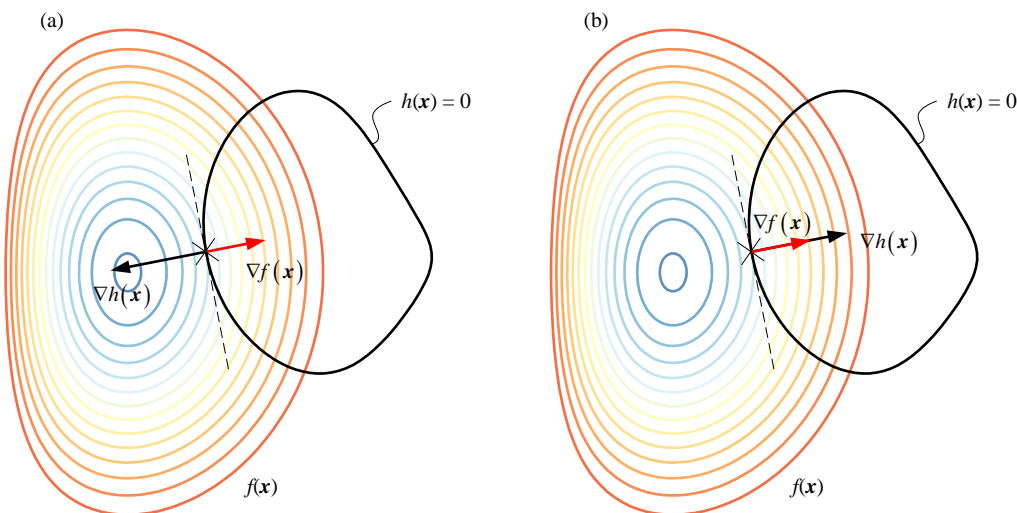


图 5. 非线性等式约束条件拉格朗日算子几何意义

## 进一步判断

用拉格朗日乘子计算出来的驻点到底是极大值、极小值、鞍点，还需要进一步判断。

图 6 给出四种极值常见情况。如图 6 (a) 所示， $f(x)$  自身为凹函数， $f(x)$  等高线和  $h(x) = 0$  相切于  $A$  点和  $B$  点。在  $h(x) = 0$  约束条件下， $f(x)$  在  $A$  点取得极大值，在  $B$  点取得极小值。进一步判断， $A$  为最大值点， $B$  为最小值点。

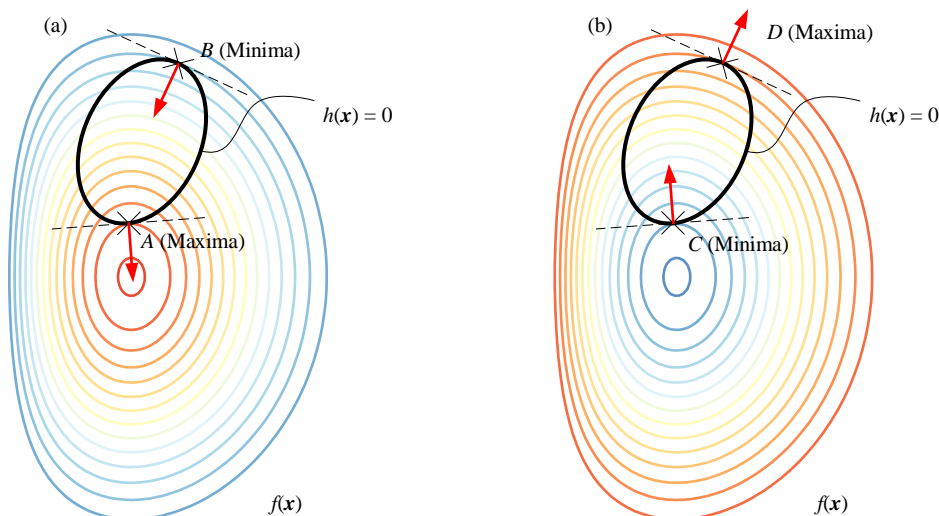


图 6. 四种极值情况

而在图 6 (b)， $f(x)$  自身为凸函数， $f(x)$  等高线和  $h(x) = 0$  相切于  $C$  点和  $D$  点；在  $h(x) = 0$  约束条件下， $f(x)$  在  $C$  点取得极小值，在  $D$  点取得极大值。进一步判断， $C$  为最小值点， $D$  为最大值点。

⚠ 这里请大家注意，如果  $h(x) = 0$  为等式约束，不需要关注  $h(x)$  自身函数值变化趋势。但是，不等式约束  $g(x) \leq 0$  就必须考虑  $g(x)$  函数自身变化趋势，本章后续将讨论这个话题。



说个题外话，天文中的**拉格朗日点** (Lagrangian point) 很可能比本章介绍的拉格朗日乘子法更出名。

两个天体环绕运行，比如太阳-地球(日-地)、地球-月亮(地-月)，在空间中可以找到满足两个天体引力平衡五个点，如图 7 所示的  $L_1 \sim L_5$ 。这五个点叫做拉格朗日点。欧拉于 1767 年推算出前三个拉格朗日点，拉格朗日于 1772 年推导证明剩下两个。



在  $L_1 \sim L_5$  这五个点任意一点放置质量可以忽略不计的第三个天体，使其和另外两个天体以相同模式运转，这就是所谓的**三体问题** (three-body problem)。

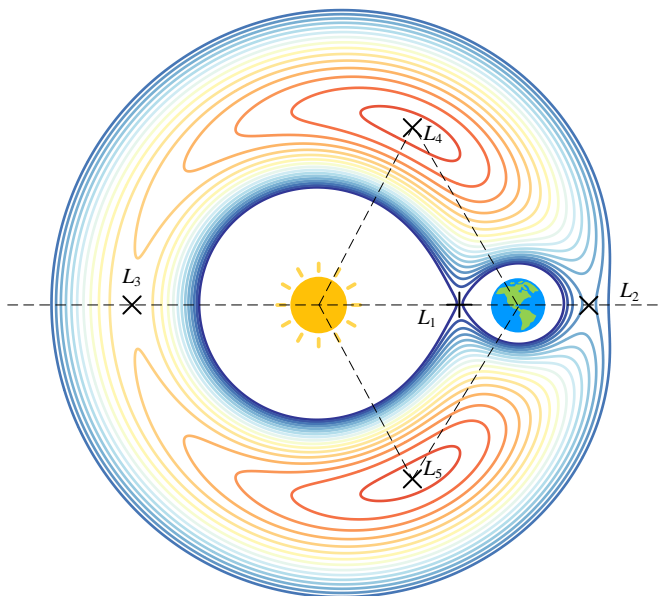


图 7. 五个拉格朗日点

实际情况，第三天体不可能在拉格朗日点保持相对静止；人造卫星一般会围绕拉格朗日点附近运转，完成观测或中继等任务，以节省大量燃料。

嫦娥二号完成探月任务后，专门飞往“日-地”拉格朗日  $L_2$  点进行科学探测。我国探月时用到的鹊桥中继星就是绕“地-月”拉格朗日  $L_2$  点运转。詹姆斯·韦伯空间望远镜绕“日-地”拉格朗日  $L_2$  点运转。

之所以聊到这个话题是因为图 7 所示拉格朗日点、引力场等高线图 and 驻点、极值、梯度向量场这些概念都有密切的关系。

## 18.3 线性等式约束

下面用一个简单例子来解释上一节介绍的等式约束优化问题。

给定一个优化问题如下：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) &= x_1^2 + x_2^2 \\ \text{subject to: } h(\mathbf{x}) &= x_1 + x_2 - 1 = 0 \end{aligned} \quad (9)$$

这是一个二次规划问题，含一个线性等式约束条件  $h(\mathbf{x}) = 0$ 。

利用矩阵运算，(9) 可以写成：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) &= \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|_2^2 \\ \text{subject to: } h(\mathbf{x}) &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \mathbf{x} - 1 = 0 \end{aligned} \quad (10)$$

根据上一章内容，请大家自行计算两个函数的梯度向量。

图 8 所示为  $h(x)$  梯度向量场。观察图像，我们发现  $h(x) = 0$  对应一条直线，直线上不同点处的梯度向量均垂直于该直线。

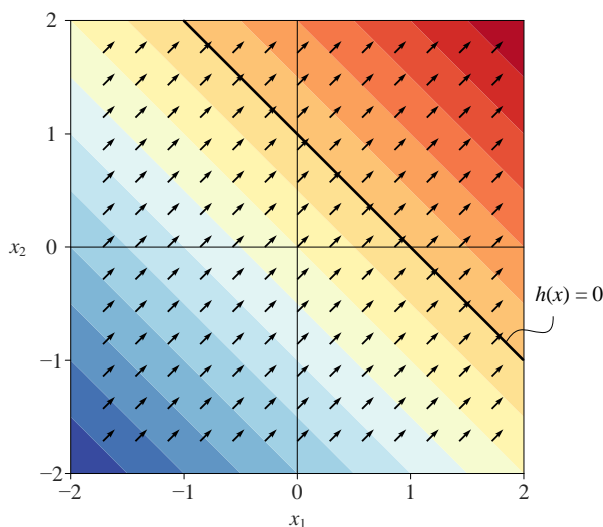
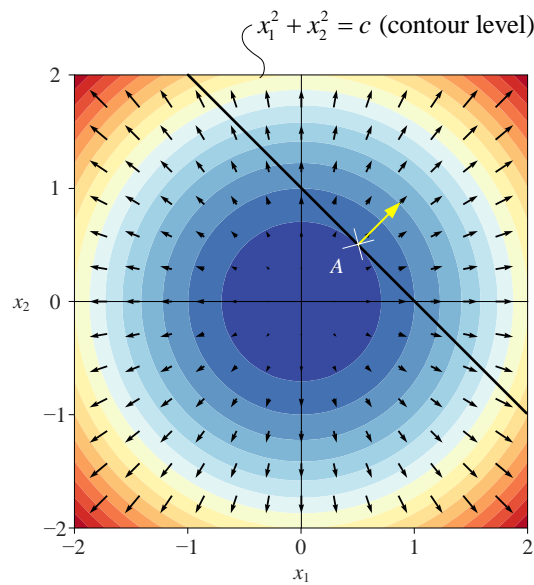


图 8.  $h(x)$  梯度向量场

如图 9 所示， $x_1x_2$  平面上，目标函数  $f(x)$  的等高线是一组同心圆。等式约束条件  $x_1 + x_2 - 1 = 0$  对应图中黑色直线。优化解只能在  $x_1 + x_2 - 1 = 0$  限定的直线上选取。

图 9 中，黄色箭头代表  $h(x)$  梯度方向，图中的黑色箭头是  $f(x)$  梯度向量场。当同心圆和等式约束相切于 A 点， $f(x)$  取得最小值。显然，A 点处  $f(x)$  和  $h(x)$  梯度方向一致，或称平行。

黑色直线 ( $h(x) = 0$ ) 上任何偏离 A 点位置都会导致目标函数  $f(x)$  增大。

图 9.拉格朗日算子求解二次规划，极值点 A 处  $f(x)$  和  $h(x)$  梯度同向， $\lambda$  小于 0

## 拉格朗日函数

构造拉格朗日函数  $L(\mathbf{x}, \lambda)$ :

$$L(\mathbf{x}, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1) \quad (11)$$

构造下列偏导为 0 等式组并求解  $(x_1, x_2, \lambda)$ :

$$\begin{cases} \frac{\partial L(\mathbf{x}, \lambda)}{\partial x_1} = 2x_1 + \lambda = 0 \\ \frac{\partial L(\mathbf{x}, \lambda)}{\partial x_2} = 2x_2 + \lambda = 0 \\ \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = x_1 + x_2 - 1 = 0 \end{cases} \Rightarrow \begin{cases} x_1 = \frac{1}{2} \\ x_2 = \frac{1}{2} \\ \lambda = -1 \end{cases} \quad (12)$$

$\lambda$  为负值，这说明在优化解处，梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  同向。

将  $\lambda = -1$  代回 (11) 得到如图 10 所示的拉格朗日函数  $L(\mathbf{x}, \lambda = -1)$  平面等高线。图 10 中我们发现  $L(\mathbf{x}, \lambda = -1)$  最小值位置就是 (12) 的优化解。

从图像角度，我们将图 9 这个含有线性等式约束的优化问题转化成图 10 这个无约束优化问题。

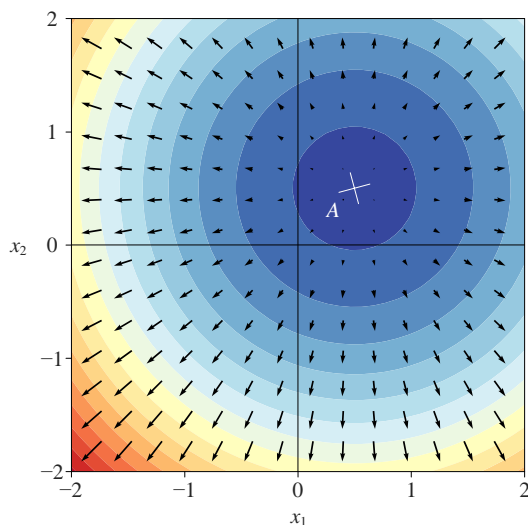


图 10. 拉格朗日函数平面等高线

### 另外一种记法

前文提过，很多文献  $\lambda$  前采用负号，拉格朗日函数  $L(\mathbf{x}, \lambda)$  则为：

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda h(\mathbf{x}) \quad (13)$$

$L(\mathbf{x}, \lambda)$  对  $\mathbf{x}$  和  $\lambda$  偏导为 0 对应等式组为：

$$\begin{cases} \nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) = \frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \nabla f(\mathbf{x}) - \lambda \nabla h(\mathbf{x}) = \mathbf{0} \\ \nabla_{\lambda} L(\mathbf{x}, \lambda) = \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = h(\mathbf{x}) = 0 \end{cases} \quad (14)$$

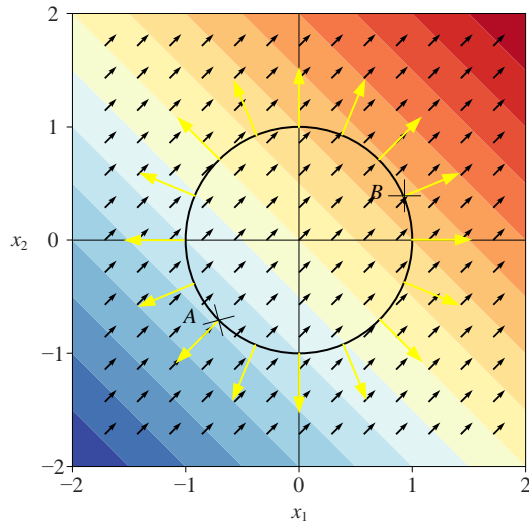
这种拉格朗日函数构造，若梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  同向， $\lambda$  为正值。如果梯度  $\nabla f(\mathbf{x})$  和梯度  $\nabla h(\mathbf{x})$  反向， $\lambda$  为负值。不管  $\lambda$  前是正还是负，都不会影响结果。本章后续也会使用 (13) 这种形式。

## 18.4 非线性等式约束

本节再看一个线性规划问题实例，它约束条件为非线性等式约束：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) &= x_1 + x_2 \\ \text{subject to: } h(\mathbf{x}) &= x_1^2 + x_2^2 - 1 = 0 \end{aligned} \quad (15)$$

图 11 所示为， $f(\mathbf{x})$  和  $h(\mathbf{x}) = 0$  梯度向量场。大家自己是否能够根据图 11 梯度向量之间的关系，判断 (15) 极大值和极小值位置。

图 11.  $f(x)$  和  $h(x)=0$  梯度向量场

## 拉格朗日函数

构造拉格朗日函数  $L(\mathbf{x}, \lambda)$  如下：

$$L(\mathbf{x}, \lambda) = x_1 + x_2 + \lambda(x_1^2 + x_2^2 - 1) \quad (16)$$

根据偏导为 0 构造如下等式组：

$$\begin{cases} \frac{\partial L(\mathbf{x}, \lambda)}{\partial x_1} = 1 + 2x_1\lambda = 0 \\ \frac{\partial L(\mathbf{x}, \lambda)}{\partial x_2} = 1 + 2x_2\lambda = 0 \\ \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = x_1^2 + x_2^2 - 1 = 0 \end{cases} \Rightarrow \begin{cases} x_1 = -\frac{1}{2\lambda} \\ x_2 = -\frac{1}{2\lambda} \\ x_1^2 + x_2^2 - 1 = 0 \end{cases} \quad (17)$$

根据上述等式组构造  $\lambda$  等式，并求解  $\lambda$ ：

$$\left(\frac{1}{2\lambda}\right)^2 + \left(\frac{1}{2\lambda}\right)^2 - 1 = 0 \Rightarrow \lambda = \pm \frac{\sqrt{2}}{2} \quad (18)$$

$\lambda$  取正值获得最小值：

$$\begin{cases} x_1 = -\frac{\sqrt{2}}{2} \\ x_2 = -\frac{\sqrt{2}}{2} \\ \lambda = \frac{\sqrt{2}}{2} \end{cases} \quad (19)$$

$\lambda$  取负值获得最大值。

图 12 所示为拉格朗日函数  $L(\mathbf{x}, \lambda = \sqrt{2}/2)$  对应的平面等高线图。同样，利用拉格朗日乘子法，我们将如图 11 所示的含有非线性等式约束的优化问题，转化成如图 12 所示的无约束优化问题。

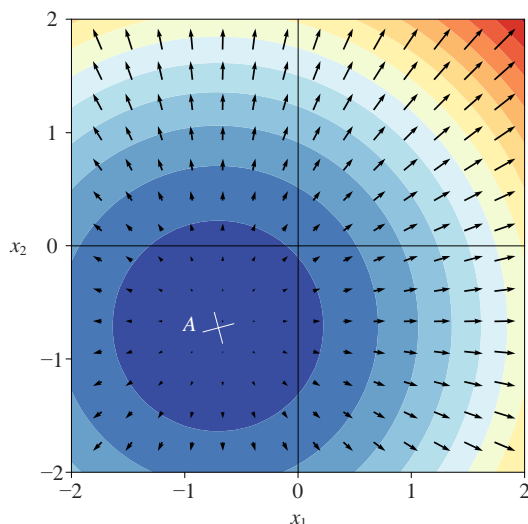


图 12. 拉格朗日函数等高线

## 18.5 不等式约束

本节介绍如何用 **KKT** (Karush-Kuhn-Tucker) 条件将本章前文介绍的拉格朗日乘子法推广到不等式约束问题。

给定如下不等式约束优化问题：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{subject to: } g(\mathbf{x}) \leq 0 \end{aligned} \quad (20)$$

其中， $f(\mathbf{x})$  和  $g(\mathbf{x})$  为连续函数。

### 几何视角

如图 13 所示，黑色曲线和图 5 一样，代表等式情况，即  $g(\mathbf{x}) = 0$ 。图 13 中浅蓝色区域代表  $g(\mathbf{x}) < 0$  情况。

优化解  $\mathbf{x}$  出现位置有两种情况：第一种情况， $\mathbf{x}$  出现在边界上（黑色线），约束条件有效，如图 13 (a)；第二种情况， $\mathbf{x}^*$  出现在不等式区域内（浅蓝色背景），约束条件无效，如图 13 (b)。

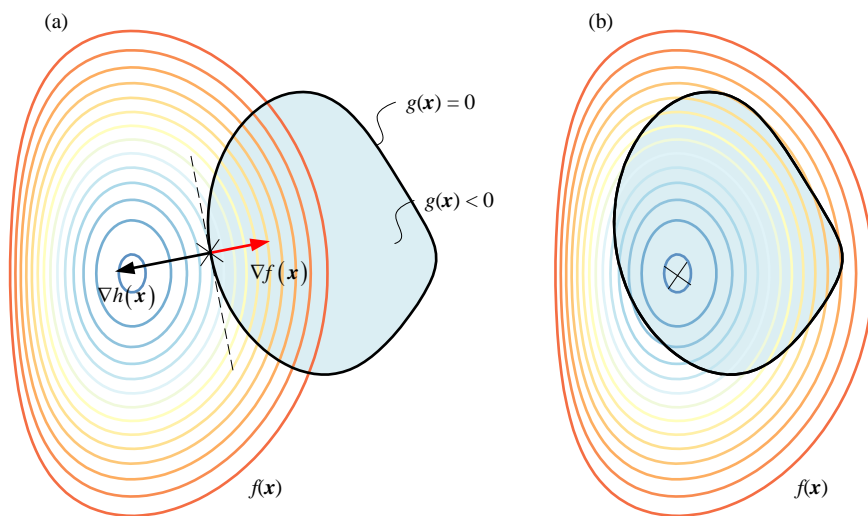


图 13. 不等式约束条件下拉格朗日乘法两种情况

在图 13(a) 中，第一种情况等价于图 5 讨论情况，即  $g(\mathbf{x}) = 0$  成立。

在图 13(b) 中，最优解  $\mathbf{x}$  出现在  $g(\mathbf{x}) < 0$  蓝色区域内。对于凸函数，如果在最优解的邻域内  $f(\mathbf{x})$  有连续的一阶偏导数，可以直接通过  $\nabla f(\mathbf{x}) = 0$  获得最优解，此时  $\lambda$  为 0。这种情况，含约束优化问题直接变为无约束问题。

结合上述两种情况， $\lambda g(\mathbf{x}) = 0$  恒成立。也就是说，要么  $g(\mathbf{x}) = 0$  (图 13(a))，要么  $\lambda = 0$  (图 13(b))。

### 判断极值点性质

进一步讨论图 13 (a) 对应的情况。如图 14 所示，不等式内部区域  $g(\mathbf{x}) < 0$ ，而边界  $g(\mathbf{x}) = 0$ 。而黑色边界外， $g(\mathbf{x}) > 0$ 。因此，在黑色边界  $g(\mathbf{x}) = 0$  上，梯度向量指向区域外部。

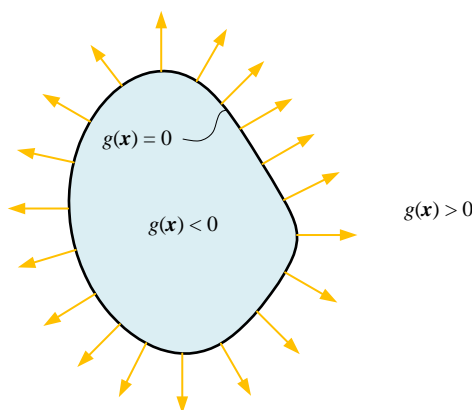


图 14. 不等式约束梯度方向

图 15 所示为  $\nabla f(\mathbf{x})$  和梯度  $\nabla g(\mathbf{x})$  反向和同向两种情况。

图 15 (a) 中,  $A$  点处,  $f(\mathbf{x})$  梯度  $\nabla f(\mathbf{x})$  是黑色箭头, 指向右上方。而  $A$  点处,  $g(\mathbf{x})$  梯度  $\nabla g(\mathbf{x})$  是橙色箭头, 和  $\nabla f(\mathbf{x})$  同向。 $A$  点为  $g(\mathbf{x}) \leq 0$  不等式条件约束下  $f(\mathbf{x})$  极大值。

图 15(b)中,  $B$  点处,  $\nabla f(\mathbf{x})$  和  $\nabla g(\mathbf{x})$  方向相反, 也就是  $\lambda > 0$ 。 $B$  点是  $g(\mathbf{x}) \leq 0$  不等式条件约束下  $f(\mathbf{x})$  极小值。

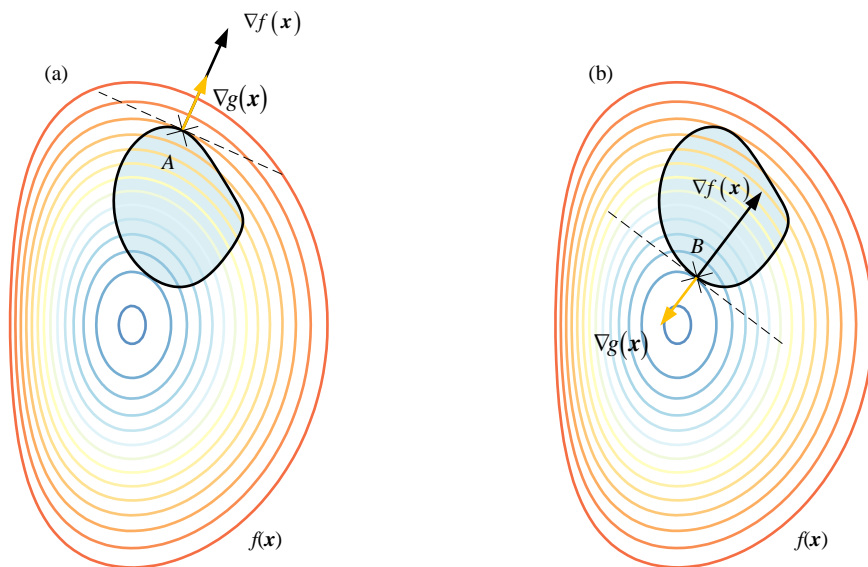


图 15. 梯度向量同方向和反方向

### KKT 条件

结合以上讨论, 对于  $g(\mathbf{x}) \leq 0$  不等式条件约束下  $f(\mathbf{x})$  最小值问题, 构造如下拉格朗日函数  $L(\mathbf{x}, \lambda)$ :

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) \quad (21)$$

极小点  $\mathbf{x}$  出现位置满足以下条件:

$$\begin{cases} \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0 \\ g(\mathbf{x}) \leq 0 \\ \lambda \geq 0 \\ \lambda g(\mathbf{x}) = 0 \end{cases} \quad (22)$$

以上这些条件合称 KKT 条件。

### 合并两类约束条件



在不等式  $g(\mathbf{x}) \leq 0$  及等式约束  $h(\mathbf{x}) = 0$  条件下，构造最小化  $f(\mathbf{x})$  优化问题：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{subject to: } g(\mathbf{x}) \leq 0, h(\mathbf{x}) = 0 \end{aligned} \quad (23)$$

构造拉格朗日函数：

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda_h h(\mathbf{x}) + \lambda_g g(\mathbf{x}) \quad (24)$$

KKT 条件如下：

$$\begin{cases} \nabla f(\mathbf{x}) + \lambda_h \nabla h(\mathbf{x}) + \lambda_g \nabla g(\mathbf{x}) = 0 \\ h(\mathbf{x}) = 0 \\ g(\mathbf{x}) \leq 0 \\ \lambda_g \geq 0 \\ \lambda_g g(\mathbf{x}) = 0 \end{cases} \quad (25)$$

### 多个约束条件

有以上讨论，把 (25) 推广到多个等式约束和多个不等式约束情况。

对于如下优化问题：

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{subject to: } \begin{cases} h_i(\mathbf{x}) = 0, & i = 1, \dots, n \\ g_j(\mathbf{x}) \leq 0, & j = 1, \dots, m \end{cases} \end{aligned} \quad (26)$$

构造如下拉格朗日函数：

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum \lambda_{h,i} h_i(\mathbf{x}) + \sum \lambda_{g,j} g_j(\mathbf{x}) \quad (27)$$

(27) 对应的 KKT 条件如下：

$$\begin{cases} \nabla_{\mathbf{x}, \lambda} L(\mathbf{x}, \lambda) = 0 \\ h_i(\mathbf{x}) = 0 \\ g_j(\mathbf{x}) \leq 0 \\ \lambda_{g,j} \geq 0 \\ \lambda_{g,j} g_j(\mathbf{x}) = 0, \quad \forall j \end{cases} \quad (28)$$

## 18.6 再谈特征值分解：优化视角

这一节介绍一些线性代数中会遇到的含约束优化问题。利用拉格朗日乘法，它们最终都可以用特征值分解求解。

### 第一个优化问题

给定如下优化问题：

$$\begin{aligned} \arg \max_{\mathbf{v}} \quad & \mathbf{v}^T \mathbf{A} \mathbf{v} \\ \text{subject to: } & \mathbf{v}^T \mathbf{v} = 1 \end{aligned} \quad (29)$$

其中， $\mathbf{A}$  为对称矩阵，列向量  $\mathbf{v}$  为优化变量。优化问题的等式约束条件是  $\mathbf{v}$  为单位向量。

构造拉格朗日函数：

$$L(\mathbf{v}, \lambda) = \mathbf{v}^T \mathbf{A} \mathbf{v} - \lambda (\mathbf{v}^T \mathbf{v} - 1) \quad (30)$$

⚠ 注意，为了满足特征值分解常用记法，(30) 中  $\lambda$  前采用负号。

$L(\mathbf{v}, \lambda)$  对  $\mathbf{v}$  偏导为  $\mathbf{0}$ ，得到等式：

$$\frac{\partial L(\mathbf{v}, \lambda)}{\partial \mathbf{v}} = 2\mathbf{A}\mathbf{v} - 2\lambda\mathbf{v} = \mathbf{0} \quad (31)$$

整理得到：

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad (32)$$

最大化问题中，最优解为  $\lambda_{\max}$ ，特征向量  $\mathbf{v}$  对应矩阵  $\mathbf{A}$  最大特征值  $\lambda_{\max}$ 。

如果是最小化问题，即：

$$\begin{aligned} \arg \min_{\mathbf{v}} \quad & \mathbf{v}^T \mathbf{A} \mathbf{v} \\ \text{subject to: } & \mathbf{v}^T \mathbf{v} = 1 \end{aligned} \quad (33)$$

最优解特征向量  $\mathbf{v}$  对应矩阵  $\mathbf{A}$  最小特征值  $\lambda_{\min}$ 。

此外，(29) 约束条件也可以写成：

$$\|\mathbf{v}\|_2 = 1, \quad \|\mathbf{v}\|_2^2 = 1 \quad (34)$$

### 第二个优化问题

给定如下优化问题：

$$\arg \max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad (35)$$

上式中， $\mathbf{A}$  为已知数据矩阵， $\mathbf{x}$  为优化变量。注意， $\mathbf{x}^T \mathbf{x}$  在分母上，因此  $\mathbf{x}$  不能为零向量  $\mathbf{x}$ 。这就是本书第 14 章讲的瑞利商。上述优化问题等价于 (29)。本书前文多次强调过，上式分子、分母都是标量。

类似 (35)，最小化优化问题：

$$\arg \min_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad (36)$$

上式等价于 (33)。

### 第三个优化问题

给定优化问题：

$$\begin{aligned} \arg \max_{\mathbf{v}} \quad & \mathbf{v}^T \mathbf{A} \mathbf{v} \\ \text{subject to: } & \mathbf{v}^T \mathbf{B} \mathbf{v} = 1 \end{aligned} \quad (37)$$

构造拉格朗日函数：

$$L(\mathbf{v}, \lambda) = \mathbf{v}^T \mathbf{A} \mathbf{v} - \lambda (\mathbf{v}^T \mathbf{B} \mathbf{v} - 1) \quad (38)$$

$L(\mathbf{v}, \lambda)$  对  $\mathbf{v}$  偏导为 0，得到等式：

$$\frac{\partial L(\mathbf{v}, \lambda)}{\partial \mathbf{v}} = 2\mathbf{A} \mathbf{v} - 2\lambda \mathbf{B} \mathbf{v} = 0 \quad (39)$$

整理得到：

$$\mathbf{A} \mathbf{v} = \lambda \mathbf{B} \mathbf{v} \quad (40)$$

如果  $\mathbf{B}$  可逆，上式相当于对  $\mathbf{B}^{-1} \mathbf{A}$  进行特征值分解。特别地，当  $\mathbf{B} = \mathbf{I}$  时对应 (29)。

### 第四个优化问题

给定优化问题：

$$\arg \max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{B} \mathbf{x}} \quad (41)$$

上式实际上是瑞利商的一般式。这个优化问题等价于 (37)。一般情况，矩阵  $\mathbf{B}$  为正定，这样  $\mathbf{x} \neq \mathbf{0}$  时， $\mathbf{x}^T \mathbf{B} \mathbf{x} > 0$ 。

令：

$$\mathbf{x} = \mathbf{B}^{-\frac{1}{2}} \mathbf{y} \quad (42)$$

代入 (41) 中的目标函数，得到：

$$\frac{\left(\mathbf{B}^{-\frac{1}{2}}\mathbf{y}\right)^{\mathrm{T}}\mathbf{A}\left(\mathbf{B}^{-\frac{1}{2}}\mathbf{y}\right)}{\left(\mathbf{B}^{-\frac{1}{2}}\mathbf{y}\right)^{\mathrm{T}}\mathbf{B}\left(\mathbf{B}^{-\frac{1}{2}}\mathbf{y}\right)} = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{B}^{-\frac{1}{2}\mathrm{T}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{y}} \quad (43)$$

如果  $\mathbf{B}$  为正定矩阵， $\mathbf{B}$  的特征值分解可以写成：

$$\mathbf{B} = \mathbf{V}\mathbf{A}\mathbf{V}^{\mathrm{T}} \quad (44)$$

而  $\mathbf{B}^{-\frac{1}{2}}$  为：

$$\mathbf{B}^{-\frac{1}{2}} = \mathbf{V}\mathbf{A}^{-\frac{1}{2}}\mathbf{V}^{\mathrm{T}} \quad (45)$$

请大家自己将 (45) 代入 (43)，并完成推导。

## 第五个优化问题

给定优化问题：

$$\begin{aligned} \arg \min_{\mathbf{v}} \quad & \|\mathbf{A}\mathbf{v}\| \\ \text{subject to: } & \|\mathbf{v}\| = 1 \end{aligned} \quad (46)$$

(46) 也等价于：

$$\begin{aligned} \arg \min_{\mathbf{v}} \quad & \mathbf{v}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{v} \\ \text{subject to: } & \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1 \end{aligned} \quad (47)$$

(46) 还等价于：

$$\arg \min_{\mathbf{x} \neq \mathbf{0}} \left( \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \right)^2 = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \quad (48)$$

注意， $\mathbf{x}$  不能是零向量  $\mathbf{0}$ 。

(46) 也等价于：

$$\arg \min_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \quad (49)$$

式中，对  $\mathbf{A}$  是否为对称矩阵没有限制，因为  $\mathbf{A}^{\mathrm{T}}\mathbf{A}$  为对称矩阵。对  $\mathbf{A}^{\mathrm{T}}\mathbf{A}$  的特征值分解，便可以解决这个优化问题。这个优化问题实际上就是我们要在本章后文要讨论的 SVD 分解的优化视角。

## 18.7 再谈 SVD：优化视角

本节从优化视角再讨论 SVD 分解。

### 从投影说起

如图 16 所示，数据矩阵  $X$  中任意行向量  $\mathbf{x}^{(i)}$  在  $\mathbf{v}$  上投影，得到标量投影结果为  $y^{(i)}$ ：

$$\mathbf{x}^{(i)} \mathbf{v} = y^{(i)} \quad (50)$$

其中， $\mathbf{v}$  为单位向量。

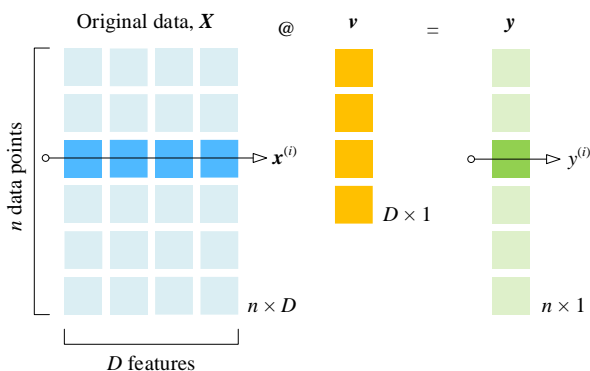


图 16. 数据矩阵  $X$  中任意行向量  $\mathbf{x}^{(i)}$  在  $\mathbf{v}$  上投影

如图 17 所示， $y^{(i)}$  就是  $\mathbf{x}^{(i)}$  在  $\mathbf{v}$  上坐标， $h^{(i)}$  为  $\mathbf{x}^{(i)}$  到  $\mathbf{v}$  的距离。

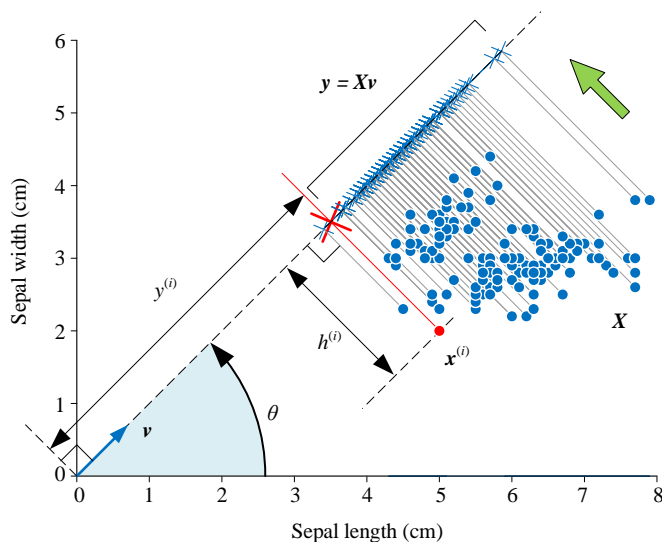


图 17.  $\mathbf{x}^{(i)}$  在  $\mathbf{v}$  上投影

整个数据矩阵  $X$  在  $\mathbf{v}$  上投影得到向量  $\mathbf{y}$ :

$$X\mathbf{v} = \mathbf{y} \quad (51)$$

数据矩阵  $X$  对应图 17 中的圆点  $\bullet$ ,  $\mathbf{y}$  对应图 17 中的叉  $\times$ 。

### 构造优化问题

从优化问题角度, SVD 分解等价于最大化  $y^{(i)}$  平方和:

$$\max_{\mathbf{v}} \sum_{i=1}^n (y^{(i)})^2 \quad (52)$$

上式相当于, 最小化  $h^{(i)}$  平方和:

$$\min \sum_{i=1}^n (h^{(i)})^2 \quad (53)$$

而如下几个式子等价,

$$\sum_{i=1}^n (y^{(i)})^2 = \|\mathbf{y}\|_2^2 = \mathbf{y}^T \mathbf{y} = (\mathbf{X}\mathbf{v})^T (\mathbf{X}\mathbf{v}) = \mathbf{v}^T \underbrace{\mathbf{X}^T \mathbf{X}}_{\mathbf{G}} \mathbf{v} \quad (54)$$

大家是否看到了 (48) 的影子。

构造如下优化问题:

$$\begin{aligned} \mathbf{v}_1 = \arg \max_{\mathbf{v}} \quad & \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \\ \text{subject to: } & \mathbf{v}^T \mathbf{v} = 1 \end{aligned} \quad (55)$$

上式中,  $X$  为已知数据矩阵,  $\mathbf{v}$  为优化变量。

(55) 等价于:

$$\mathbf{v}_1 = \arg \max_{\mathbf{v}} \frac{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \quad (56)$$

利用  $L^2$  范数, (55) 还等价于:

$$\begin{aligned} \mathbf{v}_1 = \arg \max_{\mathbf{v}} \quad & \|\mathbf{X}\mathbf{v}\| \\ \text{subject to: } & \|\mathbf{v}\| = 1 \end{aligned} \quad (57)$$

(55) 也等价于:

$$\arg \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{X}\mathbf{x}\|}{\|\mathbf{x}\|} \quad (58)$$

上式中,  $\mathbf{x}$  为优化变量。

对  $X$  进行奇异值分解得到的最大奇异值  $s_1$  满足：

$$s_1 = \|Xv_1\| = \|y_1\|_2 = \sqrt{\sum_{i=1}^n (y_1^{(i)})^2} \quad (59)$$

其中， $Xv_1 = y_1$ 。也就是说，奇异值  $s_1$  代表， $X$  行向量在  $v$  方向上投影结果  $y$  的模的最大值。

格拉姆矩阵  $X^T X$  最大特征值  $\lambda_1$  满足：

$$\lambda_1 = s_1^2 = \|Xv_1\|_2^2 = \|y_1\|_2^2 = \sum_{i=1}^n (y_1^{(i)})^2 \quad (60)$$

请大家格外注意理解这个优化视角，它阐释了奇异值分解的内核。

### 顺序求解其他右奇异向量

确定第一右奇异向量  $v_1$  之后，我们可以依次构造类似如下优化问题求解其他右奇异向量：

$$\begin{aligned} v_2 = \arg \max_v \|Xv\| \\ \text{subject to: } \|v\| = 1, v \perp v_1 \end{aligned} \quad (61)$$


上式等价于：

$$\begin{aligned} v_2 = \arg \max_v v^T X^T X v \\ \text{subject to: } \|v\| = 1, v \perp v_1 \end{aligned} \quad (62)$$

### 中心化数据

数据矩阵  $X$  中每一列数据  $x_j$  分别减去本列均值可以得到中心化数据  $X_c$ 。利用广播原则， $X$  减去行向量  $E(X)$  得到  $X_c$ ：

$$X_c = X - E(X) \quad (63)$$

 特别强调，SVD 分解中心化数据  $X_c$  得到的结果一般不同于 SVD 分解原数据矩阵  $X$ 。

如图 18 所示，中心化数据  $X_c$  在  $v$  上投影得到向量  $y_c$ ：

$$X_c v = y_c \quad (64)$$

图 18 对应的优化问题为：

$$\begin{aligned} v_{c-1} = \arg \max_v \|X_c v\| \\ \text{subject to: } \|v\| = 1 \end{aligned} \quad (65)$$

$X_c$  的最大奇异值  $s_{c-1}$  为：

$$s_{c-1} = \|\mathbf{X}_c \mathbf{v}_{c-1}\| \quad (66)$$

也就是说， $s_{c-1}$  的平方为  $\mathbf{X}_c$  所有点在  $\mathbf{v}_{c-1}$  方向上标量投影的平方值之和的最大值：

$$\begin{aligned} s_1^2 &= \|\mathbf{X}_c \mathbf{v}_{c-1}\|_2^2 = \sum_{i=1}^n (y_c^{(i)})^2 = \|\mathbf{y}_c\|_2^2 = \mathbf{y}_c^T \mathbf{y}_c \\ &= (\mathbf{X}_c \mathbf{v}_{c-1})^T (\mathbf{X}_c \mathbf{v}_{c-1}) = \mathbf{v}_{c-1}^T \underbrace{\mathbf{X}_c^T \mathbf{X}_c}_{(n-1)\Sigma} \mathbf{v}_{c-1} = (n-1) \mathbf{v}_{c-1}^T \Sigma \mathbf{v}_{c-1} \end{aligned} \quad (67)$$

相信大家已经注意到上式中的协方差矩阵。大家可能会对 (67) 感到困惑，SVD 分解怎么和协方差矩阵  $\Sigma$  扯到一起？这是本书最后三章要回答的问题。

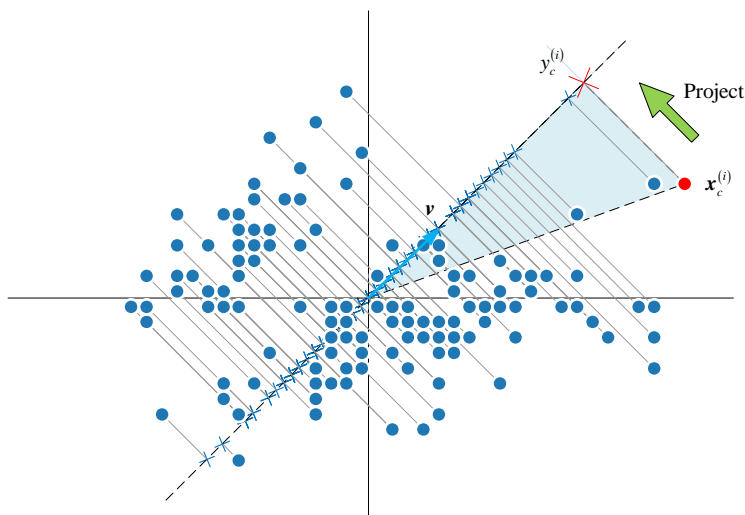


图 18. 中心化数据在  $\mathbf{v}$  上投影

## 18.8 矩阵范数：矩阵 $\rightarrow$ 标量，矩阵“大小”

有了上一节的优化视角，本节要介绍几种机器学习算法中常用的**矩阵范数** (matrix norm)。矩阵范数相当于向量范数的推广。本书第 3 章讲过向量范数代表某种“距离”，计算向量范数是某种“向量  $\rightarrow$  标量”映射。

类似向量范数，矩阵范数也是某种基于特定规则的“矩阵  $\rightarrow$  标量”映射。矩阵范数也从不同角度度量了矩阵的“大小”。

### 矩阵 $p$ -范数

形状为  $m \times n$  矩阵  $\mathbf{A}$  的  $p$ -范数定义为：



$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \quad (68)$$

大家是否已经看到类似 (49) 的形式。

本节内容以如下矩阵  $A$  为例：

$$A_{m \times n} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}_{3 \times 2} \quad (69)$$

### 矩阵 1-范数

矩阵  $A$  的 1-范数，也叫**列元素绝对值之和最大范数** (maximum absolute column sum norm)，具体定义如下：

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{i,j}| \quad (70)$$

(69) 给出的矩阵  $A$  有 2 列，先计算每一列元素绝对值之和，然后再取出其中最大值。这个最大值就是矩阵  $A$  的 1-范数：

$$\|A\|_1 = \max(0+1+1, 1+1+0) = \max(2, 2) = 2 \quad (71)$$

### 矩阵 $\infty$ -范数

矩阵  $A$  的  $\infty$ -范数，也叫**行元素绝对值之和最大范数** (maximum absolute row sum norm)，具体定义如下：

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{i,j}| \quad (72)$$

(69) 给出的矩阵  $A$  有 3 行，先计算每一行元素绝对值之和，然后再取出其中最大值。这个最大值就是矩阵  $A$  的  $\infty$ -范数：

$$\|A\|_\infty = \max(0+1, 1+1, 1+0) = \max(1, 2, 1) = 2 \quad (73)$$

### 矩阵 2-范数

矩阵  $A$  的 2-范数就要用 (49) 这个优化问题。矩阵  $A$  的 2-范数具体定义如下：

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = s_1 = \sqrt{\lambda_1} \quad (74)$$

根据本章前文所讲,  $\|A\|_2$  的值时矩阵  $A$  奇异值分解中最大奇异值  $s_1 = \sqrt{3}$ 。本书第 11 章手算过矩阵  $A$  的奇异值分解。

$\|A\|_2$  也是  $A$  的格拉姆矩阵  $A^T A$  特征值分解中最大特征值的平方根  $\sqrt{\lambda_1} = \sqrt{3}$ 。

## 矩阵 $F$ -范数

本节介绍的最后一个范数叫**弗罗贝尼乌斯范数** (Frobenius norm), 简称  $F$ -范数, 对应定义为:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2} \quad (75)$$

矩阵  $A$  的  $F$ -范数就是矩阵所有元素的平方和, 再开方。

(69) 给出的矩阵  $A$  有 6 个元素, 计算它们的平方和、再开方就是  $A$  的  $F$ -范数:

$$\|A\|_F = \sqrt{0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2} = \sqrt{4} = 2 \quad (76)$$

本书第 5 章介绍过矩阵  $A$  的所有元素平方和就是  $A$  的格拉姆矩阵的迹, 即:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2} = \sqrt{\text{tr}(A^T A)} \quad (77)$$

根据本书第 13 章介绍过矩阵的迹等于其特征值之和, 这样我们又得到了  $F$ -范数另一个计算方法:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2} = \sqrt{\text{tr}(A^T A)} = \sqrt{\sum_{i=1}^n \lambda_i} \quad (78)$$

其中,  $\sum_{i=1}^n \lambda_i$  为  $A^T A$  的特征值之和。 $A$  的形状为  $m \times n$ , 因此  $A^T A$  的形状为  $n \times n$ 。所以,

$A^T A$  有  $n$  个特征值。一些教材会把  $\sum_{i=1}^n \lambda_i$  求和上限写成  $\min(m, n)$ , 即:

$$\|A\|_F = \sqrt{\sum_{i=1}^{\min(m,n)} \lambda_i} \quad (79)$$

这是因为格拉姆矩阵  $A^T A$  非 0 特征值最多就  $\min(m, n)$ 。如果  $A$  非满秩, 非 0 特征值更少。

(69) 给出的矩阵  $A$  格拉姆矩阵  $A^T A$  有两个特征值 1 和 3, 由此计算  $A$  的  $F$ -范数:

$$\|A\|_F = \sqrt{1+3} = \sqrt{4} = 2 \quad (80)$$

由于,  $A^T A$  的特征值和  $A$  的奇异值存在等式关系  $\lambda_i = s_i^2$ , (78) 还可以写成:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2} = \sqrt{\text{tr}(A^T A)} = \sqrt{\sum_{i=1}^n \lambda_i} = \sqrt{\sum_{i=1}^n s_i^2} \quad (81)$$

对比 (74) 和 (81)，显然矩阵  $A$  的 2-范数不大于  $F$ -范数，即：

$$\|A\|_2 \leq \|A\|_F \quad (82)$$

## 18.9 再谈数据正交投影：优化视角

本章最后从优化视角再谈数据正交投影。

### 正交投影

鸢尾花数据集的前两列构造数据矩阵  $X_{150 \times 2}$ 。给定规范正交基  $V = [v_1, v_2]$ ， $v_1$  和横轴正方向夹角为  $\theta$ 。

如图 19 所示， $X$  在  $v_1$  方向标量投影结果为  $y_1 = Xv_1$ 。 $y_1$  为行数为 150 的列向量， $y_1$  相当于  $X$  在  $v_1$  方向的坐标。

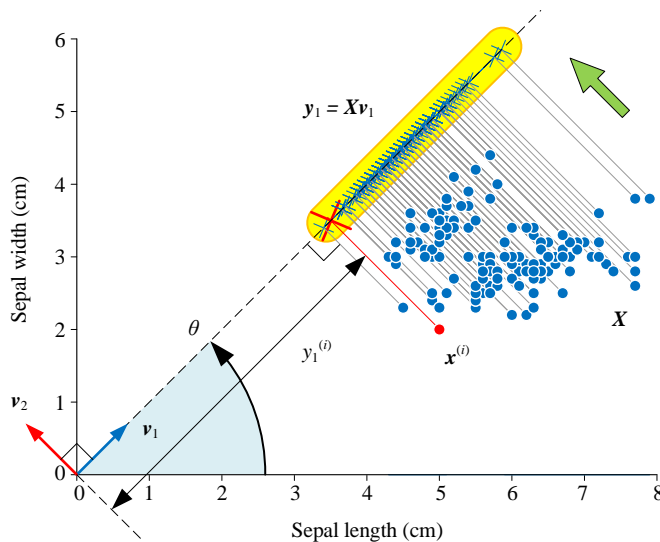
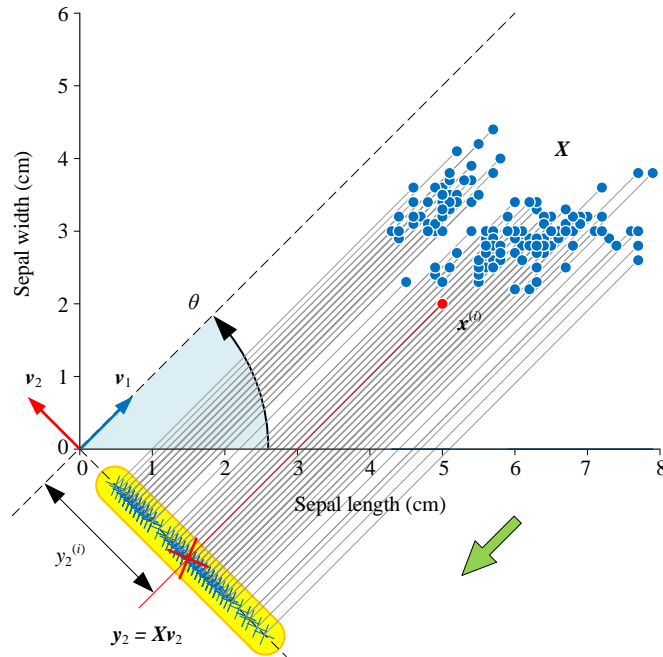


图 19.  $X$  在  $v_1$  上投影

如图 20 所示， $X$  在  $v_2$  方向标量投影结果为  $y_2 = Xv_2$ ， $y_2$  则是  $X$  在  $v_2$  方向的坐标。

图 20.  $X$  在  $v_2$  上投影

### 向量特征、向量之间关系

作为列向量， $y_1$  和  $y_2$  各自有其模 ( $\|y_1\|$ 、 $\|y_2\|$ )，即向量长度。以  $y_1$  为例， $\|y_1\|^2$  写成：

$$\|y_1\|_2^2 = y_1^T y_1 = (Xv_1)^T Xv_1 = v_1^T \underset{G}{X^T X} v_1 = v_1^T G v_1 \quad (83)$$

$y_1$  和  $y_2$  的向量内积 ( $\langle y_1, y_2 \rangle$ )、夹角 ( $\cos(y_1, y_2)$ )、夹角的余弦值 ( $\text{angle}(y_1, y_2)$ ) 可以用来度量  $y_1$  和  $y_2$  之间关系，即：

$$\langle y_1, y_2 \rangle = y_1 \cdot y_2 = y_1^T y_2, \quad \cos(y_1, y_2) = \frac{y_1 \cdot y_2}{\|y_1\| \|y_2\|}, \quad \text{angle}(y_1, y_2) = \arccos\left(\frac{y_1 \cdot y_2}{\|y_1\| \|y_2\|}\right) \quad (84)$$

观察图 19 和图 20，不难发现  $y_1$  和  $y_2$  两个列向量随  $\theta$  变化。也就是说，上述几个量值都会随着  $\theta$  变化。有了变化，就会有最大值、最小值，这就进入了优化视角。

进一步，将  $y_1$  和  $y_2$  写成  $Y = [y_1, y_2] = XV$ ， $Y$  格拉姆矩阵可以写成：

$$G_Y = Y^T Y = (XV)^T XV = V^T \underset{G_X}{X^T X} V = V^T G_X V \quad (85)$$

将  $V = [v_1, v_2]$  代入 (85)，展开得到：

$$G_Y = V^T G_X V = \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} G_X \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} v_1^T G_X v_1 & v_1^T G_X v_2 \\ v_2^T G_X v_1 & v_2^T G_X v_2 \end{bmatrix} \quad (86)$$

这个格拉姆矩阵集成了  $y_1$  和  $y_2$  各自长度 (模)、相互关系 (向量相对夹角) 两方面信息。

## 统计视角

从统计视角来看，如图 21 所示，数据矩阵  $X$  在规范正交基  $[v_1, v_2]$  投影的结果为  $y_1$  和  $y_2$ ，它俩无非就是两列各自含有 150 样本数据的集合。

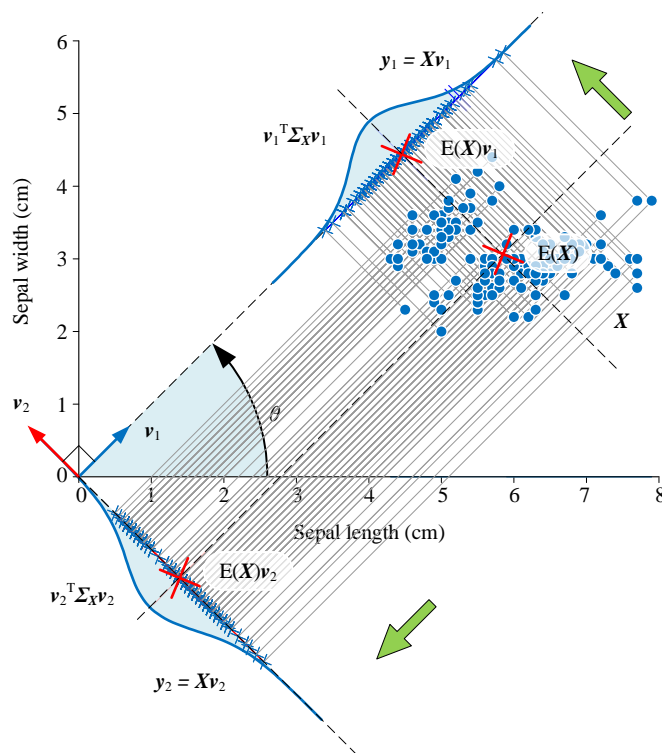


图 21.  $X$  在  $[v_1, v_2]$  上投影，统计视角

$y_1$  和  $y_2$  肯定都有自己统计量，比如均值 ( $E(y_1)$ 、 $E(y_2)$ )、方差 ( $\text{var}(y_1)$ 、 $\text{var}(y_2)$ )、标准差 ( $\text{std}(y_1)$ 、 $\text{std}(y_2)$ )。

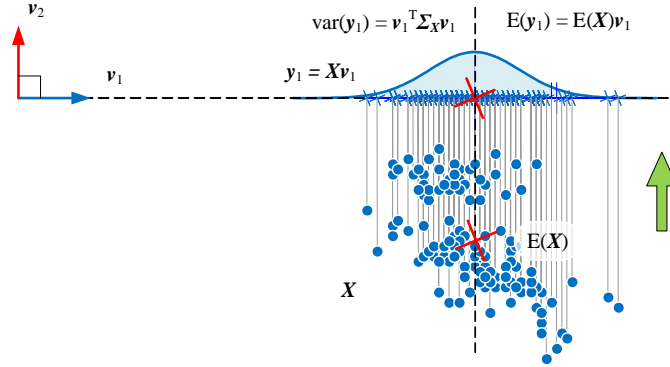
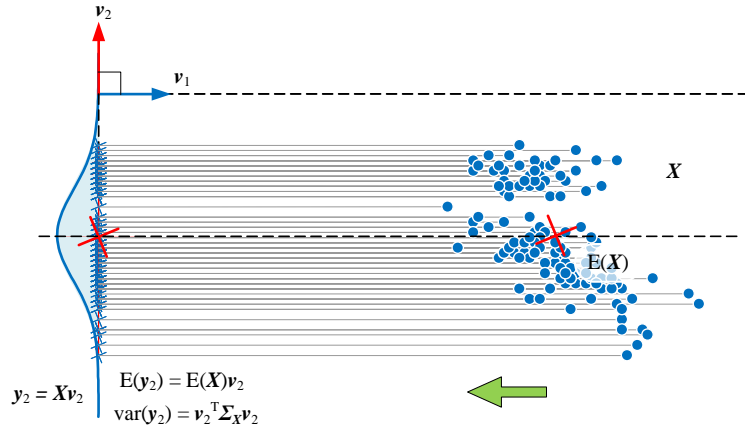
而  $y_1$  和  $y_2$  之间也存在协方差 ( $\text{cov}(y_1, y_2)$ )、相关性系数 ( $\text{corr}(y_1, y_2)$ ) 这两个重要的统计量。

而上述统计度量值同样随着  $\theta$  变化。图 22 和图 23 展示一系列重要统计运算，下面逐个来说。

$y_1$  和  $y_2$  均值 (期望值)  $E(y_1)$  和  $E(y_2)$  为：

$$E(y_1) = E(Xv_1) = E(X)v_1, \quad E(y_2) = E(Xv_2) = E(X)v_2 \quad (87)$$

这相当于数据质心  $E(X) = [E(x_1), E(x_2)]$  分别向  $v_1$  和  $v_2$  投影。

图 22.  $y_1$  的统计特征图 23.  $y_2$  的统计特征

$y_1$  和  $y_2$  的方差  $\text{var}(y_1)$  和  $\text{var}(y_2)$  分别为：

$$\text{var}(y_1) = v_1^T \Sigma_X v_1, \quad \text{var}(y_2) = v_2^T \Sigma_X v_2 \quad (88)$$

其中， $\Sigma_X$  为数据矩阵  $X$  的协方差矩阵。

$y_1$  和  $y_2$  的协方差分别为：

$$\text{cov}(y_1, y_2) = v_1^T \Sigma_X v_2 = \text{cov}(y_2, y_1) = v_2^T \Sigma_X v_1 \quad (89)$$

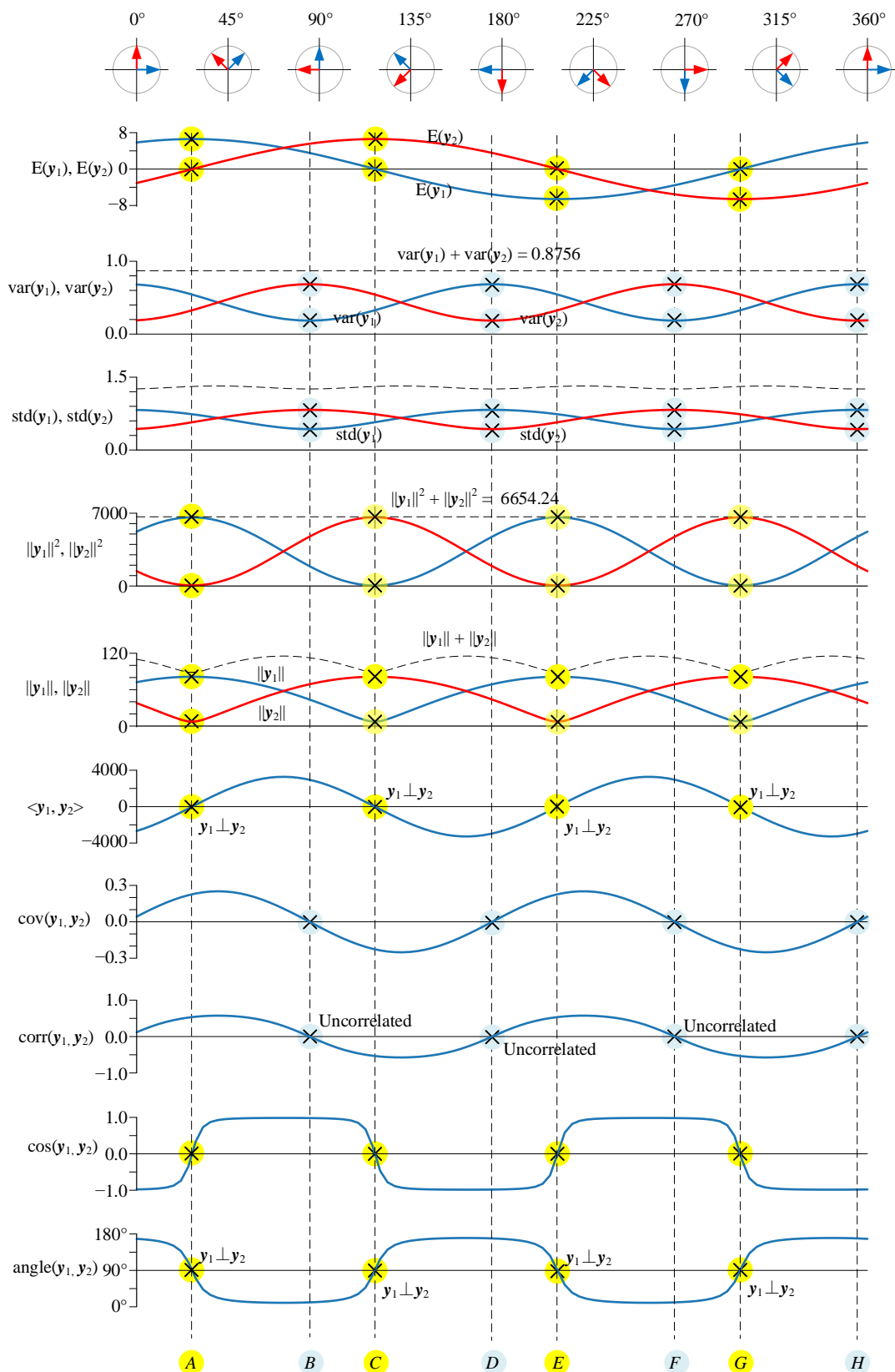
特别地，将  $y_1$  和  $y_2$  写成  $Y = [y_1, y_2]$ ， $Y$  的协方差矩阵可以写成：

$$\Sigma_Y = \begin{bmatrix} \text{var}(y_1) & \text{cov}(y_1, y_2) \\ \text{cov}(y_2, y_1) & \text{var}(y_2) \end{bmatrix} = \begin{bmatrix} v_1^T \Sigma_X v_1 & v_1^T \Sigma_X v_2 \\ v_2^T \Sigma_X v_1 & v_2^T \Sigma_X v_2 \end{bmatrix} = \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} \Sigma_X \begin{bmatrix} v_1 & v_2 \end{bmatrix} = V^T \Sigma_X V \quad (90)$$

比较 (86) 和 (90)，我们发现协方差矩阵和格拉姆矩阵存在大量相似性。本书最后三章和《概率统计》还会继续深入讨论这一话题。

### 优化视角、连续变化

下面，我们用图 24 这展示本节前文介绍的有关  $y_1$  和  $y_2$  各种量化指标随  $\theta$  变化。



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)

图 24.  $y_1$  和  $y_2$  各种量化关系随  $\theta$  变化

请大家注意图 24 中两组  $\theta$  位置  $A$ 、 $C$ 、 $E$ 、 $G$  和  $B$ 、 $D$ 、 $F$ 、 $H$ 。

当  $\theta$  位于  $A$ 、 $C$ 、 $E$ 、 $G$  时， $\|y_1\|^2$  和  $\|y_2\|^2$  取得极值，这四个位置对应  $y_1$  和  $y_2$  垂直，即  $y_1 \perp y_2$ 。

特别值得注意的是，不管  $\theta$  怎么变， $\|y_1\|^2$  和  $\|y_2\|^2$  之和为定值：

$$\|y_1\|_2^2 + \|y_2\|_2^2 = y_1^T y_1 + y_2^T y_2 = 6654.24 = \|x_1\|_2^2 + \|x_2\|_2^2 = x_1^T x_1 + x_2^T x_2 \quad (91)$$

这是因为矩阵迹的重要性质—— $\text{tr}(AB) = \text{tr}(BA)$ ，即：

$$\text{tr}(G_Y) = \text{tr}(V^T G_X V) = \text{tr}((V^T G_X) V) = \text{tr}\left(V V^T G_X\right) = \text{tr}(G_X) \quad (92)$$

$G_Y$  的迹为：

$$\text{tr}\left(\underbrace{\begin{bmatrix} y_1^T y_1 & y_1^T y_2 \\ y_2^T y_1 & y_2^T y_2 \end{bmatrix}}_{G_Y}\right) = y_1^T y_1 + y_2^T y_2 = \|y_1\|_2^2 + \|y_2\|_2^2 \quad (93)$$

而  $G_X$  的迹为：

$$\text{tr}\left(\underbrace{\begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{bmatrix}}_{G_X}\right) = x_1^T x_1 + x_2^T x_2 = \|x_1\|_2^2 + \|x_2\|_2^2 \quad (94)$$

特别地，如果 (92) 中  $V$  来自于特征值分解，则 (93) 等于  $G_X$  的两个特征值之和：

$$y_1^T y_1 + y_2^T y_2 = x_1^T x_1 + x_2^T x_2 = \lambda_1 + \lambda_2 \quad (95)$$

当  $\theta$  位于  $B$ 、 $D$ 、 $F$ 、 $H$  时， $\text{var}(y_1)$  和  $\text{var}(y_2)$  取得极值，对应  $y_1$  和  $y_2$  线性无关，即相关性系数为 0，不同于  $y_1 \perp y_2$ 。

同样值得注意的是，不管  $\theta$  怎么变， $\text{var}(y_1)$  和  $\text{var}(y_2)$  之和为定值：

$$\text{var}(y_1) + \text{var}(y_2) = 0.8756 \quad (96)$$

利用迹运算，同样得出类似结论，

$$\text{tr}(\Sigma_Y) = \text{tr}(V^T \Sigma_X V) = \text{tr}((V^T \Sigma_X) V) = \text{tr}\left(V V^T \Sigma_X\right) = \text{tr}(\Sigma_X) \quad (97)$$

$\Sigma_Y$  的迹为：



$$\text{tr} \left[ \underbrace{\begin{bmatrix} \text{var}(\mathbf{y}_1) & \text{cov}(\mathbf{y}_1, \mathbf{y}_2) \\ \text{cov}(\mathbf{y}_2, \mathbf{y}_1) & \text{var}(\mathbf{y}_2) \end{bmatrix}}_{G_y} \right] = \text{var}(\mathbf{y}_1) + \text{var}(\mathbf{y}_2) \quad (98)$$

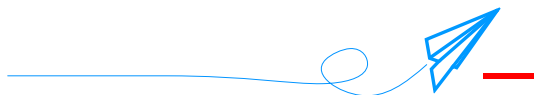
而  $\Sigma_x$  的迹为：

$$\text{tr} \left[ \underbrace{\begin{bmatrix} \text{var}(\mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{var}(\mathbf{x}_2) \end{bmatrix}}_{G_x} \right] = \text{var}(\mathbf{x}_1) + \text{var}(\mathbf{x}_2) \quad (99)$$

也就是说：

$$\text{var}(\mathbf{y}_1) + \text{var}(\mathbf{y}_2) = \text{var}(\mathbf{x}_1) + \text{var}(\mathbf{x}_2) = 0.8756 \quad (100)$$

这一点非常重要，大家将会在主成分分析看到它的应用。



约束条件影响优化问题解的位置。拉格朗日乘法可以把有约束优化问题转化为无约束优化问题。本章分别从等式约束和不等式约束两方面来展开。需要大家格外注意的是，如何利用梯度向量理解拉格朗日乘法？此外，对于不等式约束，KKT 条件中每个式子背后的数学思想是什么？

本章又从优化视角深入讨论了特征值分解、SVD 分解。请大家特别注意，SVD 分解中，分解对象可以分别为原始数据矩阵、中心化数据矩阵，甚至是  $z$  分数。它们的 SVD 分解结果有着很大差异。本书最后还会深入探讨，请大家留意。

本章最后从优化视角回顾了数据正交投影，建立了向量和统计描述之间的关系，这是本书最后四章要涉及的话题。

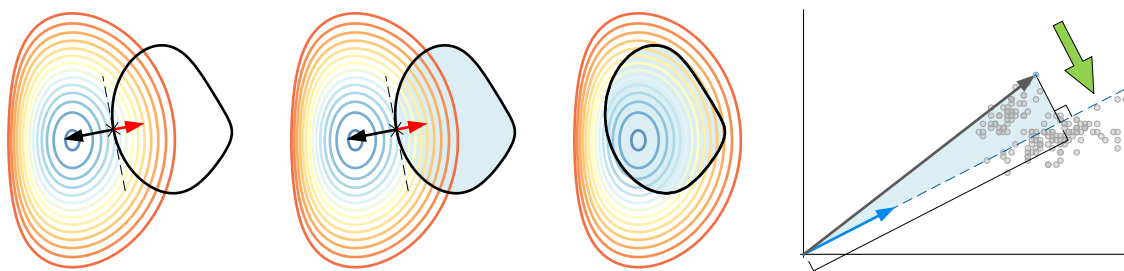


图 25. 总结本章重要内容的四副图