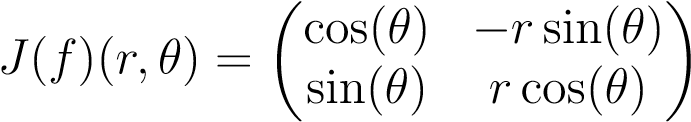
这个行列式实际上只依赖于df（a），而不是特定的基。然而，偏导数给出了计算它的方法。

当e=rn，f=rm时，对于任何函数f:rn→rm，都很容易计算偏导数）。我们简单地把函数fi:rn→r当作它的j-次参数的函数，剩下的不变，然后按照定义38.1计算导数，也就是通常的导数。

例38.2。例如，考虑函数f:r2→r2，定义如下：

f（r，θ）=（r cos（θ），r sin（θ））。

那么，我们有了



雅可比（行列式）的值为Det（j（f）（r，θ））=r。

如果e=r（或e=c），对于任何函数f:r→f（或f:c→f），则

df（a）的雅可比矩阵是列向量。实际上，这个列向量只是d1f（a）。然后，对于每一个λ∈r（或λ∈c），

df（a）（λ）=λd1f（a）。

这一情况非常重要，足以证明有一个定义。

定义38.5.给定函数f:r→f（或f:c→f），其中f是赋范仿射空间，向量

df（a）（1）=d1f（a）

我们通常用d1f（a）的雅可比矩阵d1f（a）来识别df（a），它是d1f（a）对应的列向量。通过滥用符号，我们还让df（a）表示向量df（a）（1）=d1f（a）。

当e=r时，物理解释为f定义了一条（参数）曲线，它是一些粒子在rm中运动的轨迹，是时间的函数，向量d1f（a）是运动粒子在t=a时的速度f（t）。

把函数f[a，b]→f从一个闭区间[a，b]r考虑到一个赋范仿射空间f，及其在[a，b]上的导数df（a），通常是有用的，即使[a，b]没有打开。在这种情况下，与实值函数一样，我们在a处定义右导数d1f（a+），在b处定义左导数d1f（b-），并假定它们存在。

例38.3。

1. 当a=（0,1）和f=r3时，函数f:（0,1）→r3定义r3中的（参数）曲线。如果f=（f1，f2，f3），其雅可比矩阵在a∈r处为

.

见图38.3。

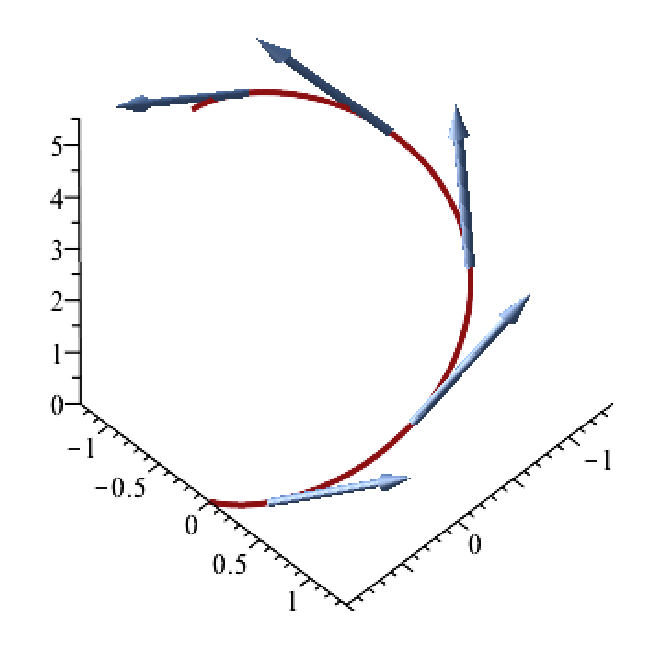


图38.3：红色空间曲线f（t）=（cos（t），sin（t），t）。

速度矢量用蓝色箭头表示。

1. 当e=r2和f=r3时，函数\_:r2→r3定义参数曲面。式中：\_=（f，g，h），其雅可比矩阵a∈r2为

.

见图38.4。雅可比矩阵是。第一列是

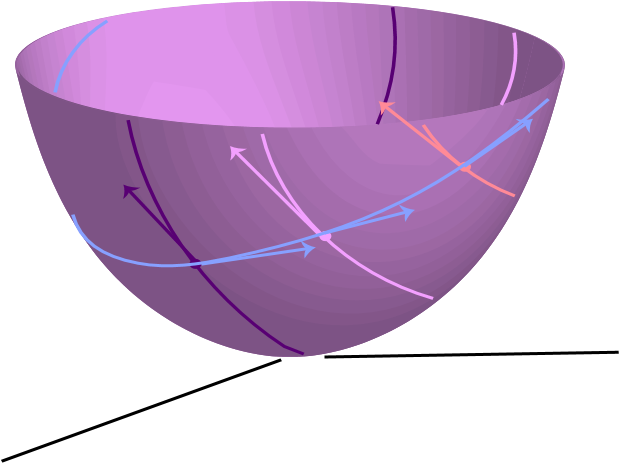


图38.4：参数曲面x=u，y=v，z=u2+v2。

向量与粉红色的U方向曲线相切，而第二列是与蓝色的V方向曲线相切的向量。

1. 当e=r3，f=r时，对于函数f:r3→r，a∈r3处的雅可比矩阵为

.

一般来说，当f:rn→r时，a∈rn处的雅可比矩阵是行向量。

.

它的转置是一个称为f在a处的梯度的列向量，用梯度f（a）或f（a）表示。那么，对于任何v∈rn，注意

D

梯度f（a）和v的标量积。

例38.4。考虑二次函数f:rn→r，由

f（x）=x>ax，x∈rn，

其中a是实n×n对称矩阵。我们声称

dfu（h）=2u>ah，表示所有u，h∈rn。

因为a是对称的，我们有

F（U+H）=（U>+H>）A（U+H）

=u>au+u>ah+h>au+h>ah=u>au+2u>ah+h>ah，

所以如果我们写的话

对于h/∈0，其中k k是2-范数，根据柯西-施瓦兹，我们有

，

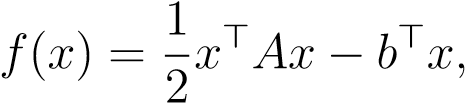
这表明lim=0。因此，

dfu（h）=2u>ah，对于所有u，h∈rn，

如要求。该公式表明，f在u处的梯度fu由下式得出：

fu=2au.

作为第一个推论，我们得到了形式函数的梯度。



其中a是对称的n×n矩阵，b是某个向量b∈rn。由于线性函数的导数是它本身，我们得到

dfu（h）=u>ah−b>h，

f的梯度由

fu=au−b.

作为第二个推论，我们得到了函数的梯度。



这是最小二乘问题中要最小化的函数，其中a是m×n矩阵。

我们有f（x）=x>a>ax−x>a>b−b>ax+b>b=x>a>ax−2b>ax+b>b，

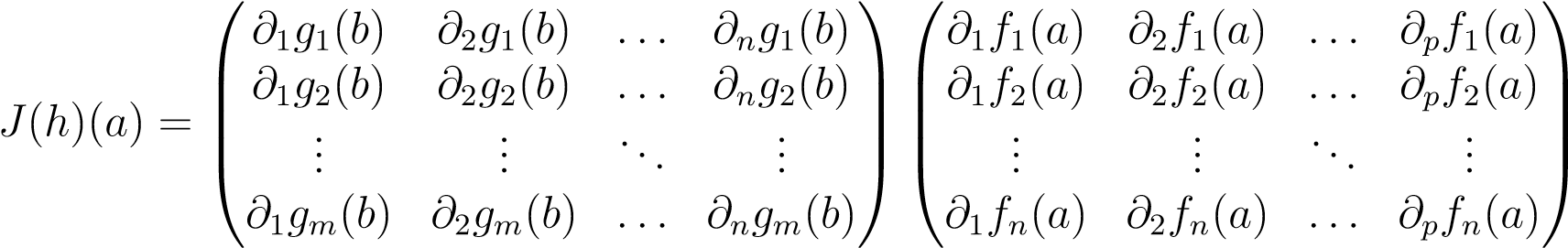
因为常数函数的导数是0，线性函数的导数是它本身，我们得到

dfu（h）=2u>a>ah−2b>ah。

因此，f的梯度由

fu=2a>au−2a>b。

当e、f和g有有限维时，（a0，（u1，…，up））是e的仿射框架，（b0，（v1，…，vn））是f的仿射框架，（c0，（w1，…，wm））是g的仿射框架，如果a是e的开放子集，b是f的开放子集，对于任何函数f:a→f和g:b→g，这样f（a）b，对于任意a∈a，假设b=f（a），h=g\_f，如果df（a）存在且dg（b）存在，根据定理38.6，雅可比矩阵j（h）（a）=j（g\_f）（a）w.r.t.基（u1，…，up）和（w1，…，wm）是雅可比矩阵j（g）（b）w.r.t.基（v1，…，vn）和（w1，…，wm）和j（f）（a）W.R.T.底座（U1，…，向上）和（V1，…，VN）：



或

.

因此，我们有了熟悉的公式

.

给定有限维的两个赋范仿射空间e和f，给定e的开子集a，如果函数f:a→f在a∈a上可微，则其雅可比矩阵是定义良好的。

应该警告一个人，相反的说法是错误的。有这样的函数，所有的偏导数都存在于某个a∈a上，但是函数在a上是不可微的，

甚至不连续。例如，考虑函数f:r2→r，定义为f（0,0）=0，以及

如果（x，y）=（06，0）。

对于任何u=06，我们有

，

以便

D

因此，所有u=06都存在duf（0,0）。另一方面，如果存在df（0,0），它将是一个线性映射df（0,0）：r2→r，由一个行矩阵（αβ）表示，我们将得到duf（0,0）=df（0,0）（u）=αh+βk，但duf（0,0）的显式公式不是线性的。事实上，函数f在（0,0）不是连续的。例如，在抛物线上

当我们接近抛物线的原点时，极限是

实际上，f（0,0）=0。

然而，对于df（a）存在的偏导数有充分的条件，即偏导数的连续性。

如果f在a上是可微的，那么f定义了函数d）。结果表明，a上偏导数的连续性是df在a上存在和连续的一个充要条件。

如果f:[a，b]→r是一个函数，它在[a，b]上是连续的，在[a，b]上是可微的，那么有一些c的a<c<b，这样

F（B）−F（A）=（B−A）f0（C）。

这个结果被称为中值定理，是罗尔定理的推广，它对应于f（a）=f（b）的情况。

不幸的是，向量值函数的中值定理失败了。例如，函数f:[0,2π]→r2由

F（t）=（成本，SINT）

f（2π）−f（0）=（0,0），但其导数f0（t）=（−sint，cost）不会在（0,2π）中消失。

如果我们考虑一个不等式（上界）而不是一个等式，中值定理对向量值函数的适当推广是可能的。这个中值定理的推广版本在证明微分学的几个主要结果中起着重要作用。

如果e是仿射空间（在r或c上），给定任意两点a，b∈e，闭段[a，b]是所有点a+λ（b−a）的集合，其中0≤λ≤1，λ∈r，而开段（a，b）是所有点a+λ（b−a）的集合，其中0<λ<1，λ∈r。

引理38.12。设e和f为两个赋范仿射空间，设a为a上的开子集，设f:a→f为a上的连续函数，给定e中任意a∈a和任意h=06，如果闭段[a，a+h]包含在a中，如果f:a→f在开段（a，a+h）的每一点上都是可微的，并且

sup kdf（x）k≤m，

X∈（A，A+H）

对于某些m≥0，则

k f（a+h）−f（a）k≤mkhk。

作为推论，如果L：→−E→→−F是一个连续的线性映射，那么

k f（a+h）−f（a）−l（h）k≤mkhk，

其中m=supx∈（a，a+h）kdf（x）−lk。

上面的引理有时被称为“中值定理”，引理38.12可以用来表示下面的重要结果。

定理38.13。给定两个赋范仿射空间e和f，其中e是有限维n，其中（a0，（u1，…，un））是e的一个框架，给定e的任何开子集a，给定任何函数f:a→f，导数d在iff上定义并连续，所有的偏导数）在a上定义并连续。j，1≤j≤n。

作为推论，如果是有限维m，并且（b0，（v1，…，vm））是f的框架，那么导数df:a→l（e；f）在iff上定义并连续，每个偏导数在a上定义并连续，对于所有i，j，1≤i≤m，1≤j≤n。

定理38.13给出了开集上函数导数的存在性和连续性的充分必要条件。应当指出，定理38.13的更一般的版本成立，假设e=（e1，a1）·······（en，an），或e=e1×······×en，并使用在命题38.11之前引入的更一般的偏导数djf。

定义38.6.给定两个赋范仿射空间e和f，以及e的开子集a，我们认为函数f:a→f是a上的类c0，或是if f上的c0函数是a上的连续函数，我们认为f:a→f是a上的类c1，或是a上的c1函数，并且是a上的连续函数。

38.3。隐函数和反函数定理

由于开集上导数的存在意味着连续性，所以C1函数当然是一个C0函数。定理38.13给出了函数f为C1函数（当e为有限维时）的一个必要和充分条件。很容易证明C1函数的组成（在适当的开集上）是C1函数。

# 38.3隐函数和反函数定理

给定三个赋范仿射空间e、f和g，给定函数f:e×f→g，给定任意c∈g，方程可能发生

f（x，y）=c

具有这样的性质：对于某些开放集a e和b f，有一个函数g:a→b，这样f（x，g（x））=c，

对于所有的x∈a，这样的情况通常是非常罕见的，但如果某个解（a，b）∈e×f使f（a，b）=c已知，在一定条件下，对于某些小的开集a e含有a和b f含有b，存在唯一的g:a→b，这样

f（x，g（x））=c，

对于所有x∈a，都可以显示。在一定条件下，还可以证明G是连续的，且是可微的。这样一个定理，称为隐函数定理，可以证明。根据Schwartz[147]的规定，我们在下面陈述了该结果的一个版本。证明（见Schwartz[147]）相当复杂，并使用不动点定理在完全度量空间中收缩映射。其他证据见lang[108]和Cartan[34]。

定理38.14。设e，f，和g，为赋范仿射空间，设Ω为e×f的开子集，设f:Ω→g为定义在Ω上的函数，设（a，b）∈Ω，设c∈g，并假定f（a，b）=c。如果下列假设成立

1. 功能F：Ω→G在Ω上是连续的；
2. f是一个完全赋范仿射空间（g也是）；

存在于每（x，y）∈Ω，并且是连续的；

是，和的双射体；

然后以下属性保持不变：

1. 存在一些含有a的开子集a e和一些含有b的开子集b f，这样a×bΩ，并且对于每个x∈a，方程f（x，y）=c有一个单解y=g（x），因此，有一个唯一的函数g:a→b，这样f（x，g（x））=c，对于所有x∈a；
2. 功能G:A→B是连续的。

如果我们也假设

（5）导数df（a，b）存在；然后

1. 衍生的dg（a）存在，并且

d）；

如果另外

也是连续的（因此，考虑到（3），F是Ω上的C1）；

然后

1. 衍生的dg:a→l（→−e；→−f）是连续的，并且

D

对于所有x∈a。

隐函数定理在变分法中起着重要作用。我们现在考虑另一个非常重要的概念，即（局部）差异同构。

定义38.7.给定两个拓扑空间e和f，以及e的开子集a，我们认为函数f:a→f是a到f的局部同态，如果对于每一个a∈a，都有一个开集u a包含a和一个开集v包含f（a），这样f是u到v=f（u）的同态。如果b是f的开放子集，我们说f:a→f是a到b的（全局）同态，如果f是a到b=f（a）的同态。如果e和f是赋范仿射空间，我们说f:a→f是a到f的局部差分形式，如果对于每个a∈a，都有一个包含a的开集u a和一个包含f（a）的开集v，这样f是u到v的双射，f是u上的c1函数，f−1是v=f上的c1函数。（U）。我们说f:a→f是A到B的一个（全局）不同同态，如果f是A到B=f（a）的同态，f是A上的C1函数，f−1是B上的C1函数。

请注意，局部差异同构是局部同胚。另外，由于命题38.8，如果f是a上的差同态，那么df（a）是每个a∈a的线性同构，可以证明以下定理。事实上，使用定理38.14有一个相当简单的证明；见Schwartz[147]、Lang[108]、Cartan[34]和Abraham和Marsden[1]。

38.3。隐函数和反函数定理

定理38.15。设e和f为完全赋范仿射空间，设a为e的开子集，设f:a→f为a上的c1函数，其性质如下：

1. 对于每一个a∈a，如果df（a）是一个线性同构（这意味着df（a）和（df（a））-1都是线性的和连续的），那么存在一些包含a的开子集u a和一些包含f（a）的f的开子集v，这样f是从u到v=f（u）的一个差异。此外，

df−1（f（a））=（df（a））−1.

对于a的每个邻域n，其图像f（n）是f（a）的邻域，对于a中心的每个开球u a，其图像f（u）包含一些中心f（a）的开球。

1. 如果df（a）对每一个a∈a都可逆，那么b=f（a）是f的开子集，f是a到b的局部差分同态。此外，如果f是内射的，那么f是a到b的差分同态。

定理38.15的（1）部分常被称为“（局部）反函数定理”，它在流形和（一般）微分方程的研究中起着重要作用。

如果e和f都是有限维的，并且选择了一些帧，则df（a）的可逆性等于雅可比行列式det（j（f）（a））不为空的事实。df（a）只是内射的或者只是Surjective的情况对于使用隐式定义定义定义流形也很重要。

定义38.8.设e和f是赋范仿射空间，其中e和f是有限维（或e和f都是完备的），a是e的开子集。对于任何a∈a，c1函数f:a→f是在a if df（a）上的浸入。c1函数f:a→f是中频df（a）上的一个浸没。C1函数f:a→f是对a（resp.如果df（a）是可注射的（resp.（主观性）对于每一个a∈a。

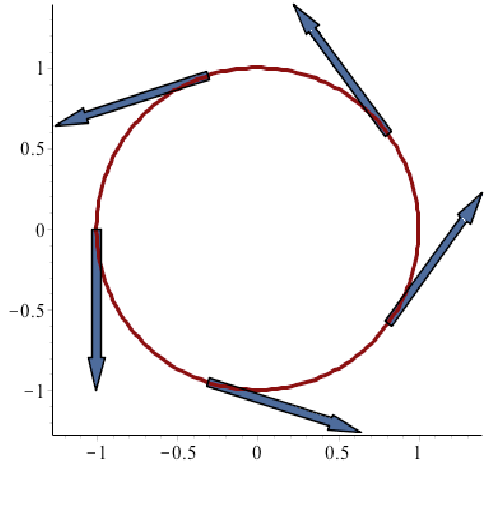
当e和f是Dim（e）=n和Dim（f）=m的有限维时，如果m≥n，则f是一个浸入式iff，雅可比矩阵j（f）（a）对所有a∈e都具有满秩n，如果n≥m，则f是雅可比矩阵j（f）（a）的一个子形式iff，对于所有a∈e都具有满秩m。例如，f:r→r2由f（t）=（cos（t），sin（t））定义的是浸入式，因为

所有t的排名为1。另一方面，f:r→r2由

f（t）=（t2，t2）不是浸没，因为t=0时消失。见图38.5。

投影图f:r2→r给出了一个淹没的例子，其中f（x，y）=x，因为。

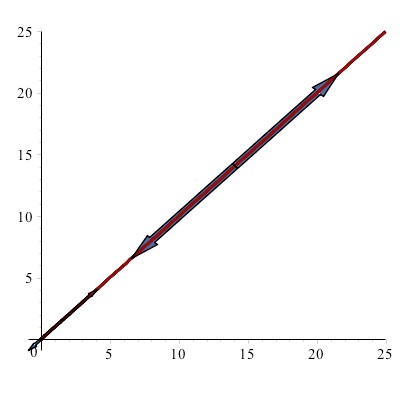
可以显示以下结果。



(

i.

)



(

ii.

)

图38.5：图（i）是将r浸入由f（t）=（cos（t），sin（t））给出的r2中。图（ii.），参数曲线f（t）=（t2，t2），不是浸入，因为切线在原点处消失。

提案38.16。设a为rn的开放子集，设f:a→rm为函数。对于每一个a∈a，f:a→rm是iff处的一个潜水器，存在一个包含a的a的开放子集u，一个开放子集w rn−m，以及一个不同的同构m\_：u→f（u）×w，这样，

F=π1\_

其中π1:f（u）×w→f（u）是第一个投影。等价地，（f\_−1）（y1，…，ym，…，yn）=（y1，…，ym）。

γ

U a/f（U）×W

nnnnn和π1

F（U）令吉

此外，f下的每个开放子集的图像都是f的开放子集（cn和cm的结果相同）。

提案38.17。设a为rn的开放子集，设f:a→rm为函数。

对于每一个a∈a，f:a→rm是一个浸入式iff，存在一个开放的u子集。

38.4。切线空间和微分

a包含a，包含f（a）的开子集v，其中f（u）v，包含0的开子集w，其中w rm−n，以及差分异构体，其中：

\_f=in1，

其中in1:u→u×w是注入图，使得in1（u）=（u，0），或等效地，（\_f）（x1，…，xn）=（x1，…，xn，0，…，0）。

f

u a/f（u）v

mmminmm1mmmm&\_

U×W

（相同的结果适用于cn和cm）。

# 38.4切线空间和微分

在本节中，我们将简要讨论导数概念的几何解释。我们考虑由可微函数定义的点集。这是（微分）流形概念的一个特例。

给定两个赋范仿射空间e和f，设a为e的开子集，设f:a→f为函数。

定义38.9.给定f:a→f，其图（f）是所有点（f）=（x，y）∈e×f x∈a，y=f（x）的集合。

如果在a上定义df，我们就说\_（f）是方程y=f（x）的e×f的微分子流形。

应该注意，这是一种非常特殊的微分流形。

例38.5。如果e=r，f=r2，则f=（g，h），其中g:r→r和h:r→r，

（f）是方程式y=g（x），z=h（x）的r3曲线。当e=r2且f=r时，\_（f）是式z=f（x，y）的r3中的曲面。

我们现在以一种非常普遍的方式定义仿射切空间的概念。接下来，我们将看到它对流形\_（f）的意义，如定义38.9所示。

定义38.10.给定一个范数仿射空间e，给定e的任意非空子集m，给定任意点a∈m，如果m中存在一个点的序列（a n）n∈n收敛到a，且一个序列（λn）n∈n，且λi∈r和λn≥0，则称向量u∈→−e在a到m处相切，从而使该序列（λn（a−a））n∈n收敛于u。

在a到m处相切的所有向量的集合称为a到m处的相切向量族，在a到m处，u属于相切向量族的形式a+u的e的所有点的集合称为a到m处的仿射相切族。

显然，0总是相切的，如果u是相切的，那么每个λu也是相切的，对于λ∈r，λ≥0。如果u=06，则序列（λn）n∈n必须趋向于+∞。我们有以下建议。

提案38.18。设e和f为两个赋范仿射空间，设a为e的开子集，设a∈a，设f:a→f为函数。如果存在→−×d→−f（a），则（a，f（a））到\_的切向量族是e f的子空间ta（），由条件（方程式）定义。

（u，v）∈ta（）iff v=df（a）（u），

（a，f（a））到\_的仿射切线族是e×f的仿射变体ta（\_），由条件（方程式）定义。

（x，y）∈ta（）iff y=f（a）+df（a）（x−a），

式中，\_是f的图形。

证据其实相当简单。我们有Ta（）=A+Ta（），并且由于Ta（）是→−E×→−F的子空间，所以集合Ta（）是仿射变体。因此，点（a，f（a））上的仿射切线空间是族对象、直线、平面等。

举例来说，当e=r2和f=r时，方程式z=f（x，y）表面（a，b，c）点处的仿射切线平面由方程式定义。

.

如果e=r和f=r2，方程式y=g（x），z=h（x）曲线的（a，b，c）处的切线由方程式定义。

.

因此，导数和偏导数作为切线空间具有理想的几何解释。当然，为了正确地处理这个问题，我们真的需要更深入地研究（微分）流形。

我们现在简单地考虑二阶和高阶导数。

# 38.5二阶及更高阶导数

给出了两个赋范仿射空间e和f，以及e的一些开子集a，如果对每一个a∈a定义了df（a），那么我们就得到了一个映射df:a→l（→−e；→−f）。由于L（→−e；→−f）是赋范向量空间，如果df存在于含有a的a的开子集u上，我们可以考虑取df在a∈a处的导数，如果d（df）（a）存在于每个a∈a，我们得到一个映射

38.5。二阶和高阶导数

d）），其中d2f（a）=d（df）（a），对于每个a∈a。如果d2f（a）存在，则对于每个u∈e，

d.

从36.61号提案中回忆起，从L（→−e；→−f）×到→−f的地图应用程序定义为，对于每一个L∈l（→−e；→−f），对于每一个V∈→−e，

约（l，v）=l（v）

是一个连续双线性映射。因此，特别是给定一个固定的→→→→→−v∈→−e，线性映射appv:l（e；f）f，定义为appv（l）=l（v）是一个连续映射。

还记得，从命题38.7，如果h:a→g是一个函数，使得dh（a）存在，k:g→h是一个连续的线性映射，那么，d（k\_h）（a）存在，并且

k（d h（a）（u））=d（k\_h）（a）（u），

即k（du h（a））=du（k\_h）（a），

将这两个事实应用于h=df和k=appv，我们得到

du（df）（a）（v）=du（appv df）（a）。

但是（appv df）（x）=df（x）（v）=dvf（x），对于每x∈a，也就是说，appv df=dvf在a上。

所以，我们有

du（df）（a）（v）=du（dvf）（a）

由于d2f（a）（u）=du（df）（a），我们得到

d2f（a）（u）（v）=du（dvf）（a）。

因此，当d2f（a）存在时，du（dvf）（a）存在，并且

d2f（a）（u）（v）=du（dvf）（a），

对于所有u，v∈→−e，我们也用d2u、vf（a）或du dvf（a）表示du（dvf）（a）。

从36.60号提案中回顾，从l2（→−e，→−e；→−f）到l（→−e；l（→−e；→−f））的映射定义为，g 7→Фiff代表每个g∈l2（→−e，→−e；→−f），

⑨（u）（v）=g（u，v）

是向量空间的同构，因此，我们将d→→→−2f（a）∈l（→−e；l（→−e；→−f））看作2（e，e；f）中的连续双线性映射，我们将写d2f（a）（u，v），而不是d2f（a）（u）（v）。

然后，上述讨论可以总结为当定义d2f（a）时，我们有

d2f（a）（u，v）=dudvf（a）。

当e有有限维且（a0，（e1，…，en））是e的框架时，我们表示dejdeif（a）

by），当i=6J时，我们表示d

利用引理38.12，可以显示出施瓦兹的以下重要引理。给出双线性映射f：→−e×→−e→→−f，回想一下f是对称的，如果

f（u，v）=f（v，u）

对于所有u，v∈→−e。

引理38.19。（施瓦兹引理）给定两个赋范仿射空间e和f，给定e的任意开子集→→→→→–−a，给定任意f:a→f，对于每个a∈a，如果d2f（a）存在，那么d2f（a）∈

l2（e，e；f）是一个连续对称双线性映射。作为推论，如果e是有限维n，（a0，（e1，…，en））是e的框架，我们有

.

注：上述引理有一个变化，它不假定d2f（a）的存在，而是假定dudvf和dvduf存在于包含a的开子集上，并且在a处是连续的，因此dudvf（a）=dvduf（a）。这只是一个不同的结果，并不意味着引理38.19，也不是引理38.19的结果。

当e=r2时，只有）和）不足以保证d2f（a）的存在。

当有限维n和（a0，（e1，…，en））的e是e的一个框架时，如果d2f（a）存在，对于→−e中的每个u=u1e1+······+unen和v=v1e1+·······+vnen，因为d2f（a）是对称双线性形式，我们有

D

可以用矩阵形式写为：

D

38.5。二阶和高阶导数

其中u是表示u的列矩阵，v是表示v的列矩阵，在帧上（a0，（e1，…，en））。

上述对称矩阵在a称为f的Hessian。如果f本身是有限维的，并且（b0，（v1，…，vm））是f的帧，那么f=（f1，…，fm）和每个分量

d2f（a）（u，v）（1≤i≤m）的d2f（a）i（u，v）可写为

D

因此，我们可以用m个对角块构成的mn×mn矩阵（即上述Hessians）和m次的行矩阵（u>、…、u>）和m次的列矩阵来描述d2f（a）（u，v）。

我们现在简要地说明如何定义高阶导数。设m≥2。给定函数f:a→f，对于任意a∈a，如果导数d if存在于a上，对于所有i，1≤i≤→lm−1，通过归纳，d−−−−→→−m−1f可以被视为连续函数dm−1f:a m−1（em−1；f），并且我们定义

dmf（a）=d（dm−1f）（a）。

然后，可以用lm中的连续m-多行图（e−→m；→−f）来识别dmf（a）。然后我们可以证明（正如我们之前所做的），如果定义了dmf（a），那么dmf（a）（u1，…，um）=du1…dumf（a）。

当有限维数n和（a0，（e1，…，en））的e为e的框架时，如果存在dmf（a），对于每个j1，…，jm∈1，…，n，我们用表示dejm…dej1f（a）

.

给定m-多行映射f∈lm（e−→m；→−f），回想一下f是对称的if

f（uπ（1），…，uπ（m））=f（u1，…，um）

对于所有的u1，…，um∈→−e，和所有的置换πon 1，…，m。接下来，施瓦兹引理的推广成立。

引理38.20。给出了两个赋范仿射空间e和f，给出了e的任意开子集a，给出了任意f:a→f，对于每个a∈a，对于每个m≥1，如果dmf（a）存在，那么dmf（a）∈lm（e−→m；→−f）是连续对称的m-多线性映射。作为推论，如果e是有限维n，（a0，（e1，…，en））是e的框架，我们有

，

对于每个j1，…，jm 1，…，n，对于每个置换πon 1，…，m。

如果e是有限维n，并且（a0，（e1，…，en））是e的帧，那么dmf（a）是对称m-多行映射，我们有

D

其中，对于任意m向量，j在所有函数j：1，…，m→1，…，n范围内。

uj=uj，1e1+····+uj，nen.

将C1函数的概念推广到cm函数的概念，并将定理38.13推广。

定义38.11.给定两个赋范仿射空间e和f，以及e的开子集a，对于任何m≥1，我们称函数f:a→f在a或cm函数上是类cm，如果dkf存在，并且在a上是连续的，对于每k，1≤k≤m。我们称f:a→f在a或c∞-函数上是类c∞。dkf存在，并且每k≥1在a上连续。c∞-函数（在a上）也称为光滑函数（在a上）。a和b之间的cm差分形式f:a→b（其中a是e的开放子集，b是b的开放子集）是a和b=f（a）之间的双射，因此f:a→b及其逆f−1:b→a都是cm函数。

等价地，f是a上的cm函数，if f是a上的c1函数，df是a上的cm−1函数。

我们有以下定理，给出了f对a上的cm函数的一个充分必要的条件。对e=（e1，a1）···（en，an）也成立的情况的推广。

定理38.21。给定两个赋范仿射空间e和f，其中e是有限维n，其中（a0，（u1，…，un））是e的框架，给定e的任何开子集a，给定任何函数，导数dmf是iff上的cm函数，每个偏导数d）在a上定义并连续，对于所有k，1≤k≤m，所有的j1，…，jk∈1，…，n。作为推论，如果f是有限维p，

38.6。泰勒公式，法阿·迪布鲁诺公式`

并且（b0，（v1，…，vp））是f的一个框架，导数dmf在iff上定义并连续。

K

每个偏导数d）在a上定义并连续，对于所有k，1≤k≤m，对于所有i，1≤i≤p，以及所有j1，…，jk∈1，…，n。

当e=r（或e=c）时，对于任意a∈e，dmf（a）（1，…，1）是→−f中的向量，称为mth阶向量导数。在m=1的情况下，我们通常用向量dmf（a）（1，…，1）来标识多行映射dmf（a）。还可以引入一些符号约定来简化高阶导数的符号，并对这些约定进行了简要的讨论。

回想一下，当e是有限维n，并且（a0，（e1，…，en））是e的帧时，dmf（a）是对称的m-多行映射，我们有

D

J 1

其中，对于任意m向量，j在所有函数j：1，…，m→1，…，n范围内。

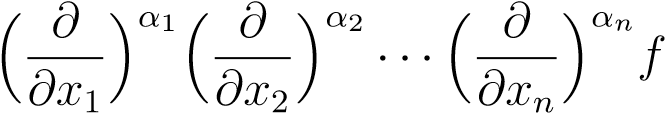
uj=uj，1e1+····+uj，nen.

然后我们可以将与相同变量xjk对应的xjk的各种出现进行分组，从而得出符号

，

式中α1+α2+····+αn=m。

如果我们简单地用α表示（α1，…，αn），那么我们表示



通过

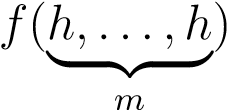
αf，或

如果α=（α1，…，αn），我们让α=α1+α2+····+αn，α！=α1！···αn！，如果h=（h1，…，hn），我们表示。

在下一节中，我们将调查泰勒公式的各种版本。

# 38.6泰勒公式，法阿迪布鲁诺公式

我们讨论了泰勒公式的几个版本，没有证据。每个版本所需的假设越来越强。第一个版本可以看作是导数概念的概括。给出了一个m-线性映射f:e−→m→→−f，对于任意向量h∈→−e，我们将其缩写为



按F（hm）。下面给出的泰勒公式的版本有时被称为泰勒-杨公式。

定理38.22。（Taylor–Young）给定两个赋范仿射空间e和f，对于任何开子集a e，对于任何函数f:a→f，对于任何a∈a，如果dkf存在于a中，对于所有k，1≤k≤m−1，并且如果dmf（a）存在，那么我们有：

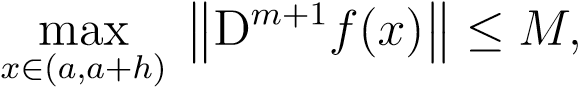
，

对于任意h，如a+h∈a，并且在哪里。

上述泰勒公式可用于研究实值函数的相对极大值（或极小值）。它还用于研究曲线和曲面的局部特性。

泰勒公式的下一个版本可以看作是引理38.12的推广。它有时被称为泰勒公式与拉格朗日余数或广义中值定理。

定理38.23。（广义中值定理）设e和f为两个赋范仿射空间，设a为→−e的开子集，设f:a→f为a上的函数，给定a∈a和e中任意h=06，如果闭段[a，a+h]包含在a中，dkf存在于a中，1≤k≤m，dm+1f（x）存在于a中t开口段的每个点x]a，a+h[，和



对于某些m≥0，则

.

作为推论，如果l:e−−m→+1→→−f是一个连续的（m+1）-线性图，那么

，

其中m=maxx∈（a，a+h）kdm+1f（x）−lk。

上述定理有时是在稍微强一点的假设下表述的，即f是a上的cm函数。如果f:a→r是实值函数，则定理38.23可以稍加改进。这个版本通常被称为泰勒-麦克劳林公式。

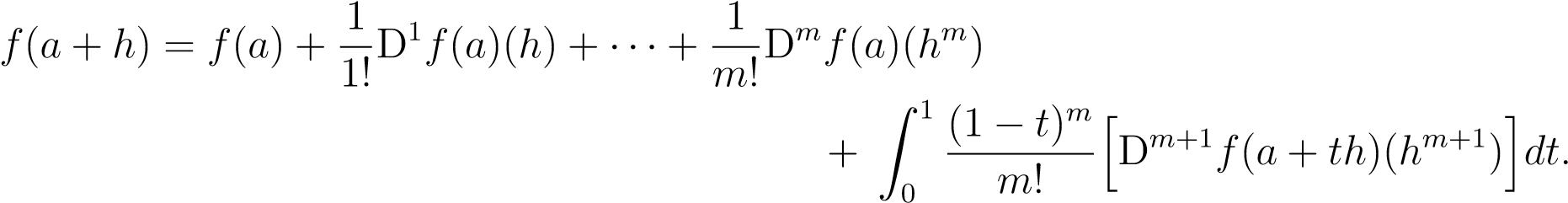
38.6。泰勒公式，法阿·迪布鲁诺公式`

定理38.24。（泰勒-麦克劳林）设e为赋范仿射空间，设a为→−e的开子集，设f:a→r为a上的实值函数，给定a∈a和e中任意h=06，如果闭段[a，a+h]包含在a中，如果dkf存在于a中，1≤k≤m，dm+1f（x）存在于开口段的每个点x]a，a+h[，则有一些θ∈r，其中0<θ<1，这样

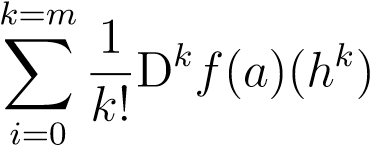
.

我们还提到了“数学文化”，在实值函数的情况下，一个带有整数余数的版本。这通常被称为带积分余数的泰勒公式。

定理38.25。（带积分余数的泰勒公式）设e为赋范仿射空间，设a为→−e的开子集，设f:a→r为a上的实值函数，给定e中任意a∈a和任意h=06，如果闭段[a，a+h]包含在a中，如果f是a上的cm+1-函数，则我们得到



上述公式的优点是给出了一个显式余数。现在我们简单地研究一下E是有限维n的情况，（a0，（e1，…，en））是e的一个框架。在这种情况下，我们得到一个更明确的表达式。



涉及泰勒公式的所有版本，其中根据惯例，d0f（a）（h0）=f（a）。如果h=h1e1+·····+hnen，那么我们有

，

使用第38.5节末尾引入的缩写符号，也可以写成

.

上述符号的高级是，它与n=1时使用的符号相同，即e=r（或e=c）时使用的符号相同。实际上，在这种情况下，泰勒-麦克劳林公式如下：

，

对于某些θ∈r，其中0<θ<1，其中dkf（a）是f在a处的k阶导数的值（因此，正如我们已经说过的多次，这是k阶矢量导数，因为f=r，它只是一个标量）。

在上述公式中，假设f:[a，a+h]→r是[a，a+h]上的cm函数，dm+1f（x）存在于每x∈（a，a+h）上。

泰勒公式有助于研究曲线和曲面的局部性质。在曲线的情况下，我们考虑一个函数f:[r，s]→f，从r的闭区间[r，s]到一些仿射空间f，导数dkf（a）（hk）对应于向量hkkf（a），其中dkf（a）是f在a处的第k个矢量导数（实际上是dkf（a）（1，…，1）），并且对于任何a∈（r，s），定理38.22得出以下公式：

，

对于任意h，如a+h∈（r，s），并且其中lim

在函数f:rn→r的情况下，可以方便地得到泰勒-杨公式和泰勒-麦克劳林公式的梯度和黑森公式。回想一下，f在a∈rn处的梯度f（a）是列向量。

，

而f0（a）（u）=df（a）（u）=f（a）·u，

对于任何u∈rn（其中·表示内积）。a∈rn处f的Hessian矩阵2f（a）是n×n对称矩阵。

，

我们有

d2f（a）（u，v）=u>2f（a）v=u·2f（a）v=2f（a）u·v，

38.7。向量场，协变导数，方括号

对于所有的u，v∈rn。然后，我们有下面三个泰勒-杨的公式2级：

.

带LIM

我们应该记住，只有第一个公式是内在的（即，不依赖于基础的选择），而其他两个公式则取决于基础，而内部产品则取决于Rn。作为练习，读者应该为泰勒-麦克劳林2阶公式编写类似的公式。

泰勒公式的另一个应用是推导一个公式，该公式给出两个函数组成的mth导数，通常称为“fa`a di Bruno's公式”。该公式在处理样条曲线和曲面的几何连续性时很有用。

提案38.26。给定任意赋范仿射空间e，对于任意函数f:r→r和任意函数g:r→e，对于任意a∈r，让b=f（a），f（i）（a）=dif（a）和g（i）（b）=dig（b），对于任意m≥1，如果f（i）（a）和g（i）（b）存在于所有i，1≤i≤m，则（g f）（m）（a）=dm（g f）（a）存在并由以下公式：

.

当m=1时，上述公式简化为熟悉的公式。

（g\_f）0（a）=g0（b）f0（a）

对于m=2，我们有

（g\_f）（2）（a）=g（2）（b）（f（1）（a））2+g（1）（b）f（2）（a）。

# 38.7向量场、协变导数、方括号

在这一部分中，我们简要地考虑向量场和向量场的协变导数。这种导数在连续力学中起着重要作用。给定一个赋范仿射空间（e，→−e），（e，→−e）上的向量场是一个函数x:e→→−e。直观地说，向量场为e中的每一点分配一个向量。这些向量可以是力、速度、加速度等。

给定两个向量场x，y在e的某个开子集Ω上定义，对于每一点a∈Ω，我们要定义x相对于y在a的导数。这是一种方向导数，当我们沿着y移动时给出x的变化，我们用dy x（a）表示。导数dy x（a）的定义如下。

定义38.12.设（e，→−e）为赋范仿射空间。给定e的任何开子集Ω，给定在Ω上定义的任意两个向量场x和y，对于任何a∈Ω，协变导数

x w.r.t.的（或Lie导数）。a处的向量场y，用dy x（a）表示，是极限（如果存在）

，

哪里

如果y是一个常量向量场，则立即验证映射

x 7→dy x（a）

是一个称为向量场x导数的线性映射，用dx（a）表示。如果f:e→r是一个函数，我们将dy f（a）定义为极限（如果存在）。

，

式中U=t∈r a+ty（a）∈Ω，t 6=0。它是f w.r.t.的方向导数，a处的向量场y，也常用y（f）（a）或y（f）a表示。

从现在开始，我们假设所有的向量场和所有考虑的函数都是光滑的（c∞）。平滑c∞-函数f:Ω→r的集合c∞（Ω）是一个环。给定一个平滑向量场x和一个平滑函数f（均大于Ω），定义向量场fx，使（fx）（a）=f（a）x（a），并立即验证它是平滑的。因此，Ω上平滑向量场的集合x（Ω）是C∞（Ω）模块。

下面的建议留作练习。结果表明，Dy x（a）是x（Ω）上的r-双线性映射，在y上是c∞（Ω）-线性的，并且满足关于x的莱布尼兹推导规则。

提案38.27。协变导数dy x（a）满足以下性质：

d（y1+y2）x（a）=dy1x（a）+dy2x（a）

dfy x（a）=f（a）dy x（a）

Dy（x1+x2）（a）=Dy x1（a）+Dy x2（a）

dy f x（a）=dy f（a）x（a）+f（a）dy x（a）

其中，x、y、x1、x2、y1、y2是Ω上的平滑向量场，f:e→r是平滑函数。

38.8。进一步阅读

在微分几何中，为了定义流形上向量场的协变导数，将上述性质作为仿射连接的公理。在许多情况下，向量场y是一些光滑曲线γ：]-η，η[→e的切向场。如果是，则以下命题成立。

提案38.28。给定一条平滑曲线γ：−η，η[→e，假设y是γ（]−η，η[）上定义的向量场，这样

，

对于γ（]-η，η[）上定义的任何向量场x，我们有

D

其中a=γ（0）。

因此，导数dy x（a）是向量场x沿曲线γ的导数，称为x沿γ的协变导数。

给定（e，→−e）的仿射帧（o，（u1，…，un）），很容易看出协变导数dy x（a）表示如下：

d.

通常，dy x（a）=6 dxy（a）。数量

[x，y]=dxy−dy x

称为向量场x和y的方括号。在微分几何中，方括号起着重要作用。在坐标方面，

.

# 38.8进一步读数

在Munkres[126]、Lang[109]、Schwartz[147]、Cartan[34]和Avez[9]中可以找到对微分学的彻底治疗。微分学技术有许多应用，特别是在曲线和曲面的几何以及一般微分几何中。为此，我们推荐Do Carmo[53，54]（两本关于这一主题的经典著作）、Kreyszig[104]、Stoker[161]、Gray[81]、Berger和Gostiaux[13]、Milnor[123]、Lang[107]、Warner[180]和Choquet Bruhat[38]。

1348第38章。微分学

第六部分

优化理论准备

一千三百四十九

第三十九章

# 实值函数的极值

## 39.1局部极值、约束局部极值和拉格朗日乘数

设j:e→r为赋范向量空间e j上定义的实值函数（或更大范围的拓扑空间）。理想情况下，我们希望找到函数的位置

最小值或最大值，至少在本地。在本章中，我们通常使用符号表示j在u处的导数，而不是dj（u）。我们的陈述与Ciarlet[41]的陈述非常接近（第7章），我们发现这是最清楚的一个。

定义39.1.e，我们说j有a if j：局部最小值e→r是在赋范向量空间（或相对最小值）上定义的实值函数，如果存在

一些包含u的开子集w e，这样

j（u）≤j（w），对于所有w∈w。

同样地，如果有一个包含u的开子集w e，我们认为j在点u∈e处有一个局部最大值（或相对最大值），这样

j（u）≥j（w），对于所有w∈w。

在这两种情况下，我们都说j在u有一个局部极值（或相对极值）。

jsome open subsethas a strict local minimumw e containing（resp.在点u∈e处，如果存在

j（u）<j（w）for all w∈w−u

（RESP）

j（u）>j（w）表示所有w∈w−u）。

一千三百五十一

通过滥用语言，我们经常说u点本身是“局部最小值”或“局部最大值”，尽管严格来说，这是没有意义的。

我们从一个众所周知的局部极值的必要条件开始。

提案39.1.设e为赋范向量空间，设j:Ω→r为函数，其中Ω为e的一些开子集。如果函数j在某个点上有局部极值u∈Ω，且

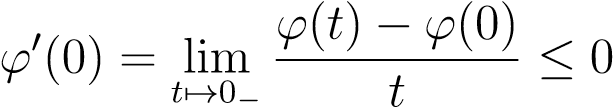
如果j在u处是可微的，那么dju=j0（u）=0。

证据。选取任意v∈e，因为Ω是开的，对于t足够小的话，我们有u+tv∈Ω，所以有一个开的区间i r，这样函数\_

⑨（t）=j（u+tv）

对于所有的t∈i都是定义明确的。通过应用链式法则，我们发现，当t=0时，a是可微的，我们得到\_（0）=dju（v）。

在不失去一般性的情况下，假设u是局部最小值。然后我们有了



和

，

这表明\_（0）=dju（v）=0。由于v∈e是任意的，我们得出dju=0。

点u∈Ω，使j0（u）=0称为j的临界点。

如果e=rn，那么条件dju=0等于系统

.

命题39.1的条件只是极值存在的必要条件，但不是充分条件。下面是一些反例。如果f:r→r是f（x）=x3给出的函数，因为f0（x）=3x2，我们得到f0（0）=0，但0既不是f的最小值，也不是f的最大值。如果g:r2→r是g（x，y）=x2−y2给出的函数，则=（0 0 0），但接近（0,0），则函数g取负值和正值。

在许多实际情况下，我们需要寻找函数j在附加约束下的局部极值。这种情况可以很方便地形式化为：我们在赋范向量空间的一些开子集Ω上定义了一个函数j:Ω→ru，但是我们也定义了Ω，我们正在寻找j关于

有一些子集集U。

在寻找某些目标函数的局部极值时，元素u∈u通常被称为关于某个子集的优化问题con-j的可行解。

由一组约束定义的。请注意，在大多数情况下，u不是打开的。实际上，u通常是闭合的。

定义39.2.如果j:Ω→r是赋范向量空间e的某个开子集Ω上定义的实值函数，如果u是Ω的某个子集，则我们认为j在点u∈u处具有局部最小值（或相对最小值），如果存在包含u的某个开子集wΩ，那么

j（u）≤j（w）表示所有w∈u w。

同样地，如果存在一些包含u的开子集wΩ，我们认为j在点u∈u处有一个局部最大值（或相对最大值），这样

j（u）≥j（w），对于所有w∈u w。

在这两种情况下，我们都说j在u处对u有一个局部极值。

注意到Ω为开的假设对39.1命题的有效性至关重要。例如，如果j是r上的恒等函数，u=[0,1]是一个封闭的子集，那么j0（x）=1表示所有x∈[0,1]，即使j在x=0处有一个最小值，在x=1处有一个最大值。

因此，为了找到函数j:Ω→r对于Ω的子集u具有局部极值的必要条件（其中Ω是开的），我们需要以某种方式将u的定义纳入这些条件中。这可以在两种情况下完成：

1. 集合U由一组方程定义，
   1. =x∈Ω\_i（x）=0，1≤i≤m，

式中，函数i:Ω→r是连续的（通常是可微的）。

1. 集合U由一组不等式定义，
   1. =x∈Ω\_i（x）≤0，1≤i≤m，

式中，函数i:Ω→r是连续的（通常是可微的）。

在（1）式中，式i（x）=0称为等式约束，在（2）式中，不等式i（x）≤0称为不等式约束。

式i（x）≥0的不等式约束等同于不等式约束−x（x）≤0。一个等同性约束\_i（x）=0相当于两个不等式约束\_i（x）≤0和−\_i（x）≤0的结合，因此不等式约束的情况包含了等同性约束的情况。然而，平等约束的情况更容易处理，在本章中，我们将限制我们对这一情况的关注。

如果函数i为凸函数，而Ω为凸函数，则u为凸函数。这是我们稍后将讨论的一个非常重要的案例。特别是，如果函数\_i是仿射的，那么对于一些m×n矩阵a和一些向量b∈rm，等式约束可以写成ax=b，不等式约束可以写成ax≤b。稍后我们还将讨论仿射约束的情况。

在等式约束的情况下，可以用拉格朗日乘子给出关于u的局部极值的必要条件。在不等式约束的情况下，对于广义拉格朗日乘子和卡鲁什-库恩-塔克条件，也存在一个关于u的局部极值的必要条件。这将在第49章中讨论。

我们首先考虑这样一种情况，即Ωe1×e2是赋范向量空间乘积的开子集，其中u是某个连续函数的零点轨迹，即

u=（u1，u2）∈Ω（u1，u2）=0。

为了简洁起见，我们说j在u处有一个约束的局部极值，而不是说j在u∈u处有一个局部极值，幸运的是，对于拉格朗日乘子，约束的局部极值有一个必要的条件。

定理39.2。（约束极值的必要条件）设Ωe1×e2为赋范向量空间乘积的开子集，其中e1 a Banach空间（e1是完整的），设\_：Ω→e2为c1函数（这意味着d\_（ω）存在且对所有ω∈Ω都是连续的），并设

u=（u1，u2）∈Ω（u1，u2）=0。

另外，让u=（u1，u2）∈u是这样一个点

而且，

设j:Ω→r是一个在u上可微的函数，如果j在u上有一个约束的局部极值，那么有一个连续的线性形式∧（u）∈l（e2；r），这样

dj（u）+∧（u）d（u）=0.

证据。攻击计划是使用隐函数定理；定理38.14。注意，定理38.14的假设确实得到了满足。因此，存在一些开放子集u1 e1，u2 e2，以及一个连续函数g:u1→u2，其中（u1，u2）u1×u2 oz，且使得ω（v1，g（v1））=0

对于所有v1∈u1。此外，g在u1∈u1上是可微的，并且

.

由此可知，j对（u1×u2）u的限制产生了一个单变量的函数g，其中

g（v1）=j（v1，g（v1））。

对于所有v1∈u1。现在，函数g在u1上是可微的，它在u1上有一个局部极值，所以命题39.1意味着

dg（u1）=0。

根据链式法则，

.

根据dg（u1）=0，我们推断

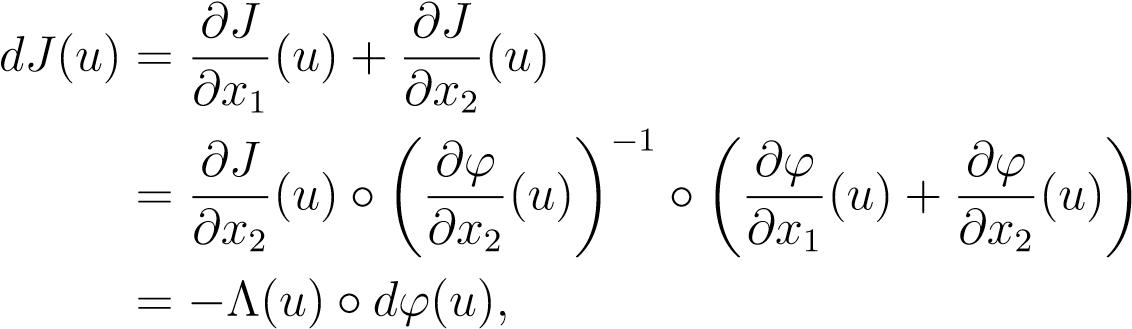
，

既然我们也有

，

如果我们让

然后我们得到



如权利要求所述，其产生dj（u）+∧（u）d（u）=0。

在大多数应用中，对于一些整数m，n，我们有e1=rn−m和e2=rm，因此1≤m<n，Ω是rn，j:Ω→r的开放子集，并且我们有m个定义子集的函数i:Ω→r

u=v∈Ω\_i（v）=0，1≤i≤m。

定理39.2给出了以下必要条件：

定理39.3。（拉格朗日乘子的约束极值的必要条件）设Ω为Rn的开子集，考虑m c1函数\_i:Ω→r（与

1≤m<n），让

u=v∈Ω\_i（v）=0，1≤i≤m，

设u∈u为一点，使得导数d\_i（u）∈l（r n；r）是线性无关的；等价地，假设m×n矩阵具有秩m，如果j:Ω→r是在u∈u处可微的函数，如果j在u处具有局部约束极值，则存在m个数λ。i（u）∈r，唯一定义，这样

dj（u）+λ1（u）d\_1（u）+····+λm（u）d\_m（u）=0；

相当地，

j（u）+λ1（u）1（u）+····+λ1（u）m（u）=0.

证据。m线性形式d i（u）的线性独立性等于m×n矩阵具有秩m的事实。通过重新排序列，我们可以假设第一个m列是线性独立的。如果我们让\_：Ω→Rm是由\_（v）=（\_1（v），…，\_m（v）定义的函数）

对于所有v∈Ω，则我们看到\_/x2（u）是可逆的，且\_/x2（u）及其逆式都是连续的，因此定理39.2适用，并且存在一些（连续）线性形式∧（u）l（rm；r），因此dj（u）+∧（u）d（u）=0。

然而，∧（u）是由一些m-元组（λ1（u），…，λm（u））定义的∈Rm，并且考虑到τ的定义，上述方程等价于

dj（u）+λ1（u）d\_（u）+····+λm（u）d m（u）=0.

λi（u）的唯一性是d i（u）的线性独立性的结果。

定理39.3中涉及的数λi（u）被称为与约束极值u相关的拉格朗日乘数（同样，还有一些轻微的语言滥用）。线性形式d\_i（u）的线性独立性等于雅可比矩阵具有秩m的事实。如果m=1，则d\_i（u）的线性独立性减小到条件\_1（u）=0.6。

重新表述拉格朗日乘子的一个有效方法是引入拉格朗日的概念，它与我们的约束极值问题有关。这就是功能

L:Ω×Rm→R由给出

L（V，λ）=J（V）+λ1\_1（V）+····+λm\_m（V），

用λ=（λ1，…，λm）。然后，观察存在一些μ=（μ1，…，μm）和一些u∈u，这样

dj（u）+\_1（u）+···+\_m（u）=0

如果且仅当

dl（u，μ）=0，

或同等

L（U，μ）=0；

也就是说，iff（u，λ）是拉格朗日L的一个临界点。

实际上，dl（u，μ）=0，如果等于

，

从那以后



和

，

我们得到dj（u）+\_1（u）+···+\_m（u）=0

\_1（u）=·····=\_m（u）=0，

也就是说，u∈u。

如果我们明确地写出条件

dj（u）+\_1（u）+···+\_m（u）=0，

我们得到了N×M系统

，

重要的是要注意，这个系统的矩阵是在u处的\_雅可比矩阵的转置。如果我们写jac（对于j的雅可比矩阵（at

u），则上述系统以矩阵形式写为

j（u）+（jac（j）（u））>λ=0，

其中，λ被视为列向量，拉格朗日等于

L（u，λ）=j（u）+（\_（u），…，m（u））λ。

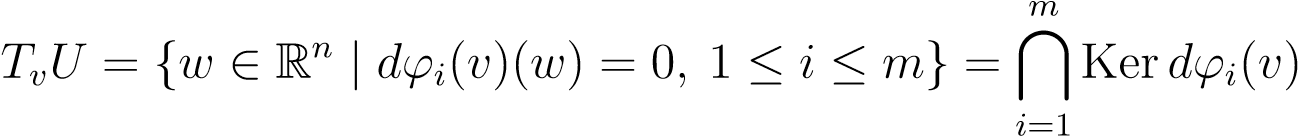
注：如果雅可比矩阵jac（所有v∈u都有秩m

（相当于线性形式d\_i（v）的线性独立性），那么我们说

0∈Rm是一个正则值。在这种情况下，我们知道

U=V∈Ω\_（V）=0

是RN尺寸n−m的光滑子流形。此外，这套



是V处U的切线空间（尺寸为N-m的向量空间）。然后，条件

dj（v）+\_1（v）+····+\_m（v）=0

意味着dj（v）在切线空间tvu上消失。相反，如果所有w∈tvu的dj（v）（w）=0，这意味着dj（v）与tvu是正交的（定义10.3（vol.i））。由于（根据定理10.1（b）（vol.i））tvu的正交是d\_（v），…，d m（v）所跨越的线性形式的空间，因此dj（v）必须是d i（v）的线性组合。因此，当0是一个正则值时，定理39.3断言，如果u∈u是j的局部极值，则dj（u）必须在切线空间tuu上消失。我们可以说得更多。Ω的z（j）子集由下式给出

Z（J）=V∈ΩJ（V）=J（U）

（j（u）级的水平集）是Ω中的超曲面，如果dj（u）=06，dj（u）的零轨迹是u（n-1维的向量空间）上的切线空间tuz（j）到z（j），其中

tuz（j）=w∈rn dj（u）（w）=0。

因此，定理39.3断言

tuu tuz（j）；

这是一个几何条件。

拉格朗日的美在于，约束\_i（v）=0已被纳入函数l（v，λ）中，并且约束j局部极值存在的必要条件被简化为非约束j局部极值存在的必要条件。D·L

但是，应仔细检查是否满足定理39.3的假设（尤其是线性形式d i的线性独立性）。例如，让

j:r3→r由下式给出

J（x，y，z）=x+y+z2

且g:r3→r×g（x，y，z）=x2+y2。

由于g（x，y，z）=0 iff x=y=0，我们得到u=（0,0，z）z∈r和j对

u由给出

j（0,0，z）=z2，

其最小值为z=0。然而，对拉格朗日乘子的“盲”使用要求存在一些λ，以便

，

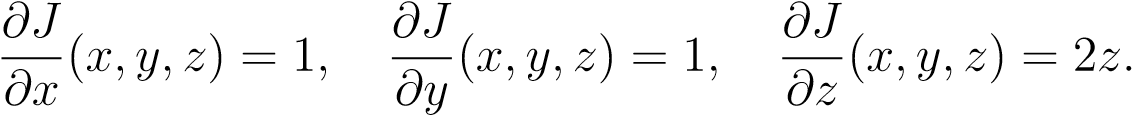
从那以后

，

首先，偏导数在x=y=0时消失，所以在局部极值时，我们也应该

，

但这是荒谬的，因为



读者应该喜欢找出论点中缺陷的原因。

我们还应该记住，定理39.3只给出了一个必要的条件。（u，λ）可能不符合局部极值！因此，总是有必要分析J在临界点U附近的局部行为。这通常是困难的，但在J是仿射或二次的情况下，约束是仿射或二次的情况下，这是可能的（尽管并非总是容易的）。

让我们将上述方法应用于以下示例，其中e1=r，e2=r，

Ω=r2，和

J（x1，x2）=-x2（x1，x2）=x21+x22−1.注意



是单位圆，从

，

很明显，单位圆上每个点的（x1，x2）=06=（x1，x2）。如果我们形成拉格朗日

-至-

定理39.3指出，j有一个约束局部极值的必要条件是l（x1，x2，λ）=0，因此下列方程必须成立：

2λx1=0

−1+2λx2=0

X21+X22=1.

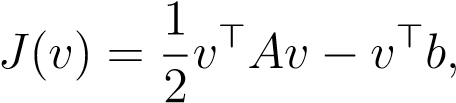
第二个方程意味着λ=06，然后第一个方程得出x1=0，所以第三个方程得出x2=±1，我们得到两个解：

（x1，x2）=（0,1）

，（X01，X02）=（0，−1）。

我们可以立即检查第一个解决方案是最小的，第二个是最大的。读者应该寻找这个问题的几何解释。

现在让我们考虑一个情况，其中j是形式的二次函数



其中a是n×n对称矩阵，b∈rn，约束由形式的线性系统给出。

cv=d，

39.2。利用二阶导数求极值

其中c是m×n矩阵，m<n，d∈rm。我们还假设c的等级为m。

在这种情况下，函数\_由下式给出：

⑨（v）=（cv−d）>，

因为我们将（v）视为行向量（v视为列向量），并且

d\_（v）（w）=c>w，

满足了在u处\_的雅可比矩阵具有秩m的条件。这个问题的拉格朗日方程是

，

其中，λ被视为列向量。现在，因为a是对称矩阵，所以很容易证明

.

因此，求逆局部极值的必要条件是

av+c>λ=b

cv=d，

可以用矩阵形式表示为

，

其中系统的矩阵是对称矩阵。我们不应该惊讶地发现第41节的系统，除了对涉及的矩阵和向量进行了一些重命名。从41.2节我们知道，函数j有一个最小的iff a是正定的，所以一般来说，如果a只是一个对称矩阵，拉格朗日的临界点不对应j的极值。

我们现在研究了涉及j的二阶导数的极值存在的条件。

## 39.2使用二阶导数求极值

为了简洁起见，我们只考虑局部极小的情况；对于局部极大值，得到了类似的结果（用−j替换j，因为maxu j（u）=−minu−j（u））。我们从一个非约束局部极小值的必要条件开始。

提案39.4.设e为赋范向量空间，设j:Ω→r为一个函数，用Ω表示e的一个开子集。如果函数j在Ω中可微，如果j在某一点上有一个二阶导数d2j（u）∈Ω，如果j在u处有一个局部最小值，那么d2j（u）（w，w）≥0表示所有w∈e。

证据。选取任意一个非零向量w∈e，由于Ω是开的，对于t足够小，u+tw∈Ω和j（u+tw）≥j（u），所以有一些开区间i r，这样

u+tw∈Ω和j（u+tw）≥j（u）

对于所有的t∈i，利用泰勒-杨公式和我们必须有dj（u）=0的事实，因为j在u处有一个局部最小值，我们得到

，

当lim=0时，这意味着

d2j（u）（w，w）≥0.

因为这个论点适用于所有的w∈e（如果w=0，则很简单），所以证明了这个命题。

应该注意的是，与前面的命题没有矛盾。例如，函数f:x 7→x3在0处没有局部最小值，而df（0）=0和d2f（0）（u，v）=0。同样，读者应检查函数f:r2→r是否由

F（x，y）=x2−3y3

在（0,0）时没有局部最小值；然而df（0,0）=0和d2f（0,0）（u，v）=2u2≥0。

当e=rn时，命题39.4表示具有局部最小值的必要条件是Hessian 2j（u）是正半定的（它总是对称的）。

我们现在给出了局部极小值存在的充分条件。

定理39.5。设e为赋范向量空间，设j:Ω→r为e的某个开子集为Ω的函数，并假定j在Ω中可微，dj（u）=0在某点u∈Ω。以下属性保留：

1. 如果d2j（u）存在并且有一些数α∈r，那么α>0和

d2j（u）（w，w）≥αkwk2，对于所有w∈e，

那么j在u有一个严格的局部最小值。

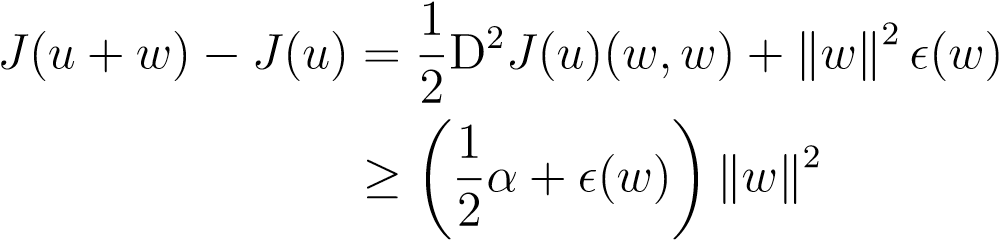
1. 如果d2j（v）存在于所有v∈Ω，并且如果有一个以u为中心的球bΩ，那么

d2j（v）（w，w）≥0，对于所有v∈b和所有w∈e，

那么j的局部最小值为u。

39.2。利用二阶导数求极值

证据。（1）使用泰勒-杨公式，对于每个足够小的向量w，我们可以写



当lim=0时。因此，如果我们选取r>0足够小，对于kwk<r的所有w，那么j（u+w）>j（u）对于所有u+w∈b，其中b是中心u和半径r的开球，这证明j在u处有一个局部严格最小值。

（2）泰勒-麦克劳林公式表明，对于所有u+w∈b，我们有

，

对于一些v∈（u，w+w）。

定理39.5的两个断言没有相反的结果。然而，在d2j（u）上有一个条件暗示了第（1）部分的条件。由于当e=rn时这种情况更容易说明，我们从这个例子开始。

回想一下，如果x>ax>0，所有x∈rn−0，n×n对称矩阵a是正定的。特别是，a必须是可逆的。

提案39.6.对于任何对称矩阵a，如果a是正定的，那么有一些α>0，这样

x>ax≥αkxk2，所有x∈rn。

证据。选择RN中的任何规范（记住RN上的所有规范都是等效的）。由于单位球面sn−1=x∈rn kxk=1是紧凑的，并且由于函数f（x）=x>ax在sn−1上从不为零，因此函数f在sn−1上具有最小α>0。使用x=kxk（x/kxk）的惯用技巧，对于每一个非零向量x∈rn，以及对于x=0，命题的不等式是平凡的事实，从

x>ax≥α，对于所有x，kxk=1，

我们得到

x>ax≥αkxk2，对于所有x∈rn，

如要求。

我们可以把定理39.5和命题39.6结合起来，得到严格局部极小存在的一个有用的充分条件。首先让我们介绍一些术语。

定义39.3.如前所述，给定函数j:Ω→r，如果dj（u）=0，并且hessian矩阵2j（u）是可逆的，那么点u∈Ω是非退化临界点。

提案39.7。设j:Ω→r为某个开放子集中定义的函数。如果j在Ω中是可微的，并且如果某点u∈Ω是非退化临界点，使得2j（u）是正定的，那么j在u处有一个严格的局部极小值。

注：通过对非退化临界点概念的适当推广，可以将39.7命题推广到无限维空间。首先，我们假设e是一个Banach空间（一个完全赋范向量空间）。然后，我们将e的对偶e0定义为e上的一组连续线性形式，因此e0=l（e；r）。遵循lang，我们对连续线性形式的空间使用符号e0，以避免与从e到r的所有线性映射的空间e=hom（e，r）混淆。连续双线性映射在l2（e，e；r）中得到由e到e0的映射

Φ（u）=u，

式中，ωu∈e0为线性形式，定义如下：

u（v）=a（u，v）。

可以很容易地检查，u是连续的，地图Φ是连续的。那么，我们认为，当Φ：e→e0是Banach空间的同构时，它是非退化的，这意味着Φ是可逆的，并且Φ和Φ-1都是连续线性映射。给定一个在Ω上可微的函数j:Ω→r（其中，Ω是e的开子集），如果d2j（u）存在于某个u∈Ω，那么当dj（u）=0并且d2j（u）是非退化临界点时，我们说u是非退化临界点。当然，d2j（u）是正定的，如果d2j（u）（w，w）>0表示所有w∈e−0。

利用上述定义，命题39.6可以推广到非退化正定双线性形式（在Banach空间上），定理39.7也可以推广到J:Ω→R在Banach空间的开子集上定义的情形。有关详细信息和证据，请参阅《卡丹》[34]（第一部分第8章）和《阿维斯》[9]（第8章和第10章）。

在下一节中，我们将利用凸性；既在域Ω上，也在函数上。

J本身。

## 39.3利用凸性求极值

我们首先回顾凸集和凸函数的定义。

定义39.4.在任意实向量空间e下，如果c=∅或对于每对点u，v∈c，连接u和v的线段包含在c中，则e的子集c是凸的，即：

（1−λ）u+λv∈c表示所有的λ∈r，使0≤λ≤1。

给定任意两点的u v∈e，直线段[u，v]是集合

[u，v]=（1−λ）u+λv∈eλ∈r，0≤λ≤1。

显然，非空集c是凸的iff[u，v]c，当u，v∈c时。凸集的例子见图39.1。

（a）

（b）

图39.1：图（a）显示了一个球体在r3中不是凸的，因为绿色虚线不在其表面。图（b）显示实心球在r3中是凸的。

定义39.5.如果c是e的非空凸子集，则函数f:c→r是凸的。

（在c上）如果对于每对点u，v∈c，f（（1-λ）u+λv）≤（1-λ）f（u）+λf（v），对于所有λ∈r，使0≤λ≤1；

函数f是严格凸的（在c上），如果对于每对不同的点u，v∈c（u=6v），f（（1-λ）u+λv）<（1-λ）f（u）+λf（v）对于所有的λ∈r，使0<λ<1；

见图39.2。函数f:a→r的上图epi（f）定义在

Rn是Rn+1的子集，定义为

epi（f）=（x，y）∈rn+1 f（x）≤y，x∈a。

在凸子集C上定义的函数f:c→r是凹的（分别为如果（−f）为凸形（分别为严格凸形）。

显然，函数f如果凸iff其上图epi（f）是rn+1的凸子集。

u

v

l = (1-

λ

)

f(u) +

λ

f(v)

f

（a）

u

v

l = (1-

λ

)

f(u) +

λ

f(v)

f

（b）

图39.2：图（a）和（b）是实值函数的图。图（a）是凸函数图，因为蓝线位于F图的上方。图（b）显示了非凸函数图。

向量空间e的子空间v e是凸的；仿射子空间，即形式为u+v的集合，其中v是e的子空间，u∈e是凸的。球（开的或闭的）是凸的。给定任意线性形式，对于任意标量c∈r，闭半空间

，

是凸面的。任何半空间的交集都是凸的。一般来说，凸集的任何交集都是凸的。

线性形式是凸函数（但不是严格凸函数）。任何范数k:e→r+都是凸函数。max函数，

最大值（x1，…，xn）=max x1，…，xn

在RN上是凸形的。对于任何c=0（6c∈r），指数x 7→ecx都是严格凸的。对数函数在r+−0上是凹的，对数行列式函数logdet在对称正定矩阵集上是凹的。该函数在凸优化中起着重要作用。Boyd[29]对凸性及其在优化中的应用作了很好的阐述。

这是一个关于凸子集u的函数具有局部最小值的必要条件。

定理39.8。（凸子集上局部最小值的必要条件）设j:Ω→r为赋范向量空间e的某些开子集Ω上定义的函数，设uΩ为非空凸子集。给定任何u∈u，如果dj（u）存在，并且j在u中对u有局部极小值，那么

dj（u）（v−u）≥0，所有v∈u。

证据。设v=u+w为u中的任意点，因为u是凸的，所以所有t都有u+tw∈u，所以0≤t≤1。因为dj（u）存在，我们可以写



当lim=0时。但是，因为0≤t≤1，



由于u是u的局部最小值，我们得到j（u+tw）−j（u）≥0，所以我们得到

.

上面暗示dj（u）（w）≥0，因为否则我们可以选择t>0足够小，以便

，

矛盾。由于该论点适用于所有v=u+w∈u，因此证明了该定理。

观察到u的凸性是拉格朗日乘子的一个代替品，但是我们现在必须处理一个不等式而不是一个等式。

考虑u是e的子空间的特殊情况，在这种情况下，由于u∈u，我们有2u∈u，对于任何u+w∈u，我们必须有2u−（u+w）=u−w∈u。前面的定理暗示dj（u）（w）≥0和dj（u）（−w）≥0，即dj（u）（w）≤0，所以dj（u）=0。由于讨论的是w∈u（因为u是一个子空间，如果u，w∈u，那么u+w∈u），我们得出如下结论：

dj（u）（w）=0表示所有w∈u。

我们现在将描述凸函数的一阶导数或二阶导数。

提案39.9.（凸性和一阶导数）设f:Ω→r是赋范向量空间e的某些开子集Ω上可微的函数，设uΩ是非空凸子集。

1. 函数f在u iff上是凸的。

f（v）≥f（u）+df（u）（v−u）表示所有u，v∈u。

u

v

f

(

u, f(u

))

(

v, f(v

))

(

y,v

)

v - u

y - v

y = f(u) + df(u)(v-u)

图39.3：凸值函数f的图示。因为f是凸的，所以它总是位于其切线之上。

1. 函数f在u iff上是严格凸的。

f（v）>f（u）+df（u）（v−u）表示所有u，v∈u，u=6v。

见图39.3。

证据。设u，v∈u为任意两个不同点，选取λ∈r，取0<λ<1。如果函数f是凸的，那么

，

会产生

接下来是

.

如果f是严格凸的，则上述推理不起作用，因为严格不等式不必通过“传递到极限”来保持。我们采用以下技巧：对于任何ω，如果0<ω<1，则观察

.

如果我们假设0<λ≤ω，则f的凸性产生

.

如果我们减去两边的f（u），我们得到

.

现在，由于0<ω<1和f是严格凸的，

f（u+ω（v−u））=f（（1−ω）u+ωv）<（1−ω）f（u）+ωf（v），

这意味着

，

因此我们得到

.

如果我们把λ设为0，通过传递到极限，我们得到

，

这就产生了期望的严格不等式。

现在让我们来考虑（1）的逆命题；也就是说，假设

f（v）≥f（u）+df（u）（v−u）表示所有u，v∈u。

对于任意两个不同的点u，v∈u，对于任何一个λ0<λ<1，我们得到

f（v）≥f（v+λ（v−u））−λdf（v+λ（u−v））（u−v）

f（u）≥f（v+λ（u−v））+（1−λ）df（v+λ（u−v））（u−v）、

如果我们将第一个不等式乘以1−λ，第二个不等式乘以λ，它们加起来就得到了不等式。

（1−λ）f（v）+λf（u）≥f（v+λ（u−v））=f（（1−λ）v+λu），

证明f是凸的。

（2）倒数的证明是相似的，只是不平等被严格的不平等所代替。

我们现在用F的二阶导数建立一个凸性准则，这个准则比前一个准则更容易检验。

提案39.10。（凸性和二阶导数）设f:Ω→r是赋范向量空间e的某些开子集Ω上的两次可微函数，设uΩ为非空凸子集。

1. 函数f在u iff上是凸的。

d2f（u）（v−u，v−u）≥0表示所有u，v∈u。

1. 如果

d2f（u）（v−u，v−u）>0表示所有u，v∈u，u=6v，

那么f是严格凸的。

证据。首先，假设满足条件（1）中的不等式。对于任意两个不同的点u，v∈u，泰勒-麦克劳林公式得出

，

对于一些w=（1−λ）u+λv=u+λ（v−u），0<λ<1，且λ）>0，因此v−u=ρ（v−w）。由于d2f（u）（v−w，v−w）≥0对于所有u，w，我们通过应用命题39.9（1）得出结论。

同样，如果（2）成立，上述推理和命题39.9（2）意味着f是严格凸的。

为了证明（1）中的必要条件，定义G：Ω→R

g（v）=f（v）−df（u）（v）

其中u∈u是被认为是固定的点。如果f是凸的，因为

G（V）−G（U）=F（V）−F（U）−DF（U）（V−U），

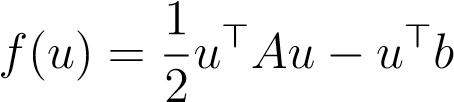
命题39.9意味着f（v）u）≥0，这意味着g在u处对所有值具有局部最小值。因此，我们得到dg（u）=0。观察到g在Ω和d2g（u）=d2f（u）中是两倍可微的，所以泰勒-杨公式对于每个v=u+w∈u和所有t，0≤t≤1，

，

如果lim=0，并且t足够小，我们必须有d2f（u）（w，w）≥0，如权利要求所述。

命题39.10（2）的逆命题是错误的，正如我们通过考虑f（x）=x4给出的函数f所看到的那样。

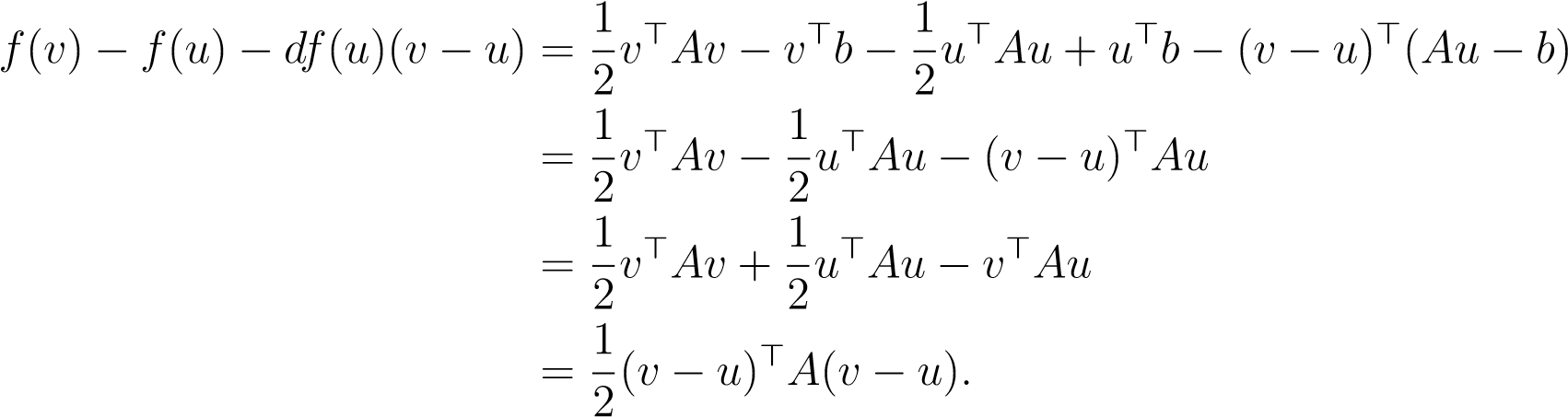
例39.1。另一方面，如果f是形式的二次函数



其中a是对称矩阵，我们知道

df（u）（v）=v>（au-b），

所以



因此，命题39.9意味着，如果a是正半定的，那么f是凸的，如果a是正定的，那么f是严格凸的。反过来，则是39.10号提案。

我们通过将前面的定理应用于凸子集上定义的凸函数来结束这一节。在这种情况下，局部最小值局部最大值）是全局最小值（分别为全球最大值）。

定义39.6.设f:e→r为在某个赋范向量空间上定义的任何函数（更一般地说，是任何集合）。对于任何u∈e，我们说f在u（resp）中有一个最小值。如果f（u）≤f（v）（resp.f（u）≥f（v））表示所有v∈e。

我们说f在u（resp）中有严格的最小值。严格的最大值，单位为u）如果

f（u）<f（v）（分别f（u）>f（v））表示所有v∈e−u。

如果u e是e和u∈u的子集，我们就说f在u（resp）中有一个最小值。关于u的严格最小值，如果f（u）≤f（v），对于所有v∈u（resp.f（u）<f（v）表示所有v∈u−u），

同样地，最大值单位为u（resp.u）中的严格最大值），关于u，其中≤变为≥和<变为>。

有时，我们说全局最大值（或最小值）来强调最大值（或最小值）不仅仅是局部最大值（或最小值）。

定理39.11。给定任意赋范向量空间e，设u为e的任意非空凸子集。

1. 对于任意凸函数j:u→r，对于任意u∈u，如果j在u中有局部极小值，则j在u中有（全局）极小值。
2. 任何严格凸函数j:u→r最多有一个极小值（单位为u），如果有，则它是一个严格极小值（单位为u）。
3. 设j:Ω→r为e的某个开子集Ω上定义的任何函数，其中uΩ，并假定j在u上是凸的。对于任何点u∈u，如果dj（u）存在，则j在u上对u iff具有最小值。

dj（u）（v−u）≥0，所有v∈u。

1. 如果（3）中的凸子集u是开的，则上述条件等于

dj（u）=0。

证据。（1）假设v=u+w是u中的任意点，因为j是凸的，所有t

0≤t≤1，我们有

j（u+tw）=j（u+t（v−u））≤（1−t）j（u）+tj（v），

会产生

j（u+tw）−j（u）≤t（j（v）−j（u））。

因为j在u中有一个局部最小值，所以有一些t0和0<t0<1，这样

0≤j（u+t0w）−j（u），

这意味着j（v）−j（u）≥0。

1. 如果j是严格凸的，则上述w=06的推理表明存在一些t0，其中0<t0<1，因此

0≤j（u+t0w）−j（u）<t0（j（v）−j（u）），

这表明u是一个严格的全局最小值（u），因此它是唯一的。

1. 从定理39.8我们已经知道，所有v∈u的条件dj（u）（v-u）≥0是必要的（即使j不是凸的）。相反地，由于j是凸的，仔细检查39.9号提案第（1）部分的证明表明，只有dj（u）存在这个事实才能证明

j（v）−j（u）≥dj（u）（v−u），对于所有v∈u，

如果

dj（u）（v−u）≥0，对于所有v∈u，

然后

j（v）−j（u）≥0，对于所有v∈u，

如要求。

1. 如果u是开的，那么对于每一个u∈u，我们可以找到一个以u为圆心的开球b，其半径足够小，使b u。那么，对于任何w=06这样的情况，我们都有v=u+w∈b和v0=u−w∈b，因此条件（3）意味着

dj（u）（w）≥0和dj（u）（−w）≥0，

得出dj（u）（w）=0。

由于上述公式适用于所有w=06，因此dj（u）是线性的，我们将其留给读者来填写dj（u）=0的证明细节。

定理39.11可用于重新推导线性系统的最小二乘解ax=b（其中a是m×n矩阵）由正态方程a>ax=a>b给出的事实。

为此，我们考虑二次函数

，

我们的最小二乘问题等价于求Rn上j的极小值。计算表明，因此

-

由于a>a是正半定的，函数j是凸的，定理39.11（4）表明j的极小值是方程的解。

a>au−a>b=0。

本章的考虑揭示了求导数图零点的方法的必要性

DJ：Ω→E0，

式中，Ω是赋范向量空间e的一些开子集，e0是e上所有连续线性形式的空间（e的子空间）。牛顿方法的推广产生了这样的方法，它们是下一章的目标。

## 39.4总结

本章的主要概念和结果如下：

* 最大局部最小值、局部最大值、局部极值、严格局部最小值、严格局部
* 包含导数的局部极值的必要条件；临界点。
* 关于子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子极小子
* 约束局部极值。
* 约束极值的必要条件。
* 拉格朗日乘子约束极值的必要条件。
* 拉格朗日
* 拉格朗日的临界点。
* 包含二阶导数的无约束局部极小的必要条件。•包含二阶导数的局部最小值的充分条件。
* 涉及非退化临界点的充分条件。
* 凹函数凸集，凸函数，，凹函数，严格凸函数，严格
* 包含导数的凸集上局部极小的必要条件。
* 关于一阶导数的条件的函数的凸性。
* 关于二阶导数条件的函数的凸性。
* 凸集上凸函数的极小值。

第四十章

# 牛顿方法及其推广

## 40.1牛顿实变函数法

在第39章中，我们研究了确定赋范向量空间e的某些开子集Ω上定义的函数j:Ω→r何时具有局部极值的问题。命题39.1给出了当j可微时的一个必要条件：如果j在u∈Ω处有一个局部极值，那么我们必须

j0（u）=0。

因此我们得到了求导数零点的问题。

j0:Ω→e0，

式中，e0=l（e；r）是从e到r的一组线性连续函数，即e的对偶，如命题39.7后的注释所定义。

这使我们以一种更一般的形式来考虑这个问题，即：给定一个函数f:Ω→y，从赋范向量空间x的开子集Ω到赋范向量空间y，发现

1. 保证函数f的零点存在的充分条件；即元素a∈Ω，使f（a）=0。
2. 一种近似这样一个A的算法，也就是说，其极限是A的Ω点序列（XK）。

当x=y=r时，我们可以使用牛顿方法。我们选取一些初始元素X0∈R“足够接近”F的零点A，并用

，

对于所有k≥0，前提是f0（xk）=06。其思想是将xk+1定义为x轴与点（xk，f（xk））处函数x 7→f（x）图形的切线的交点。实际上，这条切线的方程是

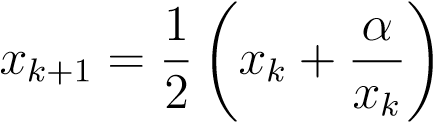
Y−F（XK）=F0（XK）（X−XK）

当y=0时，得到它与x轴的交点，得出

，

如要求。

例如，如果α>0且f（x）=x2−α，牛顿方法得出序列



计算α的平方根√α。可以看出，对于任何X0>0，该方法收敛到√α。实际上，当x0<0时，该方法也会收敛！找出限制是什么。

实数函数的情况表明，用Ωx求函数f:Ω→y的零点的方法如下：给定起点X0∈Ω，序列（Xk）由xk+1=xk−（f0（xk））−1（f（xk））定义。

所有k≥0。

为了使上述内容合理化，必须确保

1. 所有点XK保持在Ω范围内。
2. 功能F在Ω范围内可区分。
3. 导数f0（x）是所有x∈Ω的x到y的双射。

这些条件相当苛刻，但有足够的条件保证它们得到满足。另一个实际问题是，在每个迭代步骤中计算（f0（xk））-1可能非常昂贵。在下一节中，我们将研究牛顿方法的推广，它解决了我们刚刚讨论的问题。

## 40.2牛顿法的推广

假设F:Ω→Rn由n个函数fi:Ω→R给出，其中，ΩRn。在这种情况下，找到f的零点a等于解系统。

f1（a1…，an）=0 f2（a1…，an）=0

…

fn（a1…，an）=0.

牛顿方法的一次迭代就是求解线性系统。

，

然后设置

，

式中）是x k处f的雅可比矩阵。

一般来说，在每次迭代中计算j（f）（xk），然后求解相应的线性系统是非常昂贵的。如果该方法收敛，则连续向量xk与相应矩阵j（f）（xk）之间的差异应较小。因此，我们得到了牛顿方法的一种变体，它包括保持p连续步骤的相同矩阵（其中p是一些固定整数≥2）：

|  |  |
| --- | --- |
| 网络错误 |  |
| 网络错误 | 网络错误 |

也可以设置p=∞，也就是说，对所有迭代使用相同的矩阵f0（x0），这将导致表单的迭代。

XK+1=XK−（F0（X0））−1（F（XK）），K≥0，

或者甚至用一个特别的矩阵a0来代替f0（x0），这个矩阵a0很容易反转：

xk+1=xk−a−0 1f（xk），k≥0。

在最后两种情况下，如果可能的话，我们使用f0（x0）或a0的lu因子分解来加速该方法。在某些情况下，甚至可以设置a0=i。

上述考虑使我们得到了广义牛顿法的定义，如Ciarlet[41]所述（第7章）。回想一下，线性映射f∈l（e；f）被称为同构，如果f是连续的、双射的，并且f−1也是连续的。

定义40.1.如果x和y是两个赋范向量空间，如果f:Ω→y是x的某个开子集Ω的函数，则求f的零点的广义牛顿方法包括

1. 从x到y的线性同构族（a k（x））的序列，对于所有x∈Ω和所有整数k≥0；
2. 某起点X0∈Ω；
3. Ω点的序列（XK），定义如下：

x k+1=xk−（ak（x`）−1（f（xk）），k≥0，

其中，对于每个整数k≥0，整数满足条件

0≤`≤K。

函数ak（x）通常取决于f0。

定义40.1给了我们足够的灵活性来捕捉我们之前讨论过的所有情况：

|  |  |
| --- | --- |
| 网络错误 | 网络错误 |
| 网络错误  网络错误 |  |

ak（x）=a0，

其中a0是从x到y的线性同构。第一种情况对应于牛顿的原始方法，其他情况对应于我们刚才讨论的变体。我们也可以得到AK（x）=AK，一个独立于x∈Ω的固定线性同构。

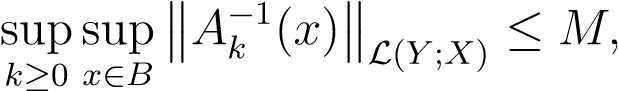
以下由牛顿-康托罗维奇定理启发而来的定理给出了保证由广义牛顿方法构造的序列（XK）收敛到接近X0的F零点的充分条件。尽管技术上很好，但这些条件并不令人惊讶。

定理40.1。设x为Banach空间，设f:Ω→y在开子集Ωx上可微，并假设有常数r、m、β>0，这样如果我们

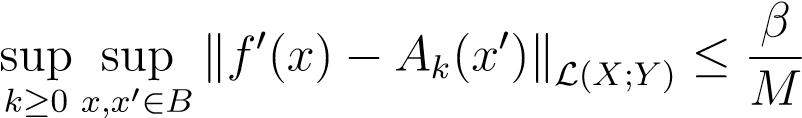
B=X∈X KX−X0K≤RΩ，

然后

（1）



（2）β<1和



（3）

.

然后，序列（XK）由

x k+1=xk−a−k 1（x`）（f（xk）），0≤`≤k

完全包含在b中并收敛到f的零点a，这是f中唯一的零点。

B.此外，收敛是几何的，这意味着

.

定理40.1的证明可在Ciarlet[41]中找到（第7.5节）。这不是很困难，但技术性很强。

如果假设我们已经知道某个元素a∈Ω是f的零点，那么下一个定理给出了一个广义牛顿方法特殊版本收敛的充分条件。对于这种特殊的方法，线性同构ak（x）独立于x∈Ω。

定理40.2。设x为Banach空间，设f:Ω→y在开子集Ωx上是可微的。如果a∈Ω是一个点，使得f（a）=0，如果f0（a）是线性同构，并且如果有一些λ具有0<λ<1/2，这样

，

然后有一个中心A的闭球B，这样对于每一个X0∈B，序列（XK）定义为

xk+1=xk−a−k 1（f（xk）），k≥0，

完全包含在b中并收敛到a，a是b中f的唯一零。此外，收敛是几何的，这意味着

KXK−AK≤βK KX0−AK，

对于某些β<1。

定理40.2的证明也可在Ciarlet[41]中找到（第7.5节）。

为了完整起见，我们提出了牛顿-康托罗维奇定理的一个版本，它对应于ak（x）=f0（x）的情况。在这种情况下，可以得到更强大的结果，特别是关于上界，我们提出了一个版本，由于GRAGG和TAPIA出现在CIARLET问题7.5-4中[41]。

定理40.3。（Newton–Kantorovich）设x为Banach空间，设f:Ω→y在开子集Ωx上可微。假设存在三个正常数λ、礹、礹和一个点x0∈Ω，这样

，

如果我们让

，

那么bΩ，f0（x0）是l（x；y）的同构，并且

.

那么，f0（x）是l（x；y）对于所有x∈b的同构，并且序列定义为

xk+1=xk−（f0（xk））−1（f（xk）），k≥0

完全包含在球B中，并收敛到F的0 A，即Ω+中F的唯一0。最后，如果我们写θ=ρ−/ρ+，那么我们有以下界限：

如果

如果，

和

.

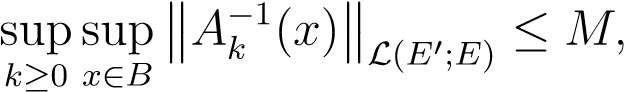
我们现在可以专门化定理40.1和40.2来搜索函数j:oz→r的导数j0:oz→e0的零点，其中有oz e。j的第二个导数j00是一个连续的双线性形式j00:e×e→r，但便于在l（e，e0）中将其看作一个线性映射；continuU线性形式j00（u）由j00（u）（v）=j00（u，v）给出。在下一个定理中，我们假设ak（x）是l（e，e0）中的同构。

定理40.4.设e为Banach空间，设j:Ω→r在开子集Ωe上是两倍可微的，并假设有常数r，m，β>0，这样如果我们

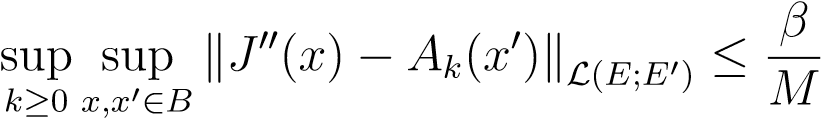
B=X∈E KX−X0K≤RΩ，

然后

（1）



（2）β<1和



（3）

.

然后，序列（XK）由

x k+1=xk−a−k 1（x`）（j0（xk）），0≤`≤k

完全包含在b中，并收敛到j0的零点A，这是b中j0的唯一零点。此外，收敛是几何的，这意味着

.

在下一个定理中，我们假设ak（x）是l（e，e0）中的同构，独立于x∈Ω。

定理40.5。设e为Banach空间，设j:Ω→r在开子集Ωe上是两次可微的。如果a∈Ω是一个点，使得j0（a）=0，如果j00（a）是一个线性同构，并且如果有一些具有0<λ<1/2的λ，那么

，

然后有一个中心为a的闭球b，这样对于每一个x0∈b，由xk+1=xk−a−k 1（j0（xk））定义的序列（xk），k≥0，

完全包含在B中并收敛到A，这是B中j0的唯一零。此外，收敛是几何的，这意味着

KXK−AK≤βK KX0−AK，

对于某些β<1。

当e=rn时，定理40.4给出的牛顿法给出了形式的迭代步骤。

x k+1=xk−a−k 1（x`）j（xk），0≤`≤k，

式中j（xk）是j在xk处的梯度（这里，我们用rn标识e0）。特别是牛顿的原始方法取ak=j00，迭代步骤为

XK+1=XK−（2J（XK））−1 J（XK），K≥0，

式中2j（xk）是j在xk的黑森。

如Ciarlet[41]所述（第7.5节），广义牛顿方法具有非常广泛的适用性。例如，各种版本的梯度下降法可以看作是牛顿法的实例。示例见第48.9节。

牛顿方法在凸优化特别是内点法中也起着重要作用。提出了牛顿法处理等式约束的一种新方法。我们请读者参考博伊德和范登堡[29]第10章和第11章，以全面阐述这些主题。

## 40.3总结

本章的主要概念和结果如下：

* 牛顿函数法f:r→r。
* 广义牛顿方法。
* 牛顿-康托罗维奇定理。

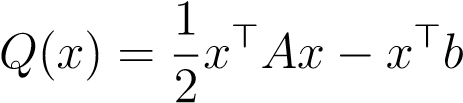
第四十一章

# 二次优化问题

## 41.1二次优化：正定情况

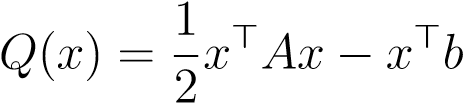
在本章中，我们考虑了工程和计算机科学（尤其是计算机视觉）中经常出现的两类二次优化问题：

1. 最小化



在所有x∈rn上，或受线性或仿射约束。

1. 最小化



在单位范围内。

在这两种情况下，a都是对称矩阵。我们还寻求F具有全局最小值的必要和充分条件。

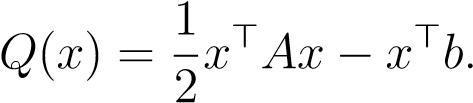
物理和工程中的许多问题可以说是某种能量函数的最小化，有或没有约束。事实上，自然的作用是使能量最小化，这是力学的一个基本原理。此外，如果物理系统处于稳定平衡状态，那么处于该状态的能量应该是最小的。例如，放置在球体顶部的小球处于不稳定的平衡位置。一个小动作使球滚下来。另一方面，由于势能很小，放置在球体内部和底部的球处于稳定的平衡位置。

最简单的能量函数是二次函数。这样的函数可以方便地在表单中定义

q（x）=x>ax−x>b，

一千三百八十三

其中a是对称n×n矩阵，x，b是rn中的向量，视为列向量。实际上，由于很快就会明白的原因，最好在二次项前面加一个因子，这样



问题是，在什么条件下（a）q（x）具有全局最小值，最好是唯一值？

我们分两个阶段完整回答上述问题：

1. 在这一节中，我们证明，如果a是对称正定的，那么q（x）在

ax=b。

1. 在第41.2节中，我们给出了一般情况下关于a的伪逆的必要和充分条件。

我们从定义20.2的矩阵版本（第一卷）开始。

定义41.1.对称正定矩阵是特征值严格为正的矩阵，对称正定半矩阵是特征值非负的矩阵。

等价标准在下面的命题中给出。

提案41.1.考虑到尺寸n的任何欧几里得空间e，下列性质成立：

1. 每个自伴线性映射f:e→e为正定iff

Hf（x），Xi＞0

对于所有x∈e，x=06。

1. 每一个自伴线性映射f:e→e都是半正定iff。

Hf（x），Xi超0

对于所有x∈e。

证据。（1）首先，假设f是正定的。回想一下，每个自伴线性映射都有一个特征向量的正交基（e1，…，en），并让λ1，…，λn作为相应的特征值。关于这个基础，每x=x1e1+····+xnen=06，我们有

，

这是严格正的，因为对于某些i，λi>0表示i=1，…，n，和x2i>0，因为x 6=0。

相反，假设HF（x），Xi＞0。

所有x=06。那么对于x=ei，我们得到

hf（ei），eii=hλiei，eii=λi，

因此，对于所有i=1，…，n，λi>0。

（2）如（1）所述，我们已经

，

由于λi超过0，对于i＝1，…，n，因为f是半正定的，我们有HF（x），Xi超过0，如所要求的。相反，如（1）所示，除了hf（ei），eii≥0，我们只得到λi≥0。

一些特殊的符号（特别是在凸优化领域）通常表示对称矩阵是正定或半正定的。

定义41.2.对于任意n×n对称矩阵，我们写的是半正定的，我们写的是正定的。

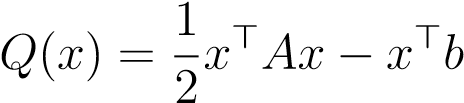
应该注意的是，我们可以定义这种关系



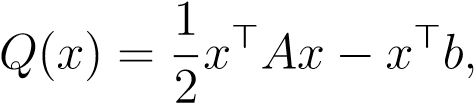
在任意两个n×n矩阵（对称与否）之间，iff a−b是对称半正定的。很容易确认这种关系实际上是矩阵上的一个偏序，称为半正定锥序；有关详细信息，请参阅Boyd和Vandenberghe[29]第2.4节。

如果a是对称正定的，很容易检查−1也是对称正定的。另外，如果c是对称正定m×m矩阵，a是秩n的m×n矩阵（因此m≥n，且映射x 7→ax被投影到rm上），则a>ca是对称正定的。

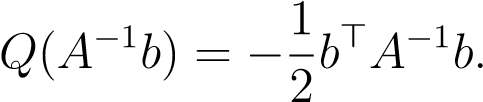
我们现在可以证明



当a是对称正定时具有全局最小值。提案41.2.给定二次函数

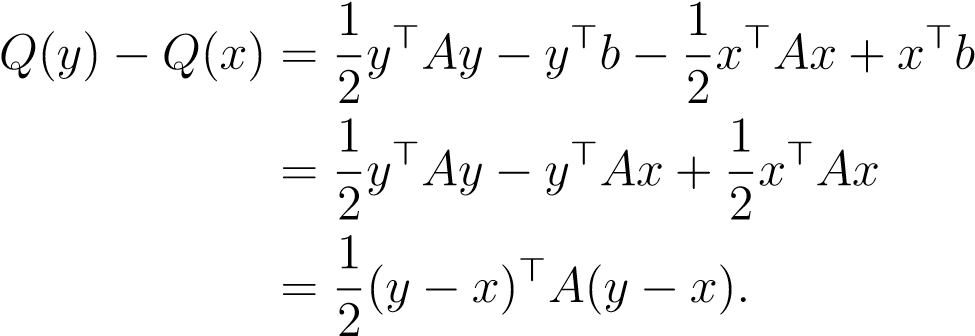


如果a是对称正定的，则q（x）对于线性系统的解ax=b具有唯一的全局最小值。q（x）的最小值为



证据。因为a是正定的，所以它是可逆的，因为它的特征值都是严格正的。

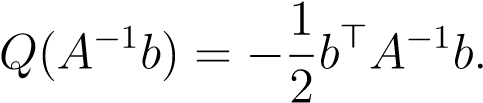
设x=a−1b，并计算任意y∈rn的q（y）−q（x）。因为ax=b，我们得到



因为a是正定的，所以最后一个表达式是非负的，因此

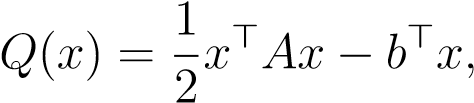
q（y）≥q（x）

对于所有y∈rn，证明x=a−1b是q（x）的全局最小值。简单的计算得出



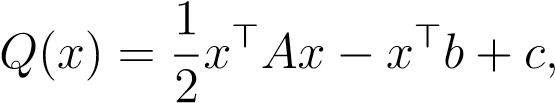
评论：

1. 二次函数q（x）也由

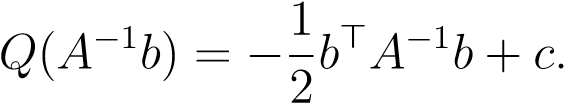


但是，使用x>b的定义对于41.2号提案的证明更为方便。

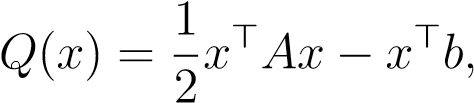
1. 如果q（x）包含一个常数项c∈r，那么



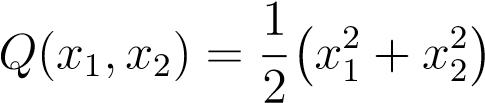
命题41.2的证明仍然表明，对于x=a−1b，q（x）具有唯一的全局最小值，但最小值为



因此，当系统的能量函数q（x）由二次函数给出时



如果a是对称正定的，求q（x）的全局最小值等于解线性系统ax=b。有时，将线性问题ax=b重设为变分问题（求某个能量函数的最小值）是有用的。然而，通常情况下，最小化问题会带有额外的约束，必须满足所有可接受的解决方案。例如，我们可能想最小化二次函数



受约束的

2x1−x2=5.

q（x1，x2）最小的溶液不再是（x1，x2）=（0,0），而是（x1，x2）=（2，−1），如下文所示。

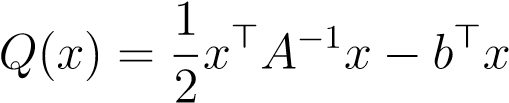
几何上，r3中z=q（x1，x2）定义的函数图是旋转轴oz的p的抛物面。约束

2x1−x2=5

与平行于Z轴的垂直平面H相对应，并且在xy平面中包含公式2x1−x2=5的线。因此，q的约束最小值位于抛物面上，即抛物面p与平面h的交点。

解决上述约束最小化问题的一个好方法是使用第39.1节中讨论的拉格朗日乘子方法。但首先，让我们精确地定义我们打算解决的最小化问题。

定义41.3.二次约束最小化问题包括二次函数的最小化。



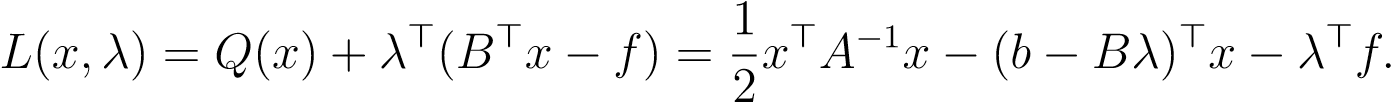
受线性约束

b>x=f，

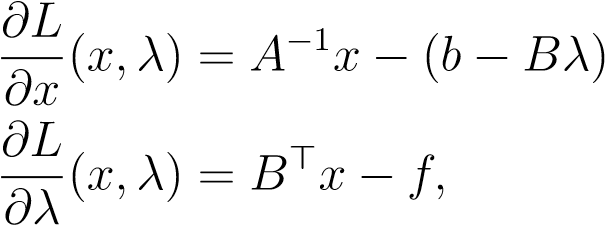
其中a−1是m×m对称正定矩阵，b是n阶m×n矩阵（使m≥n），其中b，x∈rm（视为列向量），f∈rn（视为列向量）。

使用−1而不是a的原因是约束最小化问题被解释为一组平衡方程，其中自然产生的矩阵是a（见Strang[164]）。由于a和a−1都是对称正定的，所以这没有任何区别，但最好还是坚持strang的表示法。

如第39.1节所述，拉格朗日乘数的方法包括将n个约束b>x=f合并到二次函数q（x）中，通过引入称为拉格朗日乘数的额外变量λ=（λ1，…，λn），每个约束一个。我们组成拉格朗日



从定理39.3可知，约束优化问题有解的一个必要条件是l（x，λ）=0。自从



我们得到了线性方程组。

A−1X+Bλ=B，

b>x=f，

可以用矩阵形式写为

.

我们将在下面的命题41.3中证明，我们的约束极小化问题有一个由上述系统实际给出的唯一解。

注意这个系统的矩阵是对称的。我们的解决方法如下。从第一个方程中减去x

A−1X+Bλ=B，

我们得到

x=a（b−bλ），

代入第二个方程，我们得到

b>a（b−bλ）=f，

也就是说，

b>abλ=b>ab−f。

然而，根据前面的注释，由于a是对称正定的，b的列是线性无关的，b>ab是对称正定的，因此是可逆的。因此我们得到了解决方案

λ=（b>ab）−1（b>ab−f），x=a（b−bλ）。

注意，这种解决系统的方法需要先解决拉格朗日乘数。

假设e=b−bλ，我们还注意到系统



相当于系统

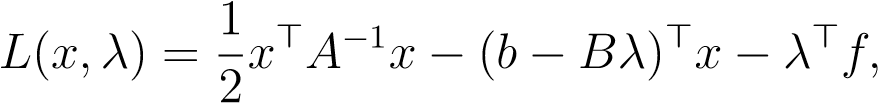
e=b−bλ，

x=ae，b>x=f。

后一个系统被Strang[164]称为平衡方程。事实上，Strang证明了许多物理系统的平衡方程可以用上述形式表示。这包括弹簧质量系统、电网和桁架，它们是由弹性杆建造的结构。在每种情况下，x、e、b、a、λ、f和k=b>ab都有物理解释。矩阵k=b>ab通常称为刚度矩阵。同样，读者也被称为Strang[164]。

为了证明我们的约束极小化问题有一个唯一的解，我们继续证明了受b>x=f约束的q（x）的约束极小化等价于另一个函数−g（λ）的无约束最大化。我们通过最小化拉格朗日L（x，λ）得到g（λ），该拉格朗日L（x，λ）仅被视为x的函数。函数−g（λ）是拉格朗日L（x，λ）的对偶函数。在这里，我们遇到了第49.7节中定义的对偶函数概念的一个特殊情况。

因为a−1是对称正定的，并且



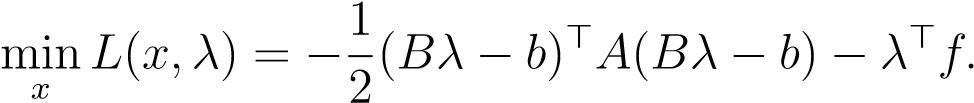
根据命题41.2，得到L（x，λ）的解x的全局最小值（关于x）。

A−1X=B−Bλ，

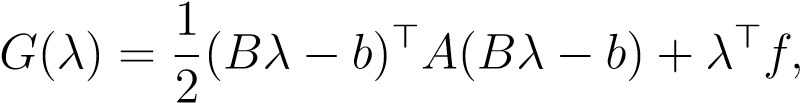
也就是说，当

x=a（b−bλ），

L（x，λ）的最小值为



出租



我们将在命题41.3中证明，受b>x=f约束的q（x）的约束极小化解等于−g（λ）的无约束最大化。这是第49.7节讨论的对偶性的一个特殊情况。

当然，因为我们最小化了l（x，λ）对x的影响，我们有

L（x，λ）≥−g（λ）

所有x和所有λ。但是，当约束b>x=f成立时，l（x，λ）=q（x），因此对于任何容许的x，也就是说b>x=f，我们有

最小值（x）≥最大值−g（λ）。

x·s

为了证明受b>x=f约束问题q（x）的唯一最小值是−g（λ）的唯一最大值，我们计算q（x）+g（λ）。

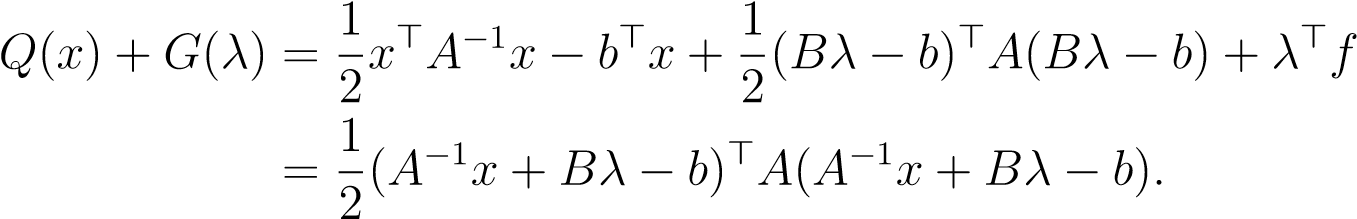
提案41.3.定义41.3的二次约束最小化问题具有由系统给出的唯一解（x，λ）。

.

此外，上述溶液的组分λ是−g（λ）最大的唯一值。

证据。如前所述，假设约束条件

b>x=f保持。去掉f，因为b>x=x>b和λ>b>x=x>bλ，我们得到



因为a是正定的，所以最后一个表达式是非负的。实际上，它是空的iff

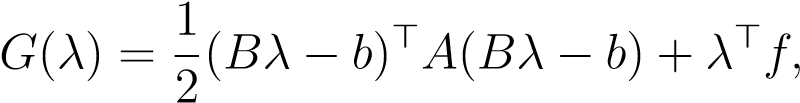
A−1X+Bλ−B=0，

也就是说，

A−1X+Bλ=B。

但是，当b>x=f和a−1x+bλ=b时，q（x）的唯一约束最小值b>x=f等于−g（λ）的唯一最大值，这证明了这个命题。

我们可以确认−g（λ）的最大值，或等于



对应于通过解系统得到的λ值

事实上，自从

而b>ab是对称正定的，通过命题41.2，当

b>abλ−b>ab+f=0，

也就是说，如我们之前发现的，λ=（b>ab）−1（b>ab−f）。

评论：

1. 在这种情况下，存在着一种二元性的形式。受b>x=f约束的q（x）的约束极小化称为原问题，而−g（λ）的无约束最大化称为对偶问题。二元性是指事实稍微松散地表述为

最小值（x）=max−g（λ）。x·s

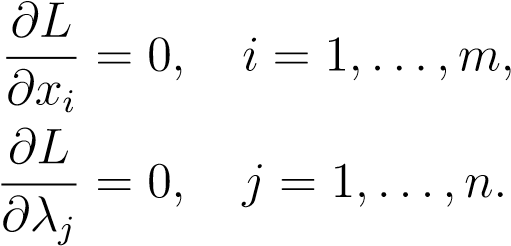
第49.7节给出了约束最小化问题对偶性的一般处理。

回顾e=b−bλ，因为

我们也可以写

这个表达式通常表示系统的总势能。同样，最佳解是使势能最小（从而使−g（λ）最大）的解。

1. 立即证明了命题41.3的方程等价于拉格朗日L（x，λ）的偏导数为空的方程：



因此，受b>x=f约束的q（x）的最小值是拉格朗日L（x，λ）的极值。如我们在命题41.3中所示，这个极值对应于同时使L（x，λ）对X最小化和使L（x，λ）对λ最大化。几何上，这一点是l（x，λ）的鞍点。第49.7节讨论了鞍点。

1. 拉格朗日乘数有时具有自然的物理意义。例如，在弹簧质量系统中，它们对应于节点位移。在一般意义上，拉格朗日乘数是满足平衡方程所需的修正项，也是满足约束条件的代价。有关更多详细信息，请参见Strang[164]。

回到约束最小化）

2x1−x2=5，

拉格朗日是

，

证明拉格朗日有鞍点的方程是

x1+2λ=0，x2−λ=0，2x1−x2−5=0。

我们得到溶液（x1，x2，λ）=（2、−1、−1）。

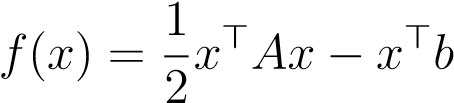
拉格朗日乘子在优化和变分问题中的应用在第49章中进行了广泛的讨论。

最小二乘法和拉格朗日乘数用于解决计算机图形和计算机视觉中的许多问题；见Trucco和Verri[172]、Metaxas[121]、Jain、Katsuri和Schunck[97]、Faugeras[60]和Foley、van Dam、Feiner和Hughes[64]。

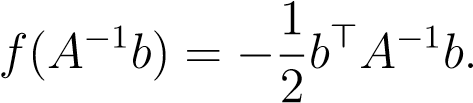
## 41.2二次优化：一般情况

在本节中，我们完成了第41.1节中开始的研究，并给出了二次函数具有全局最小值的必要和充分条件。我们从以下简单的事实开始：

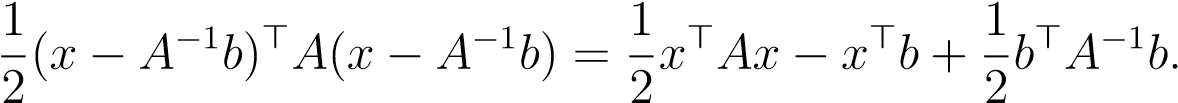
提案41.4.如果a是可逆对称矩阵，那么函数



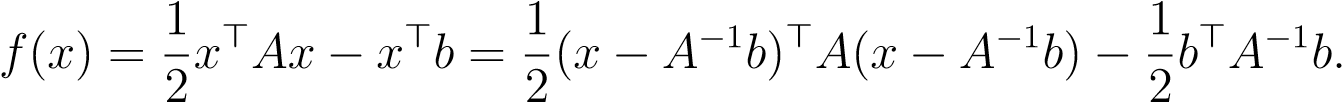
具有一个最小值iff，在这种情况下，该最佳值是针对唯一值x（即x=a−1b）获得的，并且



证据。注意

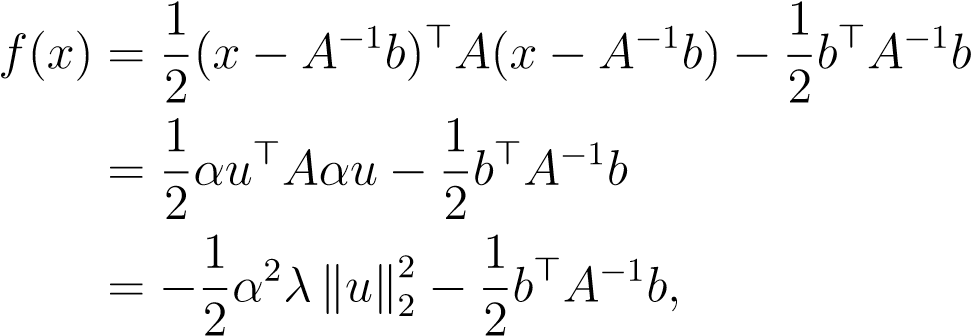


因此，



41.2。二次优化：一般情况

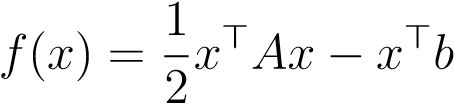
如果a有一些负特征值，比如−λ（λ>0），如果我们选取与λ相关的a的任何特征向量u，那么对于任何α∈r，α=06，如果我们让x=αu+a−1b，那么由于au=−λu，我们得到



因为α可以任意大，λ>0，我们看到f没有最小值。因此，为了使f有一个最小值，我们必须有0，因为a是可逆的，它是肯定的，所以（x−a−1b）>a（x−a−1b）>0 iff x−a−1b=06，很明显，当x−a−1b=0，即x=a−1b时，f的最小值是达到的。

现在让我们考虑一个任意对称矩阵A的情况。

提案41.5。如果a是n×n对称矩阵，那么函数



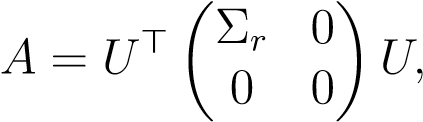
最小值iff和（i−a a+）b=0，在这种情况下，最小值为

此外，如果a被诊断为正交的，则由形式的所有x∈rn得到最优值。

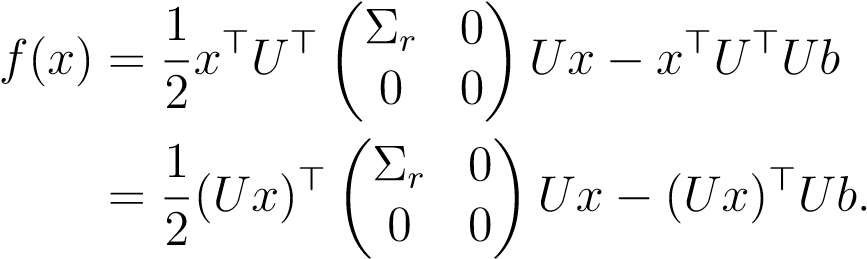
，

对于任何z∈rn−r，其中r是a的秩。

证据。A是可逆的情况由命题41.4处理，因此我们可以假定A是奇异的。如果a的秩r<n，那么我们可以将a对角化为



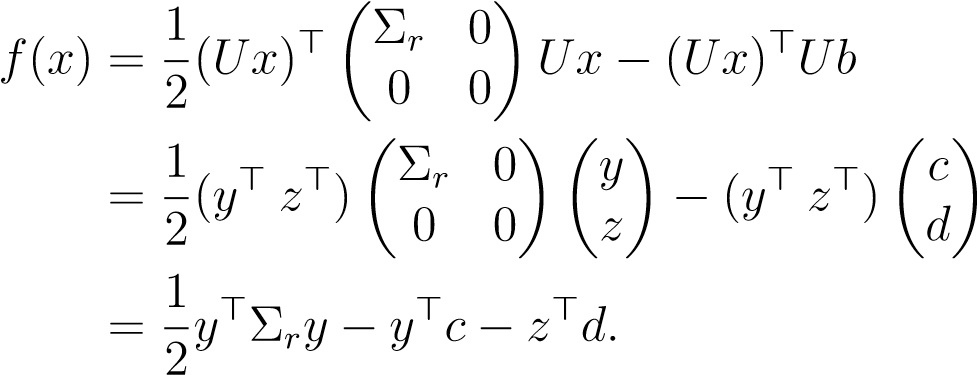
其中u是正交矩阵，∑r是r×r对角可逆矩阵。然后我们有了



如果我们写信

而且，

用y，c∈r r和z，d∈rn−r，我们得到

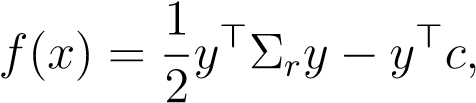


当y=0时，我们得到f（x）=-z>d，

因此，如果d=06，函数f没有最小值。因此，如果f最小，则d=0。但是，d=0意味着

，

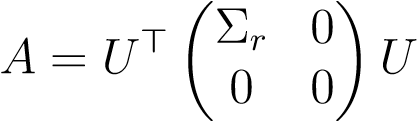
我们从命题21.5（第一卷）知道b在a的范围内（这里，u是v>），相当于（i−aa+）b=0。如果d=0，则



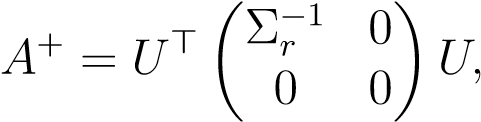
由于∑r是可逆的，根据命题41.4，函数f有一个最小的iff∑0，相当于

因此，我们已经证明，如果f有一个最小值，那么（i−a a+）b=0，反之，如果（i−aa+）b=0和0，我们刚刚做的证明f有一个最小值。

当上述条件成立时，自



是正半定的，a的伪逆a+由下式给出

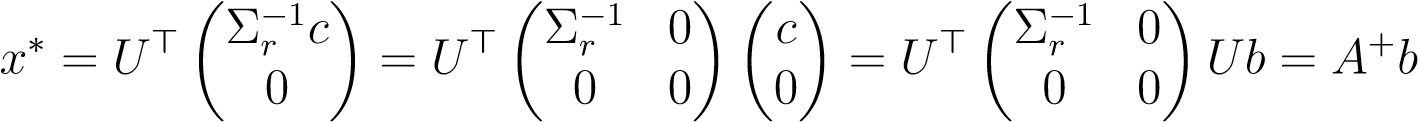


根据命题41.4，当y=∑−r 1c，z=0，d=0时，即x的最小值由

而且，

41.2。二次优化：一般情况

我们由此推断

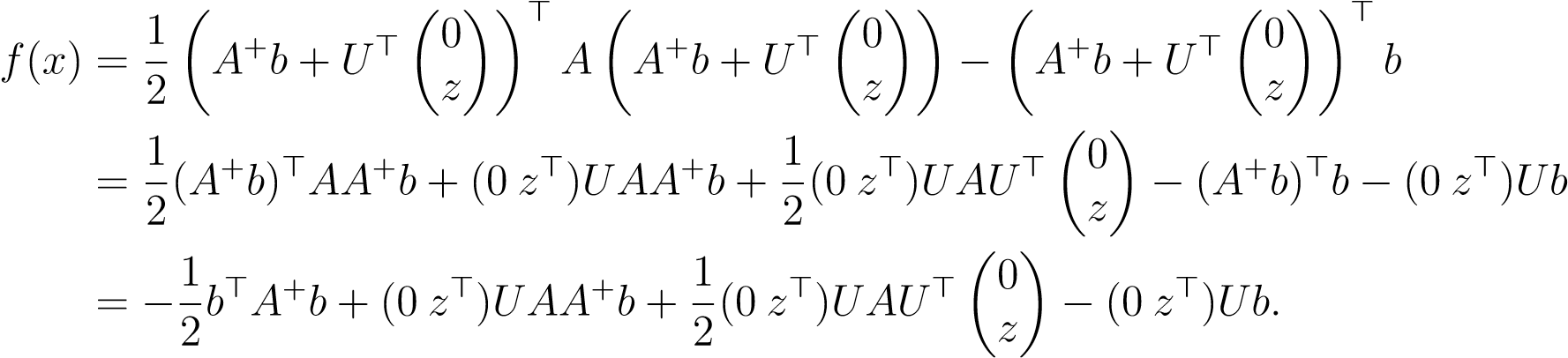


f的最小值是

对于形式的任何x∈rn

，

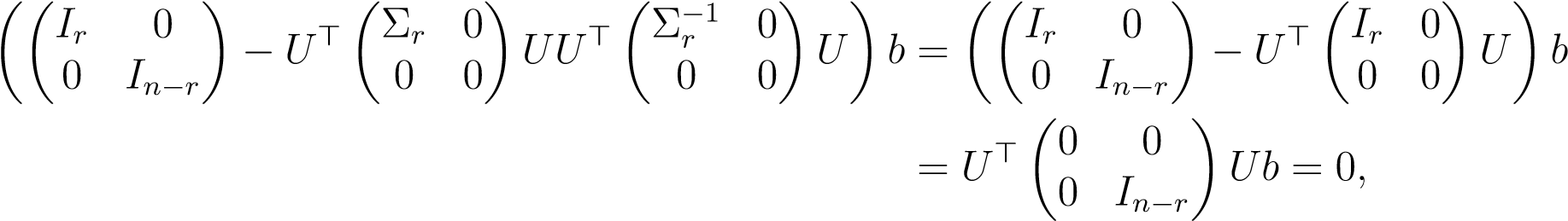
对于任何z∈rn−r，我们有



我们有

和

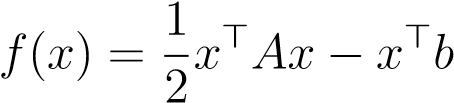
因为（i−aa+）b=0，也就是说，



所以如果

，

然后d=0。因此，最小化函数的问题



如果我们加上形式c>x=0的线性约束或形式c>x=t的仿射约束（其中t∈rm和t=0），6如果c是n×m矩阵，可以用c的qr分解将其简化为不受约束的情况。让我们演示如何对f的线性约束执行此操作。ORM C>X=0。

如果我们使用c的qr分解，通过排列c的列来确保c的前r列是线性无关的（其中r=秩（c）），我们可以假设

，

其中q为n×n正交矩阵，r为r×r可逆上三角矩阵，s为r×m-r矩阵，\_为置换矩阵（c为秩r）。那么如果我们让

，

其中y∈r r和z∈rn−r，则c>x=0变为

，

这意味着y=0，c>x=0的每一个解的形式都是

.

我们原来的问题变成了

减少

服从y=0，y∈r r，z∈rn−r。

因此，约束c>x=0被简化为y=0，如果我们写

，

其中，g11是r×r矩阵，g22是（n-r）×n-r矩阵，以及

，

我们的问题变成了

减少，

提案41.5解决了这个问题。

形式c>x=t（其中t=0）6的约束可以用类似的方式处理。在这种情况下，我们可以假定c是一个n×m矩阵，具有满秩（因此m≤n）和t∈rm。然后我们使用一个形式的QR分解

，

其中p是正交n×n矩阵，r是m×m可逆上三角矩阵。如果我们写信

，

式中y∈rm和z∈rn−m，方程c>x=t变成

（r>0）p>x=t，

也就是说，

会产生

由于r是可逆的，我们得到y=（r>）-1t，然后很容易看出，我们的原始问题根据矩阵p>ap减少到一个不受约束的问题；细节留作练习。

## 41.3在单位球面上最大化二次函数

在本节中，我们讨论了主要由计算机视觉（图像分割和轮廓分组）引起的各种二次优化问题。这些问题可归结为以下基本优化问题：给定一个n×n实对称矩阵a

最大化x>ax，以x>x=1，x∈rn为准。

根据命题21.10（第一卷），单位球面上x>ax的最大值等于矩阵A的最大特征值λ1，并且它是针对与λ1相关的任何单位特征向量U1实现的。

1. 在计算机视觉中经常遇到的上述问题的变种，包括在椭球体上用公式给出的X>AX最小化。

x>bx=1，

其中b是对称正定矩阵。因为b是正定的，所以它可以对角化为

1. =qdq>，

其中q是正交矩阵，d是对角矩阵，

d=诊断（d1，…，dn）

当di>0时，对于i=1，…，n.如果我们通过定义矩阵b1/2和b−1/2

b1/2=qdiag

和

B−1/2=qdiag，

很明显，这些矩阵是对称的，b−1/2bb−1/2=i，b1/2和b−1/2是互逆的。那么，如果我们改变变量

X=B−1/2Y，

方程x>bx=1变为y>y=1，优化问题

|  |  |
| --- | --- |
| 网络错误 | 网络错误 |
| 网络错误  网络错误 | 网络错误 |
| 网络错误 | 网络错误 |
| 网络错误 | 网络错误 |

其中y=b1/2x，其中b−1/2ab−1/2是对称的。

复杂版本的基本优化问题，其中A是一个厄米矩阵也出现在计算机视觉。也就是说，给定一个n×n复厄米特矩阵A，

最大化x ax，以x x=1，x∈cn为准。

同样，根据命题21.10（第一卷），单位球面上x ax的最大值等于矩阵A的最大特征值λ1，并且它是针对与λ1相关的任何单位特征向量U1实现的。

注：值得指出的是，如果a是一个偏厄米矩阵，也就是说，如果a=−a，那么x ax是纯虚数或零。

实际上，因为z=x ax是一个标量，我们有z=z（z的共轭），所以我们有

x ax=（x ax）=x a x=-x ax，

所以x ax+x ax=2re（x ax）=0，这意味着x ax是纯虚数或零。特别是，如果a是实矩阵，如果a是斜对称的，那么

x>ax=0。

因此，对于任何实矩阵（对称与否），

x>ax=x>h（a）x，

其中h（a）=（a+a>）/2，a的对称部分。

|  |  |
| --- | --- |
| 网络错误 |  |
| 网络错误 | 网络错误 |
| 网络错误 | 网络错误 |

在某些情况下，有必要对球体上的二次函数最大化问题添加线性约束。Golub[78]（1973）完全解决了这个问题。问题如下：给定n×n实对称矩阵a

如第41.2节所示，Golub表明线性约束c>x=0可以消除如下：如果我们使用c的qr分解，通过排列列，我们可以假设

，

其中q是一个正交n×n矩阵，r是一个r×r可逆上三角矩阵，s是一个r×r（p-r）矩阵（假设c的秩为r）。那么如果我们让

，

其中y∈r r和z∈rn−r，则c>x=0变为

，

这意味着y=0，c>x=0的每一个解的形式都是

.

我们原来的问题变成了

最小化（

以z>z=1，z∈rn−r，y=0，y∈rr为准。

因此，约束c>x=0被简化为y=0，如果我们写

，

我们的问题变成了

最小化z>g22z，服从z>z=1，z∈rn−r，

标准特征值问题。

注：有一种求g22特征值的方法，它不需要c的qrfactorization。如果我们

然后

，

### 0 G22

如果我们设置

P=Q>JQ，

然后

pap=q>jqaq>jq。

现在q>jqaq>jq和jqaq>j具有相同的特征值，所以pap和jqaq>j也具有相同的特征值。结果表明，我们的优化问题的解在k=pap的特征值中，其中r至少为0。利用CC+是C范围的投影，其中C+是C的伪逆，也可以证明

P=I−CC+，

到c>内核的投影。因此p可以直接用c来计算，特别是当n≥p和c具有全秩（c的列是线性无关的），那么我们知道c+=（c>c）−1c>和

P=I−C（C>C）−1C>。

库尔和施[42]使用了这一事实，于和施[186]也含蓄地使用了这一事实。

在实践中也出现了增加n>x=t形式的仿射约束的问题，其中t=06。乍一看，这个问题似乎并不比t=0的线性问题更困难，但事实确实如此。Gander、Golub和von Matt[75]在一篇论文中对这个问题进行了广泛的研究（1989年）。

Gander、Golub和von Matt考虑了以下问题：给定（n+m）×（n+m）实对称矩阵a（n>0），一个（n+m）×m全秩矩阵n，以及一个非零向量t∈rm（其中（n>）+表示n>的伪逆矩阵），

最小化x>ax

以x>x=1，n>x=t，x∈rn+m为准。

条件1确保问题有一个解决方案，并且不是琐碎的。作者首先证明了仿射约束n>x=t是可以消除的。一种方法是使用n.if的qr分解。

，

其中p是一个正交（n+m）×n+m矩阵，r是一个m×m可逆上三角矩阵，如果我们观察到，如果我们写下

其中b为m×m对称矩阵，c为n×n对称矩阵，\_为m×n矩阵，且

，

用y∈Rm和z∈Rn，得到

x>ax=y>by+2z>\_y+z>cz，

r>y=t，y>y+z>z=1。

因此

y=（r>）-1t，

如果我们写

s2=1−y>y>0

和

B=\_Y，

我们得到了简化的问题

最小化z>cz+2z>b，以z>z=s2，z∈rm为准。

不幸的是，如果b 6=0，21.10号提案（第一卷）不再适用。使用拉格朗日乘子仍然可以求出函数z>cz+2z>b的最小值，但这种解太过复杂，无法在此给出。感兴趣的读者将在Gander、Golub和von Matt[75]中找到一个深入的讨论。

## 41.4总结

本章的主要概念和结果如下：

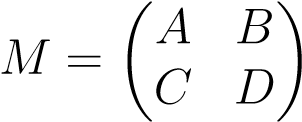
* 二次优化问题；二次函数。
* 对称正定和半正定矩阵。
* 半正定锥序。
* 当a为对称正定时，存在全局极小值。
* 约束二次优化问题。
* 拉格朗日乘数；拉格朗日。
* 原始问题和双重问题。
* 二次优化问题：对称可逆矩阵A的情况。
* 二次优化问题：对称矩阵A的一般情况。
* 添加形式c>x=0的线性约束。
* 添加形式c>x=t的仿射约束，其中t 6=0。
* 在单位球面上最大化二次函数。
* 在椭球上最大化二次函数。
* 最大化埃尔米特二次型。
* 添加形式c>x=0的线性约束。
* 添加形式n>x=t的仿射约束，其中t 6=0。

第四十二章

# 舒尔补充与应用

## 42.1舒尔补充

舒尔补数在形式的块矩阵求逆过程中自然产生。



在描述这些矩阵的对称形式是正定或半正定的时候。这些特征出现在各种二次优化问题中；见Boyd和Vandenberghe[29]，特别是附录B。在最一般的情况下，还需要伪逆。

在这一章中，我们介绍了舒尔的补充语，并描述了几种有趣的使用方法。在此过程中，我们提供了博伊德和范登堡[29]附录A.5（特别是第A.5.5节）中某些结果的一些细节和证明。设m为n×n矩阵，写为2×2块矩阵

，

其中a是p×p矩阵）。我们可以尝试求解线性系统p矩阵，d是q×q矩阵，n=p+q（b是p×q矩阵）

而c是q×。

，

也就是说，

|  |  |  |
| --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 |

一千四百零三

通过模拟高斯消去。如果我们假设d是可逆的，那么我们首先求解y，得到

y=d−1（d−cx）

在第一个方程中用这个表达式代替y之后，我们得到

ax+b（d−1（d−cx））=c，

也就是说，

（a−bd−1c）x=c−bd−1d。

如果矩阵A−bd−1c是可逆的，那么我们就得到了系统的解。

.

如果a是可逆的，那么首先使用第一个方程消除x，我们得到涉及矩阵d−ca−1b的类似公式。上述公式表明矩阵a−bd−1c和d−ca−1b起到特殊作用，并建议以下定义：定义42.1。给定任意n×n形式的分块矩阵

，

其中a是p×p矩阵，d是q×q矩阵，n=p+q（所以b是p×q矩阵，c是q×p矩阵），如果d是可逆的，那么矩阵a−bd−1c在m中称为d的schur补码。如果a是可逆的，那么矩阵d−ca−1b称为i的schur补码。N.

上述方程写为

X=（A−bd−1c）−1c−（A−bd−1c）−1bd−1d，Y=−d−1c（A−bd−1c）−1c

+（D−1+D−1C（A−BD−1C）−1BD−1）D，

根据m中d的舒尔补码，得出m的倒数公式，即

.

一瞬间的思考就可以发现

，

然后

.

通过求逆，我们得到如下结果。

42.1。舒尔补充1405

提案42.1.如果矩阵d是反转的，那么

.

上述表达式可以直接检验，且只要求d的可逆性。

备注：如果a是可逆的，那么我们可以使用a的schur补码d−ca−1b来获得m的下列因式分解：

.

如果d−ca−1b是可逆的，我们可以把上面的三个矩阵都颠倒过来，得到m的逆矩阵的另一个公式（d−ca−1b），即：

.

如果a、d和schur的两个补a−bd−1c和d−ca−1b都是可逆的，通过比较m−1的两个表达式，我们得到（不明显的）公式。

（A−BD−1C）−1=A−1+A−1B（D−CA−1B）−1CA−1。

利用这个公式，我们得到了涉及a和d的舒尔补的m的逆表达式（见Horn和Johnson[92]）：

提案42.2.如果a、d和两个schur互补a−bd−1c和d−ca−1b都是可逆的，那么

.

如果我们设置d=i，将b改为−b，我们得到

（A+BC）−1=A−1−A−1b（I−CA−1b）−1ca−1，

一个称为矩阵反演引理的公式（见Boyd和Vandenberghe[29]，附录C.4，特别是C.4.3）。

## 42.2对称正定矩阵和舒尔补

如果假设我们的块矩阵m是对称的，因此a，d是对称的，c=b>那么我们看到m表示为

，

这表明m类似于块对角矩阵（很明显，schur补码a−bd−1b>是对称的）。因此，我们有以下版本的“schur's trick”来检查对称矩阵是否为0。

提案42.3.对于任意对称矩阵m的形式

，

如果c是可逆的，则以下属性保持不变：

.

，然后。

证据。（1）注意

，

我们知道，对于任何对称矩阵t和任何可逆矩阵n，矩阵t是正定的（显然是对称的）是正定的。

0）。但是块对角矩阵是正定的，如果每个对角块是

正定，即证明的结论。

（2）这是因为对于任何对称矩阵t和任何可逆矩阵n，我们有

tnt>

第42.3号提案的另一个版本也适用，使用a的schur补语代替c的schur补语。证明使用m的因式分解，使用a的schur补码（见第42.1节）。

提案42.4.对于任意对称矩阵m的形式

，

如果a是可逆的，则以下属性保持不变：

42.3。半定矩阵与舒尔补1407

.

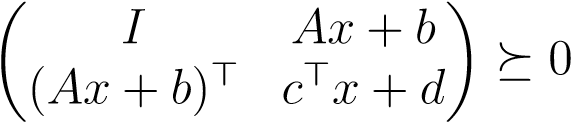
，然后。

这是42.4（2）号提案的一个例子。考虑非线性二次约束

（ax+b）>（ax+b）≤c>x+d，

是a∈mn（r），x，b，c∈rn和d∈r，由于显然i=in是可逆的，所以我们

有



如果c>x，因为矩阵（标量）c>x+d−（ax+b）>（ax+b）是上述矩阵中i的舒尔补码。

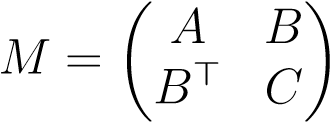
利用Schur互补将非线性不等式约束转化为包含半定义序的对称矩阵的线性约束的技巧被广泛地用于将非线性问题转化为半定程序；见Boyd和Vandenberghe。

〔29〕。

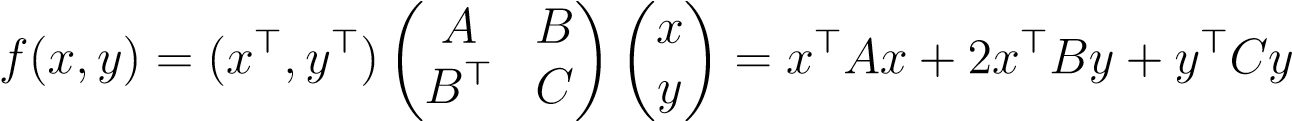
当c是奇异的（或a是奇异的）时，当上面的对称矩阵m是半正定的时，仍然可以描述它，但这需要使用包含c的伪逆的schur补码版本，即a−bc+b>（或a的schur补码，c−b>a+b）。我们使用命题41.5的标准，它告诉我们形式的二次函数何时有最小值，以及这个最佳值是什么（其中p是对称矩阵）。

## 42.3对称半正定矩阵和舒尔补

现在我们回到原来的问题，描述一个对称矩阵



是正半定的。因此，我们想知道函数



对于x和y都有一个最小值。如果我们保持y常数，则命题41.5意味着f（x，y）的最小iff为0且（i−a a+）为0，则最小值为

f（x，y）=−y>b>a+by+y>cy=y>（c−b>a+b）y。

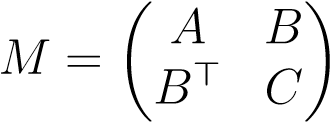
因为我们希望f（x，y）在所有x，y下都是一致的有界的，所以我们必须有（i−

aa+）b=0.现在，f（x，y）的最小iff为0。因此，我们确定f（x，y）在所有x，y iff上都有一个最小值。

.

如果我们首先对y最小化，然后对x最小化，那么也适用类似的推理，但这一次涉及到c的schur补码a−bc+b>。把这些事实综合起来，我们得到了主要的结果：

定理42.5。给定任意对称矩阵



以下条件等效：

是正半定的）。

.

如果0如定理42.5所示，那么很容易检查我们是否有以下的因式分解（使用a+aa+=a+和c+cc+=c+的事实）：

和

.

第七部分

线性优化

一千四百零九

# 第43章凸集，锥，H多面体

43.1什么是线性规划？

什么是线性规划？乍一看，人们可能会认为这是某种类型的计算机编程。毕竟，有命令式编程、函数式编程、面向对象编程等。线性编程这个术语有点误导，因为它实际上是指一种带有线性约束的规划方法，或者更准确地说，是一种优化方法，在这种方法中，目标和目标都是NCtion和约束是线性的。

线性规划是在20世纪40年代末创建的，其中一个关键的参与者是发明了simplex算法的George Dantzing。早在1939年，Kantorovitch就在线性编程方面做了一些开创性的工作。线性规划这一术语具有军事内涵，因为在20世纪50年代早期，它被用作训练部队、后勤供应、资源分配等计划或时间表的同义词。不幸的是，线性规划这一术语已经很好地建立起来，我们一直坚持使用它。

有趣的是，尽管线性规划最初大部分应用于经济学和工业工程领域，但线性规划已成为理论计算机科学和算法理论中的一个重要工具。实际上，线性规划通常是设计近似算法以解决困难问题（通常是NP困难问题）的有效工具。线性规划也是凸规划的“婴儿版”，是近年来备受关注的一种非常有效的方法。

我们在这些笔记中的目标是呈现线性规划的数学基础，特别是如果线性规划是可行和有界的，最优解的存在性，以及线性规划中的对偶定理，这是这一领域最深的成果之一。线性规划中的对偶定理也有重要的算法含义，但我们这里不讨论这个问题。我们提出了单纯形算法、双单纯形算法和原对偶算法。我们还描述了舞台形式主义

一千四百一十一

用于运行simplex算法及其变体。Tableau形式主义的一个特别好的特点是，可以使用基本行操作来更新Tableau，这与将矩阵缩减为行的梯队形式（rref）过程中使用的操作相同。不同的是选择轴的标准。

但是，我们不讨论其他方法，如椭球法或内点法。对于更多的算法问题，我们向读者介绍线性规划的标准文本。在我们看来，是最清晰的（也是最简洁的！）马托塞克和加德纳[120]；契瓦塔[40]和施里伊弗[144]是经典之作。Papadimitriou和Steiglitz[130]在组合优化的更广泛的背景下提供了非常清晰的介绍，Bertsimas和Tsitiklis[21]和Vanderbei[175]非常完整。

线性规划必须与最大化线性成本函数c1x1+·····+cnxn有关形式的m“线性”不等式。

ai1x1+·····+ainxn≤bi。

这些约束可以组合成一个m×n矩阵a=（aij），并更简洁地写为

ax≤b。

由于稍后出现的技术原因，通常最好在i＝1，n，n上加上非负乘积Xi±0，写x＝0。很容易看出，每一个线性程序都等价于另一个满足约束x≥0的程序，而代价是添加新的变量，这些变量也被约束为非负的。设p（a，b）为线性规划的可行解集，由

p（a，b）=x∈rn ax≤b，x≥0。

然后有两个基本问题：

1. p（a，b）是非空的，也就是说，我们的线性规划有机会得到一个解吗？
2. 目标函数c1x1+·····+cnxn在p（a，b）上有最大值吗？

这两个问题的答案都可以是否定的，但如果p（a，b）是非空的，并且目标函数在p（a，b）上有界，则可以证明，c1x1+······+cnxn的最大值是由x∈p（a，b）得到的。这种解称为最优解。也许令人惊讶的是，这个结果并不那么容易证明（除非有人使用单纯形方法作为处理方法）。我们将详细证明这一结果（见提案44.1）。

线性约束之所以如此重要，是因为潜在最优解P（A，B）的域是凸的。实际上，p（a，b）是一个凸多面体，它是仿射超平面截取的半空间的交集。线性的目标函数是凸的，这也是一个重要的事实。因此，我们研究了凸集，特别是那些由仿射形式定义的不等式解产生的凸集，同时也研究了凸锥。

43.2。仿射子集、凸集、超平面、半空间

我们简要介绍了这些主题。作为奖励，我们提供了几个标准来检验一个不平等系统是否

ax≤b，x≥0

是否有关于法卡斯引理的解决方案（见命题49.3和46.4）。然后我们给出了线性规划强对偶定理的完全证明（见定理46.7）。我们还讨论了互补松弛条件，并证明了它们可以用来设计一种同时使用原问题和对偶问题的线性规划求解算法。这种被称为原始对偶算法的算法，尽管目前没有被广泛使用，但它一直是一类近似算法（也被称为原始对偶算法）的灵感来源。

我们希望这些笔记将成为学习更多线性规划、凸优化以及凸几何的动力。凸面优化中的“圣经”是Boyd和Vandenberghe[29]，凸面几何的最佳来源之一是Ziegler[189]。这是一个相当高级的文本，因此读者可能希望从Gallier开始[74]。

## 43.2仿射子集、凸集、仿射超平面、半空间

我们认为RN由列向量（n×1矩阵）组成。与往常一样，行向量表示线性形式，即线性映射，如果所有x∈r n的行向量y（a 1×n矩阵）表示线性形式，则表示线性形式。我们用（rn）表示线性形式（行向量）的空间。

回想一下，RN中向量的线性组合是一个表达式

λ1x1+····+λmxm

其中，X1，…，XM\*Rn，其中，La 1，…，Lm m是R中的任意标量，给定一个向量序列S＝（x1，…，xm），其中，S中的向量的所有线性组合的集合是包含S的线性跨度的S的最小（线性）子空间，并表示跨度（S）。RN的线性子空间是线性组合下封闭的RN的任何非空子集。RN中向量的仿射组合是一个表达式。

λ1x1+····+λmxm

式中，x1，…，xm∈rn，式中，λ1，…，λm是满足条件的r中的标量。

λ1+····+λm=1.

给定一个向量序列s=（x1，…，xm）的序列，在XS中，S中所有向量的仿射组合集合是包含S的仿射壳的最小仿射子空间，并表示为AFF（S）。

（a）（b）

图43.1：（a）凸集；（b）非凸集

定义43.1.RN的仿射子空间A是仿射组合下封闭的RN的任何子集。

如果a是rn的非空仿射子空间，则可以证明va=a−b a，b∈a是rn的线性子空间，称为a的方向，并且

a=a+va=a+v v∈va

对于任意a∈a，非空仿射子空间a的维数是其方向va的维数。

凸组合是满足i=1，…，m时λi≥0的附加条件的仿射组合。

定义43.2.Rn的一个子集v是凸的，如果对于任意两点a，b∈v，对于每一点c=（1−λ）a+λb，我们有c∈v，其中0≤λ≤1（λ∈r）。对于任意两点a，b，符号[a，b]通常用于表示a和b之间的线段，即，

[a，b]=c∈rn c=（1−λ）a+λb，0≤λ≤1，

因此，当任意两点a，b∈v（a=b）允许时，集合v是凸的。凸集V的维数是它的仿射壳aff（a）的维数。

空集是平凡的凸集，每一个点集a是凸集，整个仿射空间rn是凸集。

显然，任何凸集族（有限或无限）的交集都是凸的。

定义43.3.给定RN的任何（非空）子集，包含s的最小凸集用conv（s）表示，称为s的凸壳（它是包含s的所有凸集的交集）。

43.2。仿射子集、凸集、超平面、半空间

不仅要对conv有很好的理解，而且要有很好的计算方法。我们有以下简单但关键的结果。

命题43.1.pλi a i（其中对于任何家族i∈iλi=1sand=（aλi）ii≥∈i 0of points in）是凸的hullrn，s=（ai）i∈i的setconv（v of凸包combi-）。

国家I∈I

很自然地会想，43.1命题是否可以在两个方向上进行锐化：（1）是否可以在凸组合中涉及的点的数量上有一个固定的界限？（2）是否有必要考虑所有点的凸组合，或仅考虑具有特殊性质的子集？

在这两种情况下，答案都是肯定的。在案例1中，Carath'Eodory定理断言它足以考虑n+1点的凸组合。例如，在平面r2中，一组点的凸壳是所有三角形（包括内部点）与S中顶点的结合。在案例2中，Krein和Milman的定理断言，也紧凑的凸集是其极值的凸壳（给定凸集s，如果S−A也是凸的，则点A∈S是极值。

我们不会在这里证明这些定理，但我们邀请读者咨询Gallier[74]或Berger[12]。

凸集也作为仿射超平面切出的半空间出现。

定义43.4.一个仿射式，由一些线性形式c∈（rn）和一些标量β∈r定义，因此

⑨（x）=cx+β，所有x∈rn。

如果c=06，由（c，β）指定的仿射式瓒定义仿射超平面（对于短超平面）h（瓒），由

h（\_）=x∈rn \_（x）=0=x∈rn cx+β=0，

两个（封闭的）半空间

.

当β=0时，我们称H为线性超平面。

H+（\_）和H−（）均为凸形，H=H+（）H−（）。

例如，\_：r2→r，其中\_（x，y）=2x+y+3是一个仿射形式，定义了方程y=−2x−3给出的直线。仿射形式的另一个例子是：\_:r3→r，其中\_（x，y，z）=x+y+z−1；该仿射形式定义了方程x+y+z=1给出的平面，即通过点（0,0,1）、（0,1,0）和（1,0,0）的平面。这两个超平面如图43.2所示。

y

=

-

2

x

-

3

(0

,0,

1)

(1

,0,

0)

(0

,1,

0)

x + y + z = 1

*H*

*H*

+

*H*

+

\_

*H*

\_

I.I.

图43.2：图1.说明了图2所示的超平面h（）对于\_（x，y）=2x+y+3。说明了\_（x，y，z）=x+y+z-1的超平面h（）。

对于x＝（x1，…，xn）和y＝（y1，…，yn）的任意两个向量x，y[r] rn，我们写x＝y y ff Xiωy，i＝1，…，n，x x y y y y y y×x，特别是x＝0，i＝i＝0，i＝1，…，n。

某些特殊类型的凸集称为锥和H多面体起着重要作用。线性规划的可行解集是H多面体，锥在44.1命题的证明和Farkas-Minkowski命题（命题）中起着至关重要的作用。

46.2）。

## 43.3锥体、多面体锥体和H多面体

锥体和多面体锥体定义如下。

定义43.5.给定一个非空子集s rn，由s横跨的圆锥体c=圆锥体是凸集。

锥体，（

从S来的向量的正组合。如果S由一组有限的向量组成，则C=锥称为多面体锥。图43.3显示了多面体圆锥体。

注意，如果某个非零向量u属于一个锥c，那么对于所有的λ≥0，则λu∈c，也就是说，射线λuλ≥0属于c。

注：定义43.5的锥体（和多面体锥体）总是凸的。因此，我们使用了更简单的术语“锥”而不是“凸锥”。然而，有更多非凸的圆锥（例如多面体圆锥的联合）

1)

,0,

(1

1)

(0

,0,

1)

,1,

(1

1)

,1,

(0

(1

,0,

1)

,0,

(0

1)

(1

,1,

1)

(0

,1,

1)

S

cone(S)

图43.3：设s=（0,0,1），（1,0,1），（1,1,1），（0,1,1）。多面体圆锥体，圆锥体，是一个实心“金字塔”，其顶点位于原点，横截面为正方形。

或者由图43.4中的曲线生成的线性圆锥），如果我们处理这些，我们将把定义43.5中的圆锥称为凸锥。

定义43.6.H多面体，简称多面体，是定义为有限个闭半空间Ci的交集的任何子集。一个例子

如图43.6所示，有一个闭球多面体的艾哈迈恩例子。中心多面体43.5.x和半径的图b（x）中显示了Anh多面体，其有界r>0，因此h多面体，其中p br（x）。

根据惯例，我们同意RN本身是一个H多面体。

注：定义43.6的H多面体总是凸的。因为这个原因，就像在多面体而不是凸的H多面体中一样。圆锥的情况下，我们使用简单的术语h

在代数拓扑学中，有更多的非凸多面体。

结果表明，H-多面体P等于有限多点（P的极值点）的凸壳。这是一个非常重要的结果，其证明需要大量的工作；见Gallier[74]和Ziegler[189]。

无界H多面体不等于有限点集的凸壳。为了得到一个等价的概念，我们引入了V多面体的概念。

定义43.7.v多面体是形式的任何凸子集a rn。

a=conv（y）+cone（v）=a+v a∈conv（y），v∈cone（v），

(0

,0,

1)

S

cone(S)

图43.4:z=1的平面曲线。s的线性圆锥体，由连接s到原点的所有半射线组成，不是凸形的。

其中y rn和v rn是有限的（可能是空的）。

当v=∅时，我们只得到一个多面体，当y=∅或y=0\_时，我们只得到一个圆锥体。

可以看出，每个H多面体都是V多面体，反之亦然。这是多面体理论中的一个重要定理，它的证明是非平凡的。完整的证据见Gallier[74]和Ziegler[189]。

每个多面体圆锥都是闭合的。这是一个重要事实，用于证明其他几个关键结果，如命题44.1和Farkas–Minkowski命题（命题46.2）。

虽然多面体圆锥应该是闭合的，但严格的证明并不完全是微不足道的。

实际上，多面体圆锥体是闭合的这一事实，关键在于C是由有限数量的向量构成的，因为无限集生成的圆锥体可能不会闭合。例如，考虑圆心（0,1）和半径1的闭圆盘d r2，它与原点的x轴相切。然后，圆锥体（d）由打开的上半平面加上原点（0,0）组成，但该集合不是闭合的。

提案43.2.每个多面体锥c都是闭合的。

证据。证明了这一点

1. 每个原始圆锥体都是闭合的。

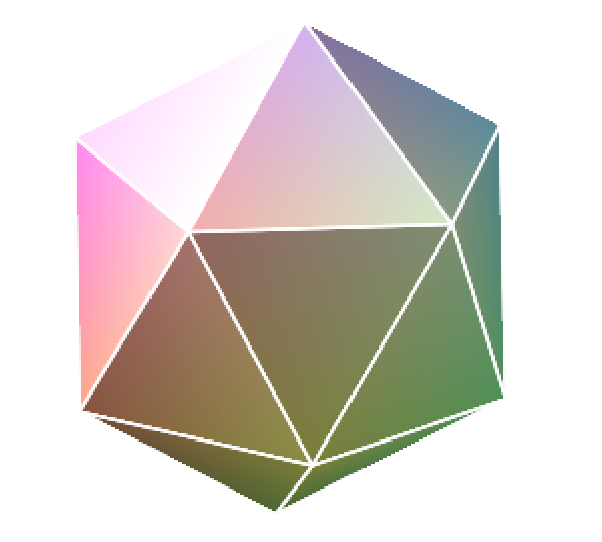
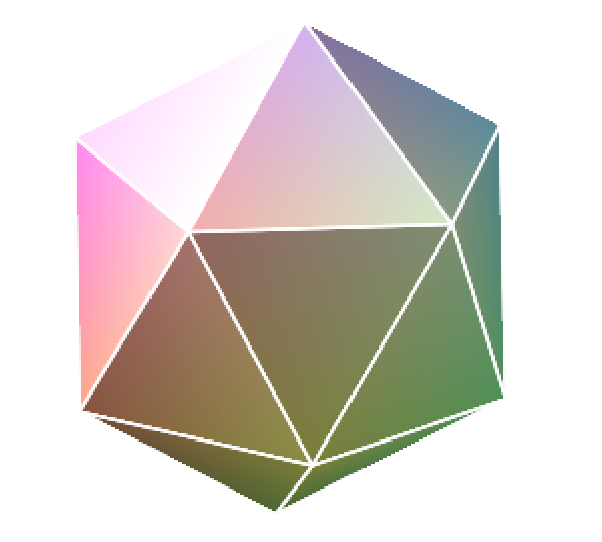
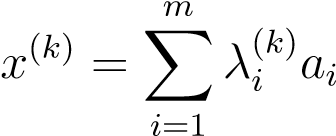


图43.5：二十面体是H-多面体的一个例子。

1. 多面体圆锥体c是有限多个原始圆锥体的联合，其中原始圆锥体是由线性无关向量跨越的多面体圆锥体。

假设（a1，…，am）在rn中是线性无关的向量，并考虑任何se-

频率（x（k））k≥0



原始圆锥体（a1，…，am）中的向量，这意味着

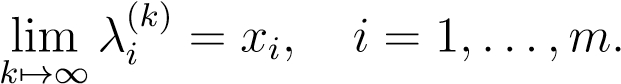
1，…，m和所有k≥0。向量x（k）属于（a1，…，am）所跨越的子空间u，u是闭合的。假设序列（（k）∈u对于所有k≥0，我们有xx（∈k））uk≥。如果我们写0收敛到一个极限x=x1a1+x········r+n.sincexmam，weu是闭的，x

希望证明i＝1，…，m的Xi超0，序列（x（k））k＝0收敛到x。

敌我识别

敌我识别

敌我识别



由于所有k＝0，对于I＝1，…，M，SO x锥（{A1，…，AM}），我们都有Xi±0。

接下来，假设x属于多面体锥c。考虑正组合。

x=λ1a1+····+λkak，（1）

对于一些非零a1，…，ak∈c，其中λi≥0且k最小。因为k是最小的，所以i=1，…，k时我们必须有λi>0。我们声称（a1，…，ak）是线性无关的。

（2,0,0）转换（Y）（-2,0,0）

2

-2

x

y

y + z = 0

c

one

(

V

)

c

o

n

v

(

Y

)

c

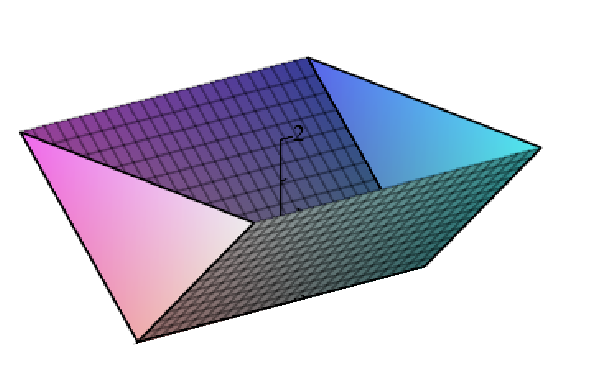
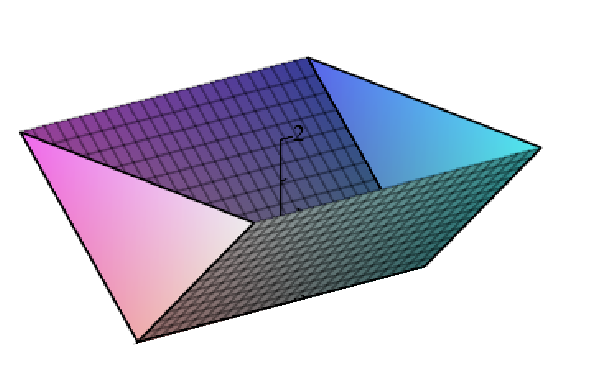
one

(

V

)

+



y - z = 0

y + z = 0

(0

,1,

1)

1)

,-1,

(0

图43.6：由不等式y−z≤0、y+z≥0和−2≤x≤2确定的“三角形槽”是H多面体和V多面体，其中y=（2,0,0）、（−2,0,0）和v=（0,1,1）、（0、−1,1）。

如果不是，有一些非平凡的线性组合

1+·····+k=0，（2）

由于ai不为零，对于一些至少一些j，μj=06。我们可以假设对于一些j，μj<0（否则，我们考虑族（−μi）1≤i≤k），那么让

j=j∈1，…，k \_礹j<0。

对于任何t∈r，由于x=λ1a1+····+λkak，使用（2）我们得到

，（3）

如果我们选择

对于i=1，…，k，我们有（λi+tμi）≥0，但是对于一些j∈j，λj+tμj=0，因此（3）是x的表达式，其系数小于k非零，包含k的最小值（1）。

因此，（A1，…，AK）是线性无关的。

由于多面体的锥C是由有限多个向量构成的，因此有有限多个原始锥（对应于线性无关的子族），并且由于每个X∈C都属于某个原始锥，因此C是有限个原始锥的并集。由于每个原始圆锥体都是封闭的，作为有限多个闭集的联合，C本身是封闭的。

上述事实也在马托塞克和加德纳[120]中得到证明（第6章第5节，引理6.5.3、6.5.4和6.5.5）。

证明多面体锥C是闭合的另一种方法是证明C也是H多面体。这需要更多的工作；见Gallier[74]（第4章，第4节，提案4.16）。LAX[110]中给出了另一个证明（第13章，定理1）。

1422第43章。凸集，锥，H多面体

第四十四章

# 线性程序

## 44.1线性规划、可行解、最优解

线性规划的目的是解决以下类型的优化问题。

定义44.1.线性规划（P）是一类优化问题：

最大化cx，以a1x≤b1为准

…

amx≤bm x≥0，

其中x∈rn，c，a1，…，am∈（rn），b1，…，bm∈r。

线性形式c定义了线性规划（p）的目标函数x 7→cx（从rn到r），其中的不等式aix≤bi和xj≥0称为线性规划（p）的约束条件。

如果我们定义m×n矩阵，其行是行向量a1，…，am和b作为列向量

b1

B=……，BM

一千四百二十三

m不等式约束aix≤bi可以写成矩阵形式

ax≤b。

因此，线性程序（P）也可以表示为线性程序（P）：

最大化cx，以ax≤b和x≥0为准。

下面是（p）类型线性程序的显式示例：

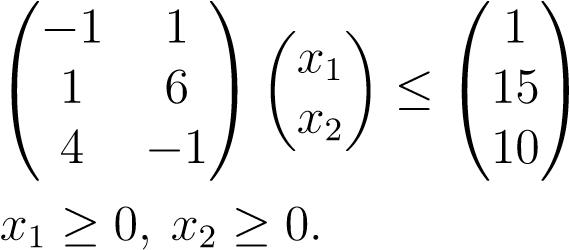
例44.1。

最大化x1+x2取决于

x2−x1≤1 x1+6x2≤15 4x1−x2≤10 x1≥0，x2≥0，

以矩阵形式

最大化受试者



结果表明，x1=3，x2=2产生目标函数x1+x2的最大值，即5。如图44.1所示。注意，满足上述约束条件的点集是由方程线确定的半平面切出的凸区域。

x2−x1=1

x1+6x2=15

4x1−x2=10 x1=0 x2=0。

44.1。线性规划、可行解、最优解

K

1

0

1

2

3

4

5

K

1

1

2

3

4

x

+

6

x

=

1

5

1

2

2)

,

(3

x

-

x

=

1

2

1

4

x

-

x

=

1

0

1

2

图44.1：与实施例44.1相关的H多面体。绿点（3，2）是唯一的最优解。

一般来说，每个约束Ax≤Bi对应于由i（x）=Ax−Bi给出的仿射形式i，并定义半空间h−（i），每个不等式xj≥0定义半空间h+（xj）。这些半空间的交集是所有这些约束的解集。它是一个（可能是空的）H多面体，表示P（A，B）。

定义44.2.如果p（a，b）=∅，我们说线性规划（p）没有可行解，否则任何x∈p（a，b）都称为（p）的可行解。

示例44.2中所示的线性程序，通过逆转示例44.1中线性程序中的不等式x2−x1≤1和4x1−x2≤10的方向获得，没有可行的解；见图44.2。

例44.2。

最大化x1+x2取决于

x1−x2≤−1 x1+6x2≤15 x2−4x1≤−10 x1≥0，x2≥0。

假设P（A，B）6为∅，使线性规划（P）有一个可行解。在这种情况下，考虑目标函数x 7→cx下p（a，b）的图像cx∈r x∈p（a，b）。

定义44.3.如果集合cx∈r x∈p（a，b）在上面是无界的，那么我们就说线性程序（p）是无界的。

K

1

0

1

2

3

4

5

K

1

1

2

3

4

5

x

+

6

x

=

1

5

1

2

x

-

x

=

1

2

1

4

x

-

x

=

1

0

1

2

图44.2：由于蓝色和紫色区域不重叠，因此没有与实施例44.2相关的H多面体。

示例44.3中所示的线性程序是从示例的线性程序中获得的。

44.1通过删除限制条件4x1−x2≤10和x1+6x2≤15，无边界。

例44.3。

最大化x1+x2取决于

.

x∈potherwise，我们将很快证明，如果（a，b），那么有一些p∈p（a，b），那么这是集合cx∈r的最小上界。|

Cp=μ，

也就是说，某个p∈p（a，b）.x→7cx的目标函数在p（a，b）上有一个最大值μ，达到了

定义44.4。（或最优，如果集合cx=∈maxr xcx∈p∈r（a，b x）∈p是非空且在上面有界的，则任何（a，b）称为；例如，最优解p。点p∈p（）a，bof）（psuch that）。最优解通常用上cp\_表示。

例44.1的线性规划有一个唯一的最优解（3,2），但观察到例44.4中目标函数为（1/6）x1+x2的线性规划有无穷多的最优解；目标函数的最大值为15/6，出现在橙色bo的点上。图44.1中的日期线。

44.1。线性规划、可行解、最优解

例44.4。

最大化受试者

x2−x1≤1 x1+6x2≤15 4x1−x2≤10 x1≥0，x2≥0。

证明如果集合cx∈r x∈p（a，b）是非空的且在上面有界的，那么有一个最优解p∈p（a，b），并不像它看起来那样简单。这取决于多面体圆锥是闭合的这一事实，如第43.3节所示。

我们还使用了一个使证明更简单的技巧，即具有不等式约束ax≤b的线性程序（p）

最大化cx，以ax≤b和x≥0为准，

等价于具有相等约束的线性程序（p2）

最大化C x BB

服从且x≥0，

乙

其中ab是m×（n+m）矩阵，bc是（rn+m）中的线性形式，xb∈rn+m，由

，和，

用x∈rn和z∈rm表示。

确实，0 iff

|  |  |
| --- | --- |
| 网络错误  网络错误 | 网络错误 |
| 网络错误 | 网络错误 |

C x=cx。BB

变量z称为松弛变量，形式（p2）的线性程序称为标准形式的线性程序。

将示例44.4的线性程序转换为标准格式的结果是示例44.5所示的程序。

例44.5。

最大化受试者

x2−x1+z1=1 x1+6x2+z2=15 4x1−x2+z3=10 x1≥0，x2≥0，z1≥0，z2≥0，z3≥0。

我们现在可以证明，如果一个线性规划有一个可行解并且有界，那么它就有一个最优解。

提案44.1.设（p2）为标准形式的线性规划，等式约束a x=b，如果p（a，b）是非空的，且在其上有界，如果μ是集合cx∈r x∈p（a，b）的最小上界，则有一些p∈p（a，b），这样

Cp=μ，

也就是说，目标函数x 7→cx在p（a，b）上有一个最大值μ，这是由一些最优解p∈p（a，b）实现的。

证据。由于μ=sup cx∈r x∈p（a，b），有一个向量的序列（x（k））k≥0

x（k）∈p（a，b），使limk7→∞cx（k）=µ。尤其是，如果我们写

我们有并且对于所有的k≥0。设ae为（m+1）×n矩阵

，

并考虑向量axe（k）的序列（axe（k））k≥0∈rm+1。我们有

，

由于假设x（k）∈p（a，b），约束条件为ax=b，x≥0。由于假设limk7→∞cx（k）=µ，序列（axe（k））k≥0收敛于矢量。现在，观察每个矢量轴（k）可以写成凸组合

，

其中，aej∈rm+1为ae的jth列。因此，axe（k）属于多先驱锥

C=锥形（，

由于这个锥是封闭的，所以limk≥∞axe（k）∈c，这意味着存在一些u∈rn，其中u≥0，这样

，

也就是说，cu=μ，au=b。因此，u是（p2）的最佳溶液。

下一个问题是，我们如何找到这样一个最优解？结果表明，对于标准形式的线性规划，约束形式为a x=b和x≥0时，总是存在一种特殊类型的最优解，称为基本可行解。

## 44.2基本可行解和顶点

如果系统ax=b有一个解，如果a的某一行是其他行的线性组合，那么相应的方程是多余的，因此我们可以假设a的行是线性无关的；也就是说，我们可以假设a的列是m，因此m≤n。

如果a是m×n矩阵，对于1，…，n的任何非空子集k，让ak是a的子矩阵，a的列的索引属于k。我们用aj表示矩阵a的jth列。

定义44.5.给定线性程序（p2）

最大化cx，以ax=b且x≥0为准，

在a有秩m的情况下，向量x∈rn是（p）的基本可行解，如果x∈p（a，b）=6∅，并且如果存在尺寸m的1，…，n的某个子集k，则

1. 矩阵AK是可逆的（也就是说，AK的列是线性无关的）。
2. xj=0表示所有j/∈k。

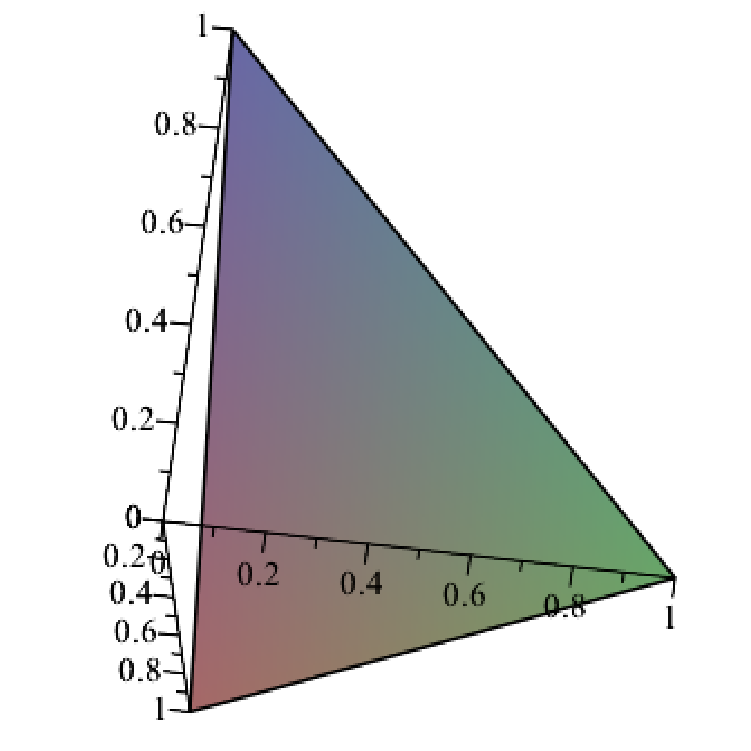
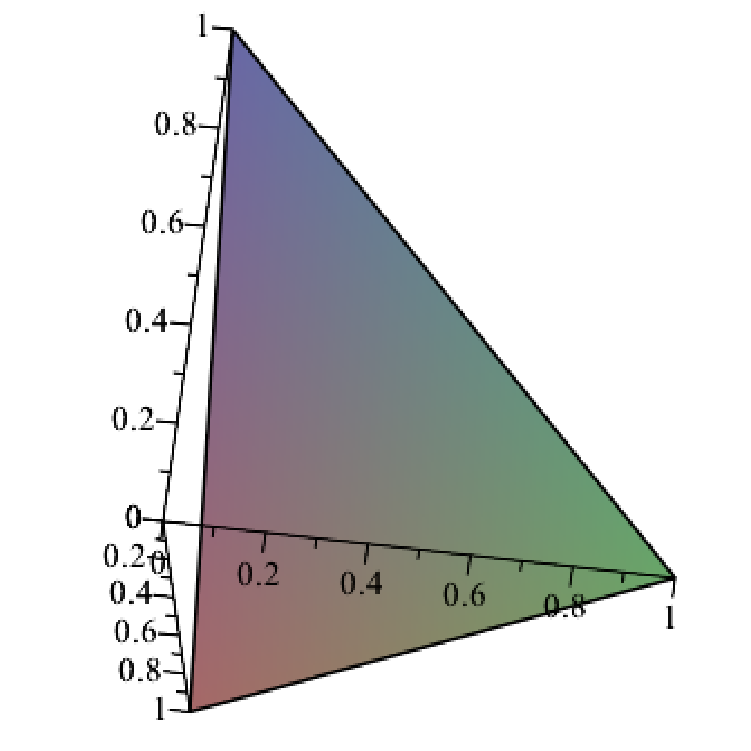
子集k称为x的基，每一个索引k∈k称为基，每一个索引j/∈k称为非基。同样，对应于指数k∈k的列ak称为基本列，对应于指数j/∈k的列aj称为非基本列。与基本指标k∈k对应的变量称为基本变量，与指标j/∈k对应的变量称为非基本变量。

例如，线性程序

最大化x1+x2

x1+x2+x3=1且x1≥0，x2≥0，x3≥0，（）

有三个基本可行解：基本可行解K=1对应点（1,0,0）；基本可行解K=2对应点（0,1,0）；基本可行解K=3对应点（0,0,1）。每个点对应图44.3所示的倾斜紫色三角形的顶点。顶点（1，0，0）和（0，1，0）以1的值优化目标函数。



x

+

y

+

z

=

1

x

+

y

=

0

.

7

图44.3：与线性程序（）相关的H-多面体。目标函数（x1→x和x2→y）由平行于紫色平面x+y=0.7的垂直平面表示，当x+y=1时达到最大值。

我们现在证明，如果定义44.5中的标准线性规划（p2）有一些可行解，并且在上面有界，那么一些基本可行解就是一个最优解。我们遵循马托塞克和加德纳[120]（第4章，第2节，定理4.2.3）。

首先，我们得到了一个更方便的基本可行解的表征。

提案44.2.给定任意标准线性规划（p2），其中a x=b，a是秩m的m×n矩阵，对于任意可行解x，如果j>=j∈1，…，n xj>0，那么x是一个基本可行解，如果矩阵aj>的列是线性无关的。

证据。如果x是一个基本可行解，那么M的子集k 1，…，n使得ak的列是线性无关的，并且x j=0表示所有j/∈k，因此根据定义，j>k，这意味着矩阵aj>的列是线性无关的。

相反，假设x是一个可行的解，使得矩阵aj>的列是线性无关的。如果j>=m，我们就完成了，因为我们可以选择k=j>，那么x是一个基本可行的解决方案。如果j>m，我们可以通过添加m−j>列索引将j>扩展到m元素子集k，以便ak的列是线性独立的，这是可能的，因为a具有秩m。

其次，我们证明了如果一个标准形式的线性规划有任何可行解X0，并且在其上有界，那么它就有一些基本可行解Xe，在cxe≥cx0的意义上，它和X0一样好。

提案44.3.设（p2）为目标函数为cx的任意标准线性规划，其中ax=b，a是m×n阶矩阵，如果（p2）在上有界，如果（p2）是（p2）的某个可行解，则存在一些基本可行解xe，使cxe≥cx0。

证据。在可行的解x中，cx≥cx0（其中x0是其中之一）选择一个最大坐标数xj等于0的解，比如xe。让

k=j>=j∈1，…，n xej>0

让我们=k。我们认为XE是一个基本可行的解决方案，通过构造cXe≥cX0。

如果ak的列是线性无关的，那么根据44.2，我们知道Xe是一个基本可行的解，我们就这样做了。

否则，ak的列是线性相关的，所以有一些非零向量v=（v1，…，vs），这样ak v=0。设w∈rn为所有j/∈k通过设置wj=0扩展v得到的向量，通过构造，

aw=ak v=0.

我们将通过展示一个可行的解决方案x（t0）得出一个矛盾，这样cx（t0）≥cx0的零坐标比x多。

e

为此，我们声称我们可以假设w满足以下两个条件：

1. cw≥0.
2. 有一些j∈k使得wj<0。

如果cw=0，如果条件（2）失败，自w=06以来，对于一些j∈k，我们有wj>0，在这种情况下，我们可以使用−w，其中wj<0。

如果c w<0，则c（−w）>0，因此我们可以假设cw>0。如果wj>0表示所有j∈k，因为x是可行的，x≥0，所以x（t）=x+tw≥0表示所有t≥0。此外，由于aw=0和e e e x是可行的，我们有e

ax（t）=ax+taw=b，e

因此x（t）对所有t≥0都是可行的。我们也有

cx（t）=cx+tcw。

e

由于cw>0，当t>0变为无穷大时，目标函数cx（t）也趋于无穷大，这与上面有界的事实相矛盾。因此，必须存在满足上述（1）和（2）条件的一些w。

结果表明，存在一些t0>0，使得cx（t0）≥cx0和x（t0）=xe+t0w是可行的，但x（t0）的零坐标比xe多，这是一个矛盾。

因为x（t）=x+tw，我们有

e x（t）i=xei+twi，

因此，如果我们让i=i∈1，…，n wi<0 k，这是非空的，因为w满足上述条件（2），如果我们选择

，

然后t0>0，因为wi<0代表所有i∈i，根据k的定义，我们得到xei>0代表所有i∈k。根据t0>0的定义，由于xe≥0，我们得到

x（t0）j=xej+t0wj≥0，对于所有j∈k，

因此x（t0）≥0，x（t0）i=0，对于某些i∈i，因为ax（t0）=b（对于任何t），x（t0）是一个可行的解，cx（t0）=cxe+t0cw≥cx0+t0cw≥cx0，

而x（t0）i=0，对于一些i∈i，我们看到x（t0）的零坐标比xe多，这是一个矛盾。

44.3号提案意味着以下重要结果。

定理44.4。设（p2）为目标函数为cx的任意标准线性规划，其中ax=b，a是m×n阶矩阵，如果（p2）有一些可行解，如果（p2）有上界，那么一些基本可行解Xe是（p2）的最优解。证据。根据命题44.3，对于任何可行解x，都有一些基本可行解x

使cxcx。但是只有有限多的基本可行解，所以其中之一

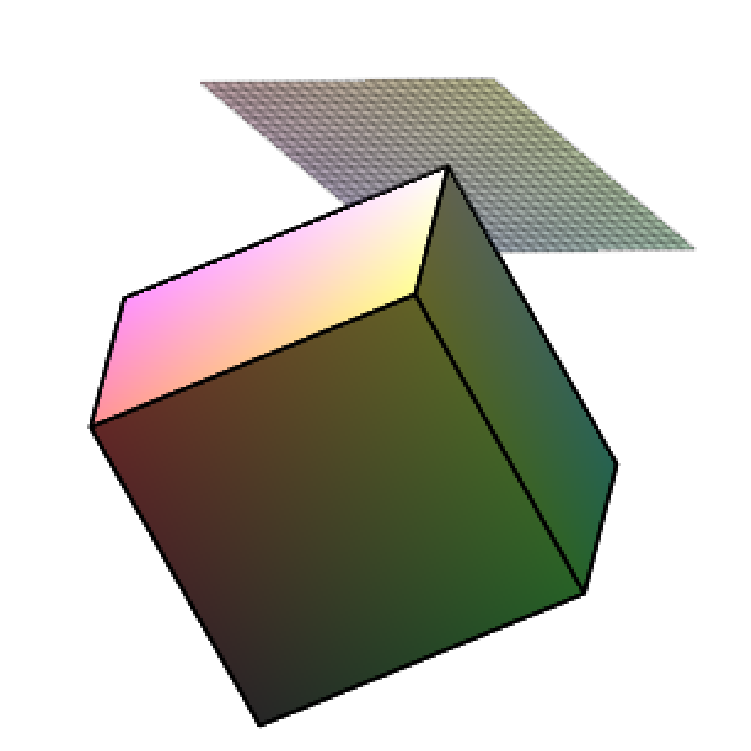
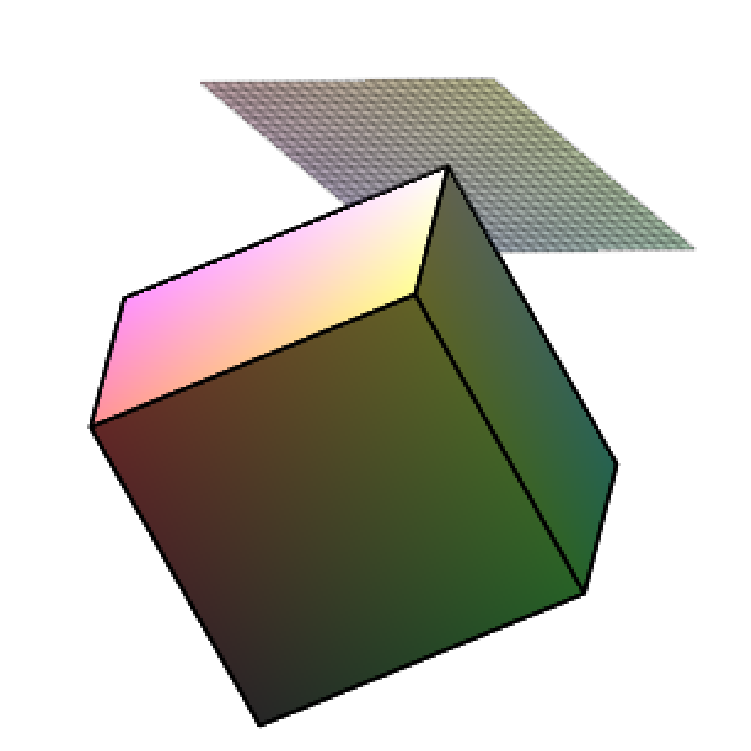
e必须得到目标函数的最大值。

几何学上，基本解就是多面体p（a，b）的顶点，我们现在定义了这个概念。

定义44.6.给定一个H多面体p rn，p的一个顶点是一个点v∈p，其性质是存在一些非零线性形式c∈（rn）和一些r，因此v是p的唯一点，其中，map x 7→cx具有最大值μ；即，对于所有y∈p−v，cy<cv=μ。几何上，这意味着方程cy=μ的超平面恰好在v处与p接触。更一般地说，p的凸子集f是p的k维面，如果f有维数k，并且如果存在某种仿射形式，那么，对于所有y∈f，cy=μ，对于所有y∈p−f，cy<μ。一维面称为边。

顶点的概念如图44.4所示，而边缘的概念如图44.5所示。

由于k维p的面f等于方程cx=祄的超平面h（）与p的交点，因此它确实是凸的，并且尺寸的概念是有意义的。



x + y + z = 3

(1

,1,

1)

图44.4：以原点为中心、对角线穿过（−1、−1、−1）和（1,1,1）的立方体有八个顶点。顶点（1,1,1）与线性形式x+y+z=3相关。

观察P的0维面是一个顶点。如果p有维d，则p的（d-1）维面称为其面。

如果（p）是标准形式的线性规划，那么它的基本可行解就是多面体p（a，b）的顶点。为了证明这一事实，我们需要以下简单的命题

提案44.5。设a x=b是一个线性系统，其中a是秩m的m×n矩阵。对于任意子集k 1，…，n的大小m，如果ak是可逆的，那么对于所有j/∈k（当然，x≥0），至多有一个基本可行解x∈rn，其中xj=0。

证据。为了使x可行，我们必须有ax=b。写n=1，…

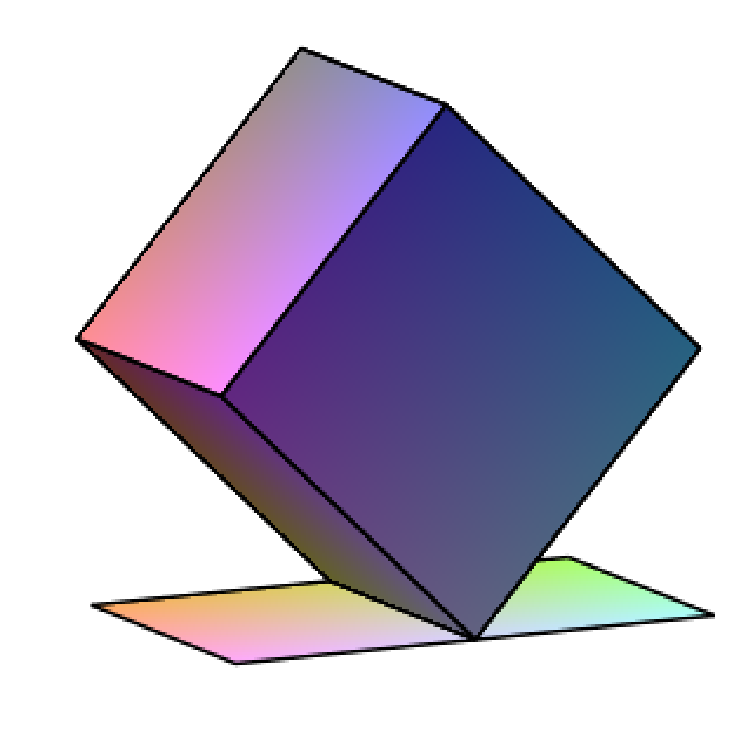
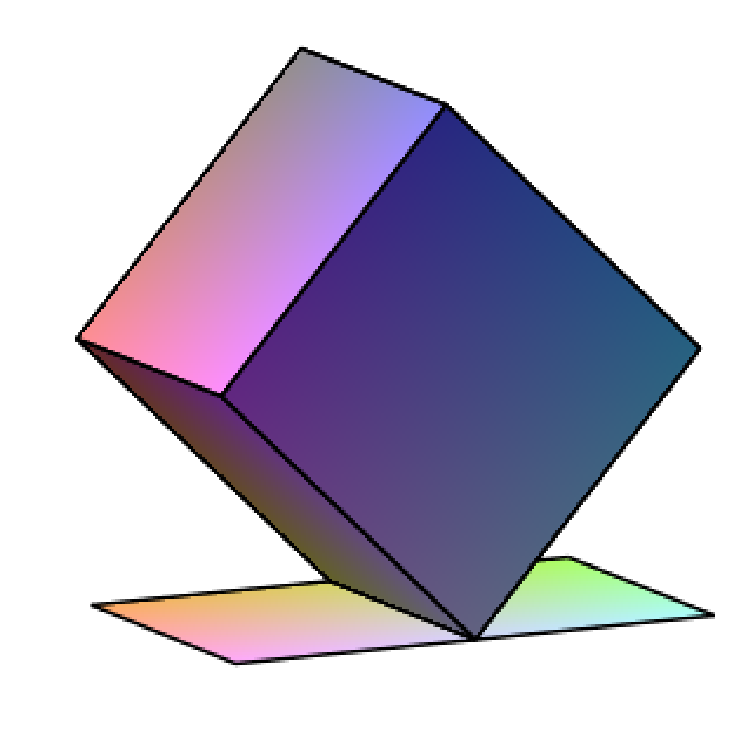
ax=akxk+anxn=b。

为了使x是一个基本可行的解，我们必须有xn=0，所以

AKXK=B。

由于假设ak是可逆的，所以xk=a−k1b是唯一确定的。如果xk≥0，那么x是一个基本可行解，否则它不是。证明了对于所有的j/∈k，至多有一个基本可行解x∈rn，xj=0。

定理44.6。设（p）为标准形式的线性规划，其中ax=b，a为秩m的m×n矩阵，对于每个v∈p（a，b），下列条件是等价的：



x

+

y

=

2

(1

,1,

1)

(1

,

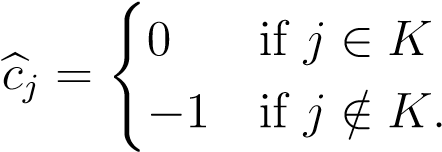
1,-1)

图44.5：立方体中心位于原点，对角线穿过（−1、−1、−1）和（1,1,1）有十二条边。从（1,1、−1）到（1,1,1）的边与线性形式x+y=2相关。

1. v是多面体p（a，b）的顶点。
2. v是线性规划（p）的基本可行解。

证据。首先，假设v是p（a，b）的顶点，并让\_（x）=cx−µ是一个线性形式，这样所有y∈p（a，b）和cv=µ的cy<µ。这意味着v是p（a，b）的唯一点，目标函数x 7→cx在p（a，b）上具有最大值μ，因此根据定理44.4，由于这个最大值是由一些基本可行解获得的，所以唯一性v必须是基本可行解。

相反地，假设v是（p）的一个基本可行解，对应于大小m的子集k 1，…，n。让bc∈（rn）是由



通过构造Cv=0且Cx≤0，因此p（a，b）b b b上的函数x→7cx在v处有最大值。此外，对于任何x≥0，Bcx<0，使得x j>0对于某些j/∈k。然而，根据44.5，向量v是唯一基本可行的解，使得vj=0对于所有j/yk，因此v是p（a，b）最大化函数x 7→cx的唯一点，

所以它是一个顶点。

在理论上，为了找到一个最优解，我们尝试了所有可能的m元素子集k，1，…，n，并求出相应的akx=b的唯一解xk，然后检查该解是否满足xk≥0，计算cxk，并返回一个目标函数最大的可行xk。这是一个完全不切实际的算法。

一种实用的算法是单纯形算法。基本上，simplex算法试图在polyhdron p（a，b）中沿着边缘从顶点到顶点“爬升”（使用基本可行解），试图最大化目标函数。我们将在下一章中介绍单纯形算法。读者也可以查阅有关线性规划的文本。特别是，我们推荐马托塞克和加德纳[120]、奇瓦塔尔[40]、帕帕迪米特里奥和斯蒂格里茨[130]、伯特西马斯和齐齐克利斯[21]、西亚莱[41]、施里伊弗[144]和范德贝[175]。

注意，定理44.4断言，如果标准形式的线性规划（p）（其中ax=b和a是m×n阶矩阵）有一些可行解，并且在其上有界，那么一些基本可行解就是最优解。根据定理44.6，多面体p（a，b）必须有一些顶点。

但是假设我们只知道p（a，b）是非空的，也就是说，我们不知道目标函数cx在上面是有界的。p（a，b）有顶点吗？

上述问题的答案是肯定的，这一点很重要，因为单纯形算法需要一个初始的基本可行解才能开始。这里我们证明，如果p（a，b）是非空的，那么它必须包含一个顶点。这个证明仍然没有构造性地生成一个顶点，但是我们将在下一章看到，如果有一个顶点，simplex算法总是找到一个顶点（前提是我们使用一个防止循环的轴规则）。

定理44.7。设（p）为标准形式的线性规划，其中ax=b，a是秩m的m×n矩阵，如果p（a，b）是非空的（有可行解），则p（a，b）有一些顶点；等价地，（p）有一些基本可行解。

证据。证明依赖于一个技巧，即添加松弛变量xn+1，…，xn+m并使用新的目标函数−（xn+1+····+xn+m）。

如果我们让ab是m×（m+n）-矩阵，x、x和xb是由

，

然后考虑标准形式的线性程序（P）

最大化−（xn+1+····+xn+m），以abxb=b和xb≥0为准。

由于所有I的Xi超0，目标函数（xn+1＋···+xn+m）有界以上0。系统abxb=b相当于系统

ax+x=b，

因此，对于每一个可行解u∈p（a，b），由于au=b，向量（u，0m）也是（pb）的可行解，实际上是一个最优解，因为x=0的目标函数−（xn+1+·····+xn+m）的值为0。根据命题44.3，线性规划（pb）有一些基本可行解（u，w），其中目标函数的值大于或等于（u，0 m）的目标函数的值，并且由于（u，0 m）是最优解，（u，w）也是（pb）的最优解。）这意味着w=0，因为否则目标函数−（xn+1+·····+xn+m）将具有严格的负值。

因此，（）是（pb）的基本可行解，因此u的非零分量对应的列是线性无关的。u的一些坐标可以等于0，但由于a具有秩m，我们可以添加a的列以获得与u相关的基k，并且u确实是（p）的基本可行解。

基本可行解的定义可适用于约束形式为a x≤b，x≥0的线性程序；见Matousek和Gardner[120]（第4章，

第4节，定义4.4.2）。

除了形式aix≤bi的约束之外，最常见的线性程序类型还允许形式aix≥bi或aix=bi的约束。变量XI也可以取负值。始终可以将此类程序转换为定义44.1中考虑的类型。我们按以下步骤进行。

每个约束aix≥bi替换为约束−aix≤−bi。每个等式约束aix=bi都被两个约束aix≤bi和−aix≤−bi替换。

如果有n个变量Xi，我们创建n个新变量Yi和n新变量Zi，用Yi—Zi替换每个变量Xi。我们还增加了2n约束yi≥0和zi≥0。如果约束由不等式ax≤b给出，我们现在有约束由

.

我们将目标函数cx替换为cy-cz。

注：我们还表明，我们可以用等式约束ax=b替换不等式约束ax≤b，通过添加约束为非负的松弛变量。

第四十五章

# 单体法

## 45.1单纯形算法背后的思想

由于Dantzig提出的单纯形算法适用于标准形式的线性规划（P），其中约束条件由a x=b和x≥0给出，矩阵m×n为秩m，目标函数c 7→cx。该算法要么报告（p）没有可行解，要么（p）没有边界，要么生成最优解。在几何上，该算法在多面体P（A，B）中逐点爬升，试图提高目标函数的值。由于顶点对应于基本可行解，因此单纯形算法实际上适用于基本可行解。

回想一下，一个基本可行解X是一个可行解，其中有一个子集k 1，…，n的大小m，这样，由指数属于k的a的列组成的矩阵ak是线性无关的，并且x j=0表示所有j/∈k。我们还让

j>（x）是一组索引

j>（x）=j∈1，…，n xj>0，

因此，对于一个与k有关的基本可行解x，我们有j>（x）k。实际上，根据命题44.2，可行解x是一个基本可行解，如果aj>（x）的列是线性无关的。

如果j>（x）对所有基本可行解x都有基数m，那么simplex算法在每一步都会取得进展，这意味着它将严格增加目标函数的值。不幸的是，对于某些基本可行解，（x）<m是可能的，在这种情况下，单纯形算法的一个步骤可能不会增加目标函数的值。更糟糕的是，在极少数情况下，算法可能进入无限循环。这种称为循环的现象可以被检测出来，但在这种情况下，算法无法给出一个确切的答案。

幸运的是，有一些方法可以阻止simplex算法循环（例如，bland的规则在后面讨论），尽管要证明这些规则正确工作是很重要的。

一千四百三十七

基本可行解的潜在“坏”行为记录在以下定义中。

定义45.1.给定一个标准形式的线性规划（P），其中约束条件由a x=b且x≥0给出，m×n矩阵为m阶，当（x）<m时，一个基本可行解x退化，否则它是非退化的。

如果原点0n是一个基本可行解，则它是退化的。对于一个不那么简单的例子，x=（0,0,0,2）是以下线性程序的退化基本可行解，其中m=2，n=4。

例45.1。

最大化x2取决于

−x1+x2+x3=0 x1+x4=2 x1≥0，x2≥0，x3≥0，x4≥0。

矩阵A和向量B由下式给出：

，

如果x=（0,0,0,2），则j>（x）=4。有两种方法可以形成包含第四列的一组两个线性无关的列。

给出了一个基本可行的解X，它与M的一个子集K有关，因为矩阵AK的列是线性无关的，通过滥用语言，我们把AK的列称为X的基。

如果u是（p）的一个顶点，即（p）的一个基本可行解，在“正常模式”下，单纯形算法试图沿着一条边从顶点u移动到相邻顶点v（带有u，v∈p（a，b）rn），它对应于一个基本可行解，其b将一个基本向量a k替换为k∈k，将另一个非基本向量a j替换为一些j/∈k，从而提高目标函数的值。

让我们用一个例子来演示这个过程。

例45.2。设（p）为下列标准形式的线性程序。

最大化x1+x2

从属于

−x1+x2+x3=1 x1+x4=3 x2+x5=2 x1≥0，x2≥0，x3≥0，x4≥0，x5≥0。

矩阵A和向量B由下式给出：

.

K

1

0

1

2

3

4

5

K

1

1

2

3

-

x

+

x

=

1

1

2

u

u

u

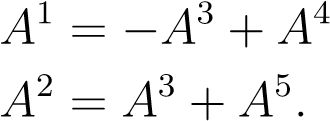
0

1

2

图45.1：与实施例45.2相关的平面H多面体。最初的基本可行解是起源。单纯形算法首先沿水平橙色线移动到顶点U1处的可行解。然后沿垂直红线移动，得到最优可行解u2。

与基k=3,4,5对应的向量u0=（0,0,1,3,2）是一个基本可行解，目标函数的对应值为0+0=0。由于K=3,4,5对应的列（a3、a4、a5）是线性独立的，我们可以将a1和a2表示为



自从

1a3+3a4+2a5=au0=b，

对于任何θ∈r，我们有

B=1A3+3A4+2A5−θA1+θA1

=1A3+3A4+2A5−θ（−A3+A4）+θA1

=θa1+（1+θ）a3+（3−θ）a4+2a5，

和

b=1a3+3a4+2a5−θa2+θa2=1a3+3a4+2a5−θ（a3+a5）+θa2=θa2+（1−θ）a3+3a4+（2−θ）a5。

在第一种情况下，向量（θ，0,1+θ，3−θ，2）是一个可行的解，iff 0≤θ≤3，目标函数的新值是θ。

在第二种情况下，向量（0，θ，1−θ，3,2−θ，1）是一个可行的解iff 0≤θ≤1，目标函数的新值也是θ。

考虑第一个案例。很自然地，我们会问是否可以通过设置（θ，0,1+θ，3−θ，2）的坐标中的一个为零来获得另一个顶点并增加目标函数，在这种情况下，通过选择θ=3来获得第四个顶点。这就产生了与基（a1，a3，a5）相对应的可行解（3，0，4，0，2），因此确实是一个基本可行解，目标函数的改进值等于3。请注意，A4离开了基准（A3、A4、A5），A1进入了新的基准（A1、A3、A5）。

我们现在可以用基（a1，a3，a5）来表示a2和a4，这很容易，因为我们已经用（a3，a4，a5）来表示a1和a2了，a1和a4被交换了。这种步骤称为旋转步骤。我们得到

A2=A3+A5

A4=A1+A3。

然后我们用u1=（3,0,4,0,2）和基（a1，a3，a5）重复这个过程。我们有

B=3A1+4A3+2A5−θA2+θA2

=3A1+4A3+2A5−θ（A3+A5）+θA2

=3A1+θA2+（4−θ）A3+（2−θ）A5，

B=3A1+4A3+2A5−θA4+θA4=3A1+4A3+2A5−θ（A1+A3）+θA4=（3−θ）A1+（4−θ）A3+θA4+2A5。

在第一种情况下，点（3，θ，4−θ，0,2−θ）是一个可行的解iff 0≤θ≤2，目标函数的新值为3+θ。在第二种情况下，点（3−θ，0,4−θ，θ，2）是一个可行的解iff 0≤θ≤3，目标函数的新值为3−θ。为了增加目标函数，我们必须选择第一种情况并选择θ=2。然后得到可行解u2=（3，2，2，0，0），对应于基（a1，a2，a3），是一个基本可行解。目标函数的新值为5。

接下来，我们用基础（a1，a2，a3）表示a4和a5。同样，这很容易做到，因为我们刚刚交换了A5和A2（一个旋转步骤），我们得到

.

我们用u2=（3,2,2,0,0）和基（a1，a2，a3）重复这个过程。我们有

B=3A1+2A2+2A3−θA4+θA4

=3A1+2A2+2A3−θ（A1+A3）+θA4

=（3−θ）A1+2A2+（2−θ）A3+θA4，

和

B=3A1+2A2+2A3−θA5+θA5=3A1+2A2+2A3−θ（A2−A3）+θA5=3A1+（2−θ）A2+（2+θ）A3+θA5。

在第一种情况下，点（3−θ，2,2−θ，θ，0）是一个可行的解iff 0≤θ≤2，目标函数的值是5−θ。在第二种情况下，点（3,2−θ，2+θ，0，θ）是一个可行的解iff 0≤θ≤2，目标函数的值也是5−θ。因为我们必须有θ≥0才能有一个可行的解，所以没有办法增加目标函数。在这种情况下，我们得到了一个最优解，在我们的情况下，u2=（3,2,2,0,0），目标函数的最大值等于5。

我们还可以将单纯形算法应用于顶点u0=（0,0,1,3,2）和向量（0，θ，1θ，1），这是一个可行的解决方案，iff 0≤θ≤1，具有新的目标函数值。选取θ=1，得到了与基（a2，a4，a5）相对应的可行解（0，1，0，3，1），这实际上是一个顶点。目标函数的新值为1。然后我们用获得的基础（a2，a4，a5）表示a1和a3。

A1=A4−A3

a3=a2−a5，

用（0,1,0,3,1）和基（a2，a4，a5）重复该过程。再经过三个步骤，我们就能得到最佳解u2=（3,2,2,0,0）。

让我们回到例子45.1的线性程序，目标函数x2，其中矩阵a和向量b由

.

回想一下，u0=（0,0,0,2）是一个退化的基本可行解，目标函数的值为0。与实施例45.1相关的H多面体的平面图见图45.2。

K

1

0

1

2

3

K

1

1

2

3

u

2

1

x

x

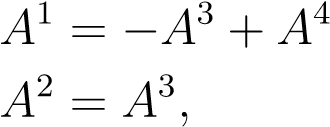
2

u

1

图45.2：与实施例45.1相关的平面H多面体。最初的基本可行解是起源。单纯形算法沿着倾斜的橙色线移动到三角形的顶点。

选择基础（A3、A4）。然后我们有了



我们得到

B=2A4−θA1+θA1

=2A4−θ（−A3+A4）+θA1

=θa1+θa3+（2−θ）a4，

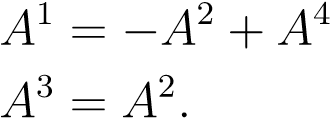
b=2a4−θa2+θa2=2a4−θa3+θa2

=θa2−θa3+2a4。

在第一种情况下，点（θ，0，θ，2−θ）是可行解iff 0≤θ≤2，目标函数的值为0，在第二种情况下，点（0，θ，−θ，2）是可行解iffθ=0，目标函数的值为θ。但是，由于在第二种情况下必须有θ=0，所以也没有办法增加目标函数。

结果表明，为了使单纯形算法所考虑的情况尽可能互相排斥，因为在第二种情况下，目标函数值的θ系数为非零，即1，我们应该选择第二种情况。我们必须选择θ=0，但是我们可以交换向量a3和a2（因为a2进来了，a3有系数−θ，这就是θ必须为零的原因），我们得到了基本可行解u1=（0,0,0,2）和新的基（a2，a4）。请注意，这个基本可行解与之前的相同顶点（0,0,0,2）对应，但基础已经改变。向量

a1和a3可以用基（a2，a4）表示为



我们现在用u1=（0,0,0,2）和基（a2，a4）重复这个过程，我们得到

B=2A4−θA1+θA1

=2a4−θ（−a2+a4）+θa1=θa1+θa2+（2−θ）a4，

和

B=2A4−θA3+θA3=2A4−θA2+θA3

=−θa2+θa3+2a4。

在第一种情况下，点（θ，θ，0,2−θ）是可行解iff 0≤θ≤2，目标函数的值是θ，在第二种情况下，点（0，−θ，θ，2）是可行解iffθ=0，目标函数的值是θ。为了增加目标函数，我们必须选择第一种情况并选择θ=2。得到了可行解u2=（2，2，0，0），其对应基为（a1，a2），目标函数值为2。

矢量a3和a4以基（a1，a2）表示为

a3=a2

a4=a1+a3，

我们用u2=（2，2，0，0）和基（a1，a2）重复这个过程。我们得到

B=2A1+2A2−θA3+θA3=2A1+2A2−θA2+θA3

=2A1+（2−θ）A2+θA3，

和

B=2A1+2A2−θA4+θA4=2A1+2A2−θ（A1+A3）+θA4=（2−θ）A1+2A2−θA3+θA4。

在第一种情况下，点（2,2−θ，0，θ）是可行解iff 0≤θ≤2，目标函数的值是2−θ，在第二种情况下，点（2−θ，2，−θ，θ）是可行解iffθ=0，目标函数的值是2。这一次没有办法改进目标函数，我们得到了目标函数最大值为2的最优解u2=（2,2,0,0）。

现在让我们考虑一个无限线性程序的例子。

例45.3。设（p）为下列标准形式的线性程序。

最大化x1取决于

x1−x2+x3=1

−x1+x2+x4=2 x1≥0，x2≥0，x3≥0，x4≥0。

矩阵A和向量B由下式给出：

.

与基k=3,4对应的向量u0=（0,0,1,2）是一个基本可行解，目标函数的对应值为0。向量a1和a2用基（a3，a4）表示

A1=A3-A4

A2=−A3+A4。

从u0=（0,0,1,2）开始，我们得到

b=a3+2a4−θa1+θa1

=a3+2a4−θ（a3−a4）+θa1

=θa1+（1−θ）a3+（2+θ）a4，

b=a3+2a4−θa2+θa2=a3+2a4−θ（−a3+a4）+θa2=θa2+（1+θ）a3+（2−θ）a4。

K

1

0

1

2

3

4

5

K

1

1

2

3

4

5

-

x

+

x

=

2

1

2

x

-

x

=

1

1

2

θ

(1

,

1)

图45.3：与实施例45.3相关的平面H多面体。最初的基本可行解是起源。单纯形算法首先沿水平靛蓝线移动到顶点（1,0）处的基本可行解。通过沿橙色箭头θ（1,1）参数化的边界线移动，可以得到任何最优可行解。

在第一种情况下，点（θ，0,1−θ，2+θ）是可行解iff 0≤θ≤1，目标函数的值是θ，在第二种情况下，点（0，θ，1+θ，2−θ）是可行解iff 0≤θ≤2，目标函数的值是0。为了增加目标函数，我们必须选择第一种情况，并选择θ=1。我们得到了与基（a1，a4）对应的可行解u1=（1，0，0，3），因此它是一个基本可行解，目标函数的值为1。

向量a2和a3是根据基（a1，a4）通过下式给出的。

.

用u1=（1,0,0,3）重复这个过程，我们得到

b=a1+3a4−θa2+θa2

=a1+3a4−θ（−a1）+θa2=（1+θ）a1+θa2+3a4，

和

b=a1+3a4−θa3+θa3=a1+3a4−θ（a1+a4）+θa3=（1−θ）a1+θa3+（3−θ）a4。

在第一种情况下，点（1+θ，θ，0,3）是所有θ≥0的可行解，如果为1+θ，则目标函数的值；在第二种情况下，点（1-θ，0，θ，3-θ）是可行解，如果为0≤θ≤1，则目标函数的值为1-θ。这一次，我们的处境是

（1+θ，θ，0,3）=（1,0,0,3）+θ（1,1,0,0），θ≥0

在可行解集合中形成无限射线，目标函数1+θ在此射线上不受上述限制。这表明我们的线性规划虽然可行，但是无界的。

现在让我们描述一下单纯形算法的一般步骤。

## 45.2一般的单纯形算法

我们假设我们已经有了一个初始顶点U0。这个顶点对应于一个基为k0的基本可行解。稍后我们将证明，总是可以找到线性规划（P）的基本可行解是标准形式，或者检测（P）没有可行解。

单纯形算法的思想是：给定一对（u，k）由基本可行解u和u的基k组成，找到另一对（u+，k+）由另一个基本可行解u+和u+的基k+组成，这样通过删除一些基本索引k−∈k从k中得到k+。并加入一些非基本指标j+∈/k，使目标函数的值增加（最好严格）。交换向量ak和aj+的步骤称为旋转步骤。

假设u是一个给定的顶点，对应于一个基本可行解的基k，由于与指数k∈k对应的m向量ak是线性无关的，它们构成了一个基，因此对于每一个非基j/∈k，我们写下

aj=xγkjak.（）

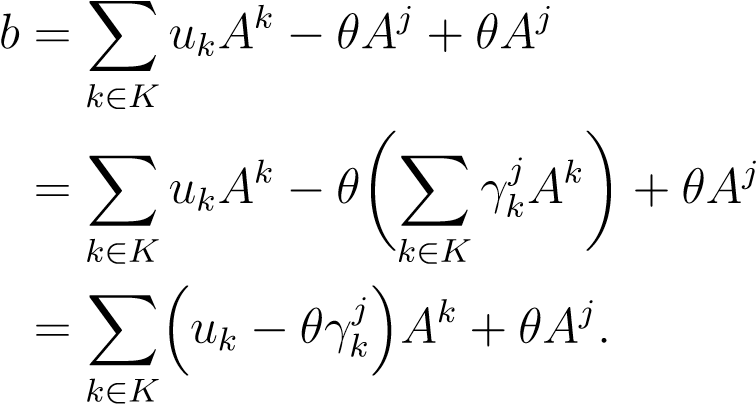
钾钾

我们让它成为向量。实际上，由于矢量依赖于k，为了非常精确，我们应该用（γkj）k表示它的分量，但是为了简化符号，我们通常写γkj而不是（γkj）k（除非出现混淆）。稍后我们将解释如何有效地计算系数γkj。

因为u是一个可行的解，我们有u≥0和au=b，也就是说，

（）

对于每一个输入基k的候选非基j/∈k，我们试图找到一个新的顶点u（θ），它改进了目标函数，为此我们在方程（）中加−θaj+θaj=0到b，然后用方程（）的右边替换−θaj中aj的出现。泰恩



因此，出现在上述方程右侧的向量u（θ）由

J

u i−θγi，如果i∈k u（θ）i=θ，如果i=j

0，如果i/∈k j

自动满足约束条件au（θ）=b，该向量是一个可行的解iff

θ≥0且Uk≥θγkj，所有k∈k。

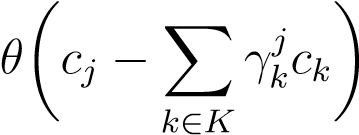
显然θ=0是一个解，如果

，

然后我们得到了0≤θ≤θj的一系列可行解，u（θ）的目标函数的值为

.

因为目标函数的潜在变化是



θ≥0，如果Cj−pk∈kγkjck≤0，则目标函数不能增加。

然而，如果某些j+∈/k为0，如果θj+>0，则可以通过选择任何θ>0严格增加目标+函数，使θ≤θj，因此选择θ=θj+，自然会使u（θ）的至少一个系数为零，这也会使目标函数的增加最大化。在这种情况下（以下情况（b2）），我们得到一个新的可行解u+=u（θj+）。

现在，如果θj+>0，那么有一些指数k∈k，这样的−0，所以我们可以为离开基K的向量ak选择这样的指数k−。我们声称k+=（k−−k−）j+是基。这是因为与ak列相关的系数不是零（实际上是0），所以方程（），即

aj+=γkj+−ak−+xγkj+ak，

K∈K−K−

得出方程

，

这些方程表明，向量（m，sincek）k∈k+所跨越的子空间是相同的。然而，K+也有M元素，它必须是一个基础。因此，k是维数m的基础，因此该子空间具有维数auk+）k=∈kuand，向量（θj+）是基本的

（A）

可行的解决方案。

上述情况是最常见的，但也可能出现其他情况。接下来，我们将讨论所有可能发生的情况。

案例（a）。

对于所有的j/∈k，我们都有0，然后证明u是一个最优解。否则，我们是在案例（b）中。

案例（b）。

对于一些j/∈k（不一定唯一），我们有Cj−pk∈kγkjck>0。有三个子类。

案例（B1）。

如果对于上面的一些j/∈k，我们也有0代表所有k∈k，因为Uk≥0代表所有k∈k，这对θ没有限制，并且目标函数在上面是无界的。示例45.3证明了这一点，其中k=3,4和j=2，因为

案例（B2）。

存在一些指数j+∈/k，因此同时

0，这意味着目标函数可能是

增加；

（2）存在一些k∈k+，使得γkj+>0，对于每一个暗示θj>0。

如果我们选择θ=θj+其中

，

0，然后Uk>0，那么可行的解决方案u+由

+=uθj+−θj+γij+ifif i i∈=kj+i

用户界面

0，如果i/∈k j+

是p（a，b）的顶点。如果我们选取任何指数k−k，使之为+，那么

K+=（K−K−）J+是U+的基础。矢量aj进入新的基k+，矢量ak-离开旧的基k。这是一个关键步骤。目标函数严格增加。示例45.2证明了这一点，其中k=3,4,5，j=1，k=4，然后

0），其中=1。因为=3，新的最佳

溶液变为U+=（3,0,1−3（−1），3−3（1），2−3（0））=（3,0,4,0,2）。

案例（B3）。

有一些指数j/∈k，使得cj−pk∈kγkjck>0，对于满足上述性质的每一个指数j/∈k，我们同时拥有

1. Cj−pk∈kγkjck>0，表示目标函数可能增加；
2. 有一些k∈k，这样γkj>0，uk=0，这意味着θj=0。

因此，目标函数不变。在这种情况下，u是一个退化的基本可行解。

我们可以将u+=u关联为一个新的基k+，如下所示：选取任何指数j+∈/k，这样

Cj+−xγkj+ck>0，

钾钾

以及任何指数k−∈k，使得

，

设k+=（k−k−−）j+。与案例（b2）一样，矢量aj+进入新的基k+，矢量ak离开旧的基k。这是一个关键步骤。然而，由于θj+=0，目标函数没有变化。示例45.1证明了这一点，其中k=3、4、j=2和k=3。

很容易证明在（a）种情况下，基本可行解u是最优解，而在（b1）种情况下，线性规划是无界的。我们已经证明了在（b2）情况下，向量u+及其基k+构成一个基本可行解，并且在（b3）情况下的证明是相似的。有关详细信息，请参阅CIARLET[41]（第10章）。

通过引入以下集合，可以方便地重新解释所考虑的各种情况：

，

和

.

那么很容易看出以下等价物成立：

箱（A）⇒B=∅，箱（B）⇒B=6∅

箱（b1）⇒b1=6∅

壳体（b2）⇒b2=6∅

箱（B3）⇒B3=6∅。

此外，案例（a）和（b）、案例（b1）和（b3）、案例（b2）和（b3）是互斥的，而案例（b1）和（b2）不是互斥的。

如果情况（b1）和情况（b2）同时出现，我们选择情况（b1），即线性程序（p）是无界的，并终止算法。

下面是一些关于这个方法的评论。

在（b2）情况下，即算法最常遵循的路径，必须对θj+>0（k+中的新索引）的索引j+∈/k做出各种选择。同样地，对于指数k−k离开k，必须做出各种选择，但这些选择通常不那么重要。

同样地，在（b3）情况下，必须为进入k+的新指数j+∈/k做出各种选择。在（b2）和（b3）情况下，做出此类选择的标准称为透视规则。

只有当u是退化顶点时，情况（b3）才会出现。但即使U退化，当γkj>0时，当Uk>0时，也可能出现情况（b2）。也可能发生u是非退化的，但由于（b2）的结果，新顶点u+退化，因为至少有两个分量，并且对于某些不同的k1，k2∈k消失。

事例（a）和（b1）对应于算法终止的情况，并且事例（b2）在单纯形算法执行期间只能出现有限次数，因为目标函数严格地从顶点增加到顶点，并且只有有限多个顶点。因此，如果在任何初始基本可行解U0上启动单纯形算法，则可能出现三种互斥情况之一：

1. 有一个有限的事例（b2）和/或事例（b3）出现的序列，以事例（a）的出现结束。最后一个顶点是最优解。这是在示例45.1和45.2中发生的。
2. 有一个有限的事例（b2）和/或事例（b3）出现的序列，以事例（b1）的出现结束。我们认为这个问题是无限的，因此没有解决的办法。这是在示例45.3中发生的。
3. 有一个有限的事例（b2）和/或事例（b3）出现的序列，后跟一个无限的事例序列（b3）。如果发生这种情况，算法将访问某个基两次。这就是所谓的循环现象。在这一点上，算法最终没有得出结论。

尽管在实践中很少出现这样的情况，但也有一些例子说明自行车运动的发生。CHVATAL[40]中给出了这样一个例子；见第3章，第31-32页，例如在某个轴规则下，七个变量和三个方程在六次迭代后循环的例子。

第三种可能性可以通过选择合适的支点规则来避免。其中两条规则是布兰德规则和词典编纂规则；见chvatal[40]（第3章，34-38页）。

布兰德的规则是：选择符合条件的入局指数j+∈/k中最小的一个，同样选择符合条件的出局指数k−k中最小的一个。

可以证明，如果将布兰德法则选为支点法则，则不会发生循环。证明是非常技术性的；见Chvatal[40]（第3章，第37-38页），Matousek和Gardner[120]（第5章，定理5.8.1），以及Papadimitriou和Steiglitz[130]（第2.7节）。因此，假设给出了一些初始的基本可行解，并使用合适的支点规则（如布兰德规则），单纯形算法总是终止，要么生成最优解，要么报告线性程序是无边界的。不幸的是，布兰德的规则是最慢的中枢规则之一。

枢轴规则的选择极大地影响了单纯形算法所经历的枢轴步骤的数量。我们不打算在这里解释各种中枢规则。我们简单地提到了以下规则，将读者引向马托塞克和加德纳[120]（第5章，第5.7节）或第43.1节中引用的文本。

1. 最大系数，或但丁法则。
2. 最大增幅。
3. 最陡的边缘。
4. 布兰德的规则。
5. 随机边缘。

最陡的边缘规则是最流行的规则之一。我们的想法是最大化比率

.

随机边规则在所有符合条件的索引中随机选取输入基向量的索引j+∈/k。

现在让我们回到单纯形算法的初始化问题。我们使用定理44.7证明过程中引入的线性程序（pb）。考虑线性程序（p2）

最大化cx，以ax=b且x≥0为准，

以标准形式表示，其中a是m×n阶矩阵。

首先，观察约束条件是方程，我们可以确保b≥0，因为每个方程aix=bi，其中bi<0可以被−aix=−bi替换。下一步是以标准形式引入线性程序（pb）

最大化受试者

其中ab和xb由

.

因为我们假设b≥0，向量xb=（0n，b）是（pb）的可行解，实际上是一个基本可行解，因为与指数n+1相关的矩阵，…，n+m是单位矩阵im。此外，由于所有I的Xi超过0，目标函数-（xn+1＋···+xn+m）被限制在0以上。

如果我们用一个防止循环的支点规则来执行单纯形算法，从基本可行解（0n，d）开始，因为目标函数是以0为界的，那么单纯形算法以由一些基本可行解（如（u，w）给出的最优解终止，其中urn和w rm。

在定理44.7的证明中，对于每一个可行解u∈p（a，b），向量（u，0m）是（pb）的最优解。因此，如果w=06，那么p（a，b）=∅，因为对于每个可行解u∈p（a，b），向量（u，0m）将产生目标函数−（xn+1+····+xn+m）等于0的值，但（u，w）将产生严格的负值，因为w=0.6

45.3。如何高效地执行旋转步骤

否则，=0，u是（p2）的可行解。由于（）是（p）的基本可行解，因此与u的非零分量对应的列是线性无关的。u的一些坐标可以等于0，但由于a具有秩m，我们可以添加a的列以获得与u相关的基k，u实际上是（p2）的基本可行解。

在线性程序pb上运行单纯形算法，得到线性程序p2的初始可行解（u0，k0），称为单纯形算法的第一阶段。在线性规划（p2）上运行具有初始可行解（u0，k0）的单纯形算法称为单纯形算法的第二阶段。如果线性规划（p2）的可行解很容易得到，则跳过第一阶段。有时，在第一阶段结束时，已经得到（p2）的最优解。

总之，我们证明了以下事实值得记录。提案45.1.对于任何线性程序（p2）

最大化cx，以ax=b且x≥0为准，

在标准格式中，如果a是m×n矩阵，秩m和b≥0，则考虑标准格式的线性程序（pb）。

最大化−（xn+1+····+xn+m），以abxb=b和xb≥0为准。

具有防止循环的透视规则的单纯形算法从（pb）的基本可行解xb=（0n，b）开始，以最优解（u，w）结束。

1. 如果w=06，则p（a，b）=∅，即线性规划（p2）没有可行解。
2. 如果w=0，那么p（a，b）=6，u是（p2）的一个基本可行解，与一些基k有关。

命题45.1表明，由方程组a x=b和不等式x≥0定义的多面体p（a，b）是否是可判定的。这个决策过程使用了simplex算法的一个防故障版本（防止循环），并且证明它总是终止并返回一个答案是非常重要的。

## 45.3如何有效地执行旋转步骤

我们现在简要讨论如何从基本可行解（u，k）中计算（u+，k+）。

为了避免应用置换矩阵，最好允许基k是一个索引序列，可能是无序的。因此，对于任意m×n矩阵a（m≤n）和任意序列k=（k1，k2，·····，km），矩阵ak表示m×m矩阵，其中i列为a的k ith列，同样，对于任意向量u∈rn（resp）。任意线性形式c（rn）），向量uk rm（线性形式ck（rm））是向量，其第i个条目是u（resp）中的kith条目。第i项是c）中kith项的线性项。

对于每一个非基j/∈k，我们有

，

所以向量由，也就是说，通过解系统

akγkj=aj.（γ）

要非常精确，因为矢量γkj依赖于k，它的分量应该用（γkj）ki表示，但正如我们前面所说，为了简化符号，我们写的不是（γkj）ki。

为了确定哪种情况适用（（a）、（b1）、（b2）、（b3）），我们需要计算所有j/∈k的cj−pk∈kγkjck。为此，观察

.

如果我们写βk=cka−k1，那么

，

我们发现，βk>是系统βk>=（a−k1）>c>k的解，这意味着βk>是系统的解。

.（β）

注：由于u是（p）的一个基本可行解，所以所有j/∈k都有uj=0，所以u是方程akuk=b的解，因此u的目标函数的值为cu=ckuk=cka−k1b，这一事实将在46.2节中起到至关重要的作用，说明n单纯形算法以线性规划（P）的最优解终止，然后它也产生了对偶线性规划（D）的最优解。

假设我们有一个基本可行解u，u的基k，我们也有矩阵ak及其逆a−k1（可能是隐式的），以及a>k的逆（a>k）−1（可能是隐式的）。下面是单纯形算法迭代步骤的描述，几乎完全遵循chvatal（chvatal[40]，第7章，框注7.1）。

（修正）单纯形法的迭代步骤

45.3。如何高效地执行旋转步骤

第1步。计算所有j/∈k的cj−pk∈kγkjck=cj−βkaj，为此，计算βk>作为系统的解。

a>kβk>=c>k。

如果所有j/∈k的cj−βkaj≤0，停止并返回最优解u（情况（a））。

第2步。如果出现（b）种情况，则使用一个轴规则来确定哪个指数j+∈/k应输入新的基k+（条件cj+−βkaj+>0应保持）。

第3步。为此计算最大值，求解线性系统

.

第4步。如果最大值为0，则停止并报告线性程序（P）未绑定（案例（B1））。

J+

第5步。如果maxj+，并使用一个支点规则来确定哪个索引k∈kγk>0，则使用allk−k∈kkksu的比率，这样

计算θ应保留k（情况（b2））。

如果max=0，则使用一个轴规则来确定哪个索引k−应该保留基k（案例（B3））。

第6步。将U、K和AK更新为U+和K+以及AK+。在这一步骤中，给定k=（k1，…，k`，…，km）序列指定的基k，其中k−=k`，则k+是通过将k'替换为输入索引j+，得到的序列，因此k+=（k1，…，j+，…，km）替换为“th slot”中的j+。

向量u很容易更新。为了从AK计算AK+，我们利用AK和AK+只有一列不同，即“第j列aj+”，由线性组合给出。

.

为了简化符号，表示并回忆一下k−=k`。如果k=（k1，…，km），那么ak=[ak1··········+aim]，因为ak+是替换

AK在AJ列，我们有

AK+=[AK1····AJ+···AIM]=[AK1···AKγ···AIM]=AKE（γ），

式中，e（γ）是从单位矩阵im中通过用γ替换其第n列得到的以下可逆矩阵：

1γ1\_

…

 

1γ`−1

E（γ）=γ`。

 

γ`+1 1 1\_

 

……γm 1

由于0，矩阵e（γ）是可逆的，很容易检查其逆矩阵是否由

1−γ`−1γ1……e（γ）−1=1−γ`−1

-

1℃，

−γ`-1γ`+1 1

……γ

-γ`-1γm 1

计算起来很便宜。我们也有

.

因此，如果AK和A−K1可用，那么AK+和可以根据AK和A−K1以及形式E（γ）的矩阵便宜地计算。然后，系统（γ）找到向量γkj可以很便宜地解决。

自从



和

（a>k+）−1=（a>k）−1（e（γ）>）−1，

矩阵和（也可以从a>k，（a>k）−1和形式e（γ）>的矩阵中廉价计算。因此，寻找线性形式βk的系统（β）也可以廉价地求解。

形式e（γ）的矩阵称为eta矩阵；见chvatal[40]（第7章）。证明了单纯形算法S步后得到的矩阵AKS可以写成

AKS=AKS−1ES

对于一些ETA矩阵，因此AKS可以写成产品

AKS=e1e2···es

s eta矩阵。这种分解称为eta分解。ETA因子分解可用于反转AKS或迭代求解AKSγ=AJ+形式的系统。哪个方法更有效取决于EI的稀疏性。

总之，从（u，k）中找到下一个基本可行解（u+，k+）的方法很便宜。我们只是想让读者了解一下这些技巧。为了有效地实现单纯形方法，我们向读者介绍了有关线性规划的文本。特别是，修改后的单纯形法在chvatal[40]、Papadimitriou和Steiglitz[130]、Bertsimas和Tsitiklis[21]和vanderbei中介绍。

〔175〕。

## 45.4使用tableaux的单纯形算法

我们现在描述一种表示单纯形算法的形式主义，即（完全）tableaux。这是所有书籍中使用的传统形式主义，模小变化。Tableau形式主义的一个特别好的特点是，可以使用基本行操作来更新Tableau，这与将矩阵缩减为行的梯队形式（rref）过程中使用的操作相同。不同的是选择轴的标准。

由于Cj−ckγkj的数量在确定Aj列应作为基础的过程中起着关键作用，因此使用符号Cj表示，这称为变量xj的降低成本。降低的成本实际上取决于k，所以要非常精确，我们应该用（ck）j表示它们，但为了简化表示法，我们编写cj而不是（ck）j。我们将很快看到（ck+）i是如何用（ck）i计算的。

观察执行单纯形算法下一步所需的数据是

1. 当前的基本解决方案UK及其基础K=（k1，…，km）。
2. 降低的成本cj=cj−cka−k1aj=cj−ckγk j，对于所有j/∈k。
3. 所有j/∈k的向量，允许我们表达每一个。

所有这些信息都可以打包成（m+1）×（n+1）矩阵，称为（完整）表，其组织如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 |
| 网络错误 |  |  |  |  |  |

······

可以方便地将第一行视为第0行，第一列视为第0列。第0行包含目标函数的当前值和降低的成本。第0列（除顶部条目外）包含当前基本解决方案UK的组件，其余列（除顶部条目外）包含向量γkj。观察到与k中指数j对应的γkj构成单位矩阵im的置换。该条目称为Pivot元素。一个表格连同新的基础k+=（k-k-）j+包含计算新的UK+、新的γkj+和新的降低成本（ck+）j所需的所有数据。

如果我们将m×n矩阵\_定义为矩阵\_=[γk1······γkn]，其jth列为γkj，c为行向量c=（c1····c n），则上表用简略的形式表示

|  |  |
| --- | --- |
| 网络错误 | 网络错误 |
| 网络错误 | 网络错误 |

我们现在证明，表格的更新可以使用基本行操作来执行，这与将矩阵缩减成行的梯队形式（RREF）过程中使用的操作相同。

如果k=（k1，…，km），j+是输入基向量的索引，k−=k`是离开基的列的索引，如果k+=（k1，…，k`-1，j+，k`+1，…，km），sincej ak+=

J

，新列γk+根据旧列γk使用（γ）和方程式计算。

.

因此，矩阵\_+由\_给出。

.

但是矩阵是这样的

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 | 网络错误 | 网络错误 | 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 | 网络错误 | 网络错误 | 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 |
| 网络错误 | | | | | | |

1. 将row`乘以1（轴的倒数），使row`和column j+上的条目等于1。
2. 从第一行减去（标准化的）第行，对于i=1，………………………………………………………，m。

正是这些基本的行操作将\_的第列减少到了标识矩阵im的第列。因此，此步骤与将矩阵转换为行缩减梯队的过程在矩阵的第列执行的步骤序列相同。唯一的区别是选择轴的标准。

由于新的基本解决方案UK+由给出，我们已经

.

这意味着U+是通过应用与\_完全相同的基本行操作从英国获得的。因此，正如将矩阵简化为rref的过程一样，我们可以对矩阵[uk\_]应用基本行操作，该矩阵由表的第1行，…，m组成。

一旦得到新的矩阵\_+，新的降低成本由以下命题给出。

45.2号提案。给定任何标准形式的线性程序（p2）

最大化cx，以ax=b且x≥0为准，

其中a是m×n矩阵的秩m，如果（u，k）是（p2）的基本（不可行）解，如果k+=（k−k−）j+，其中k=（k1，…，km）和k−k`，那么对于i=1，…，n我们有

.

使用降低的成本表示法，上述方程为

.

证据。在不损失一般性和简化符号的情况下，假设k=（1，…，m），并为j+写j，为km写j。因为，和ak+=ake（γkj），

我们有

，

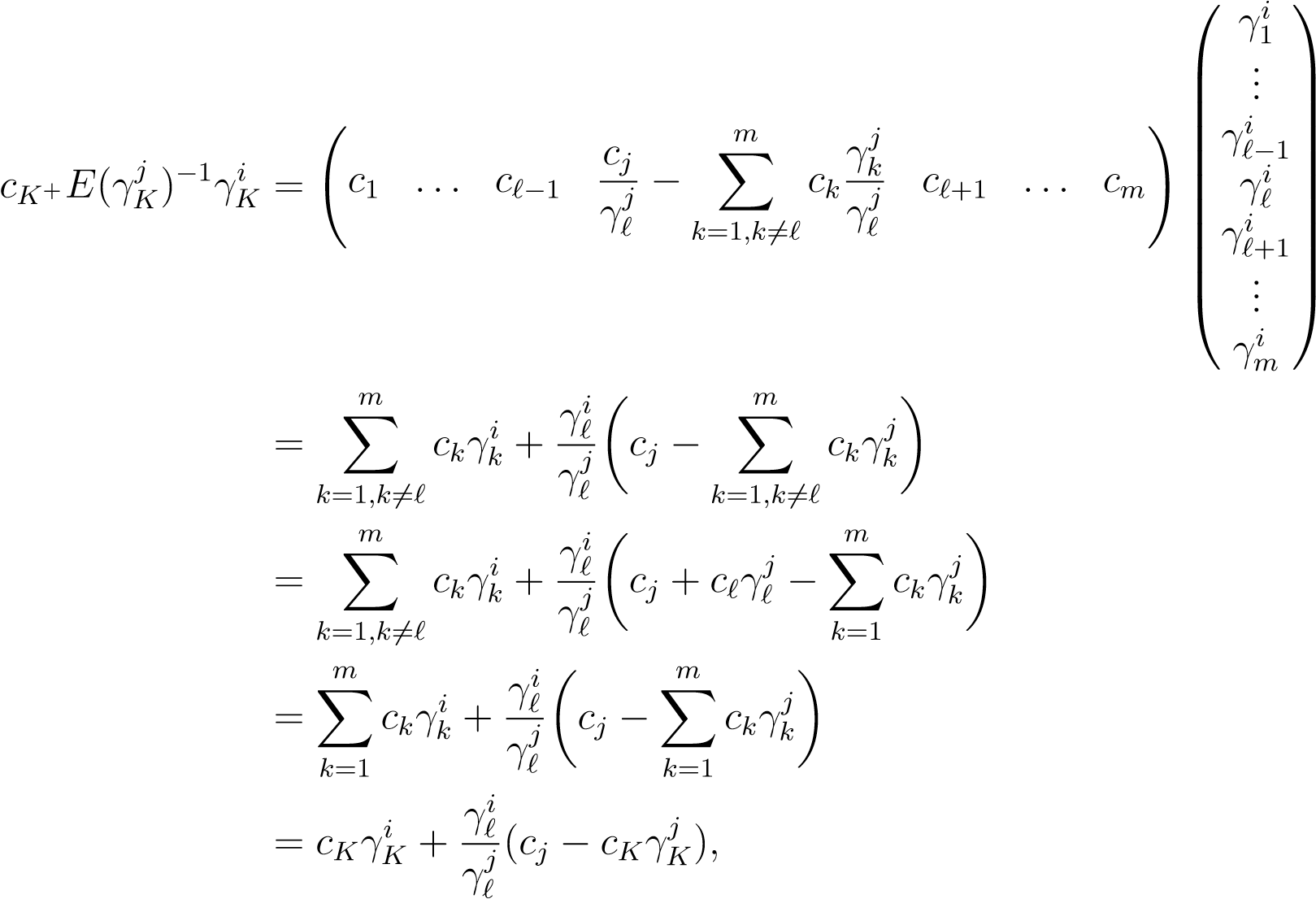
哪里

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 | 网络错误 | 网络错误 | 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 | 网络错误 | 网络错误 | 网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误  网络错误 |

其中第一栏包含γs。由于ck+=（c1，…，c`-1，cj，c`+1，…，cm），我们有

，

和



因此

，

如要求。

因为（）是\_的第几行，我们看到45.2号命题表明

，（？）

其中表示\_的第`-行，是轴。这意味着，ck+是通过基本行操作获得的，该操作包括首先将第行标准化，将其除以轴，然后从ck中减去（ck）j+×标准化行。正是这些行操作使成本降低（ck）j+0。

备注：很容易看出我们也有

ck+=c−ck+\_+。

我们在第45.2节中看到，在J+列进入和K-列离开的旋转步骤之后，目标函数的变化由下式得出：

，

哪里

如果我们用zk表示目标函数ckuk的值，那么我们可以看到

.

这意味着，目标函数的新值zk+是通过将第行标准化，除以轴得到的，然后将（ck）j+×标准化第行的第零项乘以（ck）j+，加到第0行的第零项得到的。

在更新降低的成本时，我们从第0行减去而不是增加（ck）j+×标准化行。这建议将−zk存储为第0行上的第0个条目，而不是zk，因为所有条目行0都由相同的基本行操作更新。因此，从现在开始，我们使用表格的表格

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 |
| 网络错误 | 网络错误 |  | 网络错误 |  |  |

······

simplex算法首先通过选择Cj>0的某个列来选择传入列j+，然后通过考虑0（沿着列j+）的比率来选择传出列k-，并选择k-来实现这些比率的最小值。

下面是对Papadimitriou和Steiglitz[130]的一个例子使用基本行操作的单纯形算法的说明（第2.9节）。例45.4。考虑线性程序

最大化−2x2−x4−5x7

x1+x2+x3+x4=4 x1+x5=2 x3+x6=3

3x2+x3+x7=6

x1、x2、x3、x4、x5、x6、x7≥0。

我们得到了基本可行解u=（0,0,0,4,2,3,6），其中k=（4,5,6,7）。由于ck=（−1,0,0、−5）和c=（0、−2,0、−1,0,0−5），第一个表是

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

由于cj=cj−ckγkj，第0行是通过从c=（0、−2,0、−1,0、0、−5）减去−1×第1行和−5×第4行得到的。让我们选择J+=1列作为传入列。我们有比率（对于列1上的正条目）

4/1，2/1，

因为最小值是2，所以我们把输出列选为k−=5列。枢轴1用红色表示。新的基础是k=（4,1,6,7）。接下来，我们应用行操作将列1减少到标识矩阵I4的第二个向量。为此，我们从第1行减去第2行。我们有看台

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

为了计算新的降低的成本，我们希望将c1设置为0，因此我们应用相同的行操作，并从第0行中减去第2行以获得表au。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

接下来，选择J+=3列作为传入列。我们有比率（对于第3列的正条目）

2/1，3/1，6/1，

因为最小值是2，所以我们把输出列选为k−=4列。支点1用红色表示，新基础为k=（3,1,6,7）。接下来，我们应用行操作将列3减少到标识矩阵I4的第一个向量。为此，我们从第3行和第4行中减去第1行，得到表格：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

-

为了计算新的降低的成本，我们要将c3设置为0，所以我们从第0行减去6×第1行，得到表格

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

-

接下来我们选择j+=2作为输入列。我们有比率（对于

第2列）

2/1，4/2，

因为最小值是2，所以我们把输出列选为k−=3列。支点1用红色表示，新基础为k=（2,1,6,7）。接下来，我们应用行操作将列2减少到标识矩阵I4的第一个向量。为此，我们将第1行添加到第3行，并从第4行减去2×第1行，得到表格：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

为了计算新的降低的成本，我们要将c2设置为0，所以我们从第0行减去8×第1行，得到表格

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

唯一可能的输入列对应于j+=5。我们有比率（对于第5列的正条目）

2/1，0/3，

因为最小值是0，所以我们把输出列选为k−=7列。支点3用红色表示，新基础为k=（2,1,6,5）。由于最小值为0，基k=（2,1,6,5）退化（事实上，对应于索引5的分量为0）。接下来，我们应用行操作将第5列减少到标识矩阵I4的第四个向量。为此，我们将第4行乘以1/3，然后将标准化的第4行添加到第1行，并将标准化的第4行从第2行中减去，得到表AU：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

为了计算新的降低的成本，我们希望将c5设置为0，所以我们从第0行减去13×第4行，得到表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 2  /  3 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

唯一可能的输入列对应于j+=3。我们有比率（对于第3列的正条目）

2/（1/3）=6，2/（2/3）=3，3/1=3，

因为最小值是3，所以我们把输出列选为k−=1列。支点2/3用红色表示，新基础为k=（2,3,6,5）。接下来，我们应用行操作将列3减少到标识矩阵I4的第二个向量。为此，我们将第2行乘以3/2，将第1行减去（1/3）×标准化第2行，将第3行减去标准化第2行，并将（2/3）×标准化第2行加到第4行，得到表格：

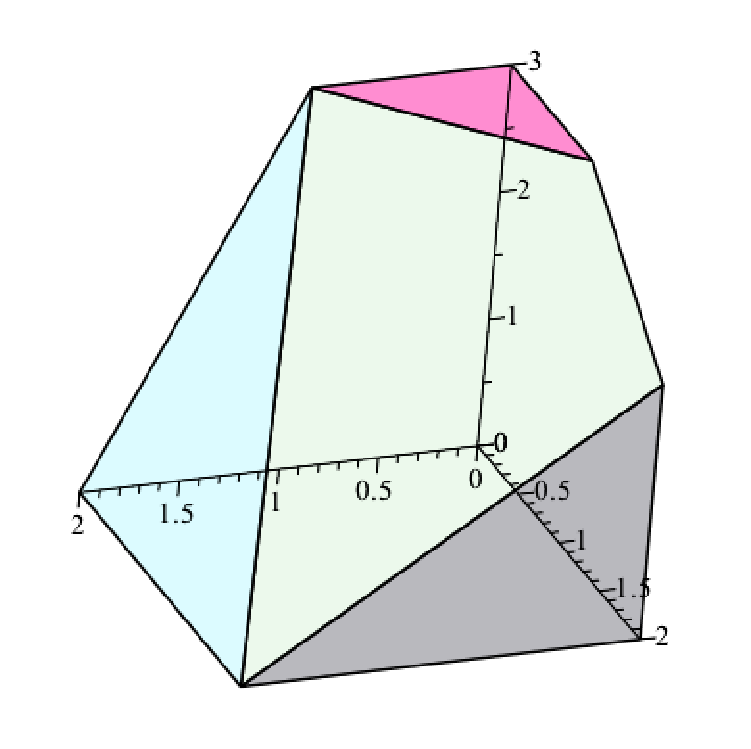
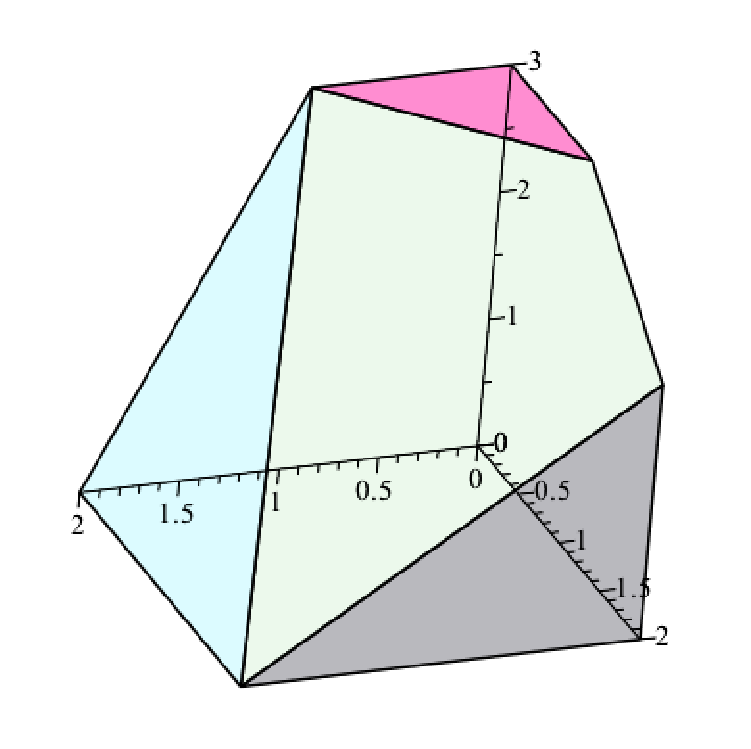
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

为了计算新的降低的成本，我们希望将c3设置为0，因此我们从第0行减去（2/3）×第2行以得到表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

由于所有降低的成本都小于等于0，我们得到了最优解，即（0,1,3,0,2,0,0,0,0），最优值为−2。

单纯形算法从一个基本可行解到另一个基本可行解的级数对应于多面体P顶点的访问，这些顶点与图45.4所示线性程序的约束有关。



x

3

x

2

x

1

0)

,2,

(2

2)

,2,

(0

0)

,2,

(0

,0,

0)

(2

3)

,0,

(1

3)

,0,

(0

,1,

(0

3)

1

x

=

2

1

3

4

=

5

6

x

+

x

+

x

=

4

1

2

3

3

x

+

x

=

6

3

2

图45.4：与通过Tableau方法优化的线性程序相关的多面体P。红色箭头路径跟踪单纯形方法从原点到顶点（0,1,3）的进程。

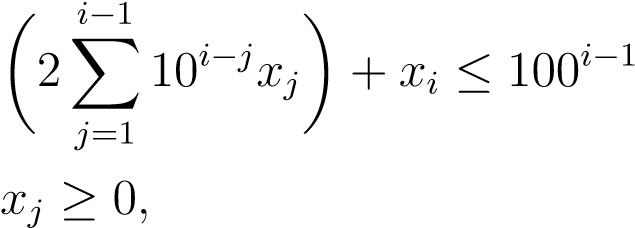
最后，如果需要运行simplex算法的第一阶段，在simplex算法以最优解（）和基k终止的情况下，一些ui=0，那么基k包含与需要的松弛变量对应的基本列aj的索引。被逐出基地。这很容易通过执行一个涉及其他列j+的旋转步骤来实现，这些列j+对应于一个原始变量（不是松弛变量），其中（γk）ji+=06。在这一步中，（γk）ji+<0或（ck）j+≤0都无关紧要。如果原始矩阵A没有多余的方程，那么这样的步骤总是可能的。否则，（γk）ji=0表示所有非松弛变量，因此我们检测到第i个方程是多余的，我们可以删除它。

Tableau方法的其他介绍见Bertsimas和Tsitiklis[21]和Papadimitriou和Steiglitz[130]。

## 45.5单纯形法的计算效率

最后，我们对单纯形算法的效率作一些评论。在实践中，Dantzig观察到，对于m<50和m+n<200的线性程序，单纯形算法通常需要小于3m/2的迭代，但最多需要3m的迭代。这一事实与最近用更大的程序进行的经验实验相一致，这些程序表明数字迭代是以3米为界的。因此，1972年，克莱和明蒂发现了一个线性程序，其中有n个变量和n个方程，其中dantzig的simplex算法的透视规则需要2n-1次迭代。本程序（摘自Chvatal[40]，第47页）转载如下：

最大化受试者



对于i=1，…，n和j=1，…，n。

如果p=max（m，n），那么，就更糟的情况而言，对于所有当前已知的透视规则，simplex算法在p中具有指数复杂性。然而，正如我们前面所说的，在实践中，诸如klee-minty示例之类的令人讨厌的示例似乎很少，并且迭代次数似乎是l。单位：m。

单纯形算法在多项式时间内以m为单位运行的轴规则（透视规则）是否仍然是一个悬而未决的问题。

Hirsch推测认为存在一些支点规则，使得simplex算法在o（p）步骤中找到最优解。到目前为止，由于Kalai和Kleitman，已知的最好的边界是m1+lnn=（2n）lnm。有关此主题的更多信息，请参见马托塞克和加德纳[120]（第5.9节）和贝尔茨马斯和齐齐克利斯[21]（第3.7节）。

研究人员研究了如果使用随机支点规则，在预期支点步数上找到上限的问题。已经找到了大于2米的界限（当然不是多项式）。

45.5。单纯形法的计算效率

理解线性编程的复杂性，特别是单纯形算法，仍然在进行中。感兴趣的读者可参考Matousek和Gardner[120]（第5章，第5.9节），了解一些要点。

在下一节中，我们将考虑确定一组约束a x≤b和x≥0是否具有解的重要理论标准。

1468第45章。单纯形算法

第四十六章

# 线性规划与对偶

## 46.1法卡斯引理的变体

如果a是m×n矩阵，如果b∈rm是向量，由线性代数可知，线性系统ax=b没有解，只要存在某种线性形式y∈（rm）使得ya=0和yb=06。这意味着y的线性消失在a1列，…，a的a列上，但不消失在b列上。由于线性形式y定义了方程y z=0的线性超平面h（用z∈rm），因此在几何上方程ax=b没有解，如果存在含有a1的线性超平面h，…，则方程ax=b没有解。不含b的d。这是一种分离定理，它表示向量a1，…，a和b可以被线性超平面h分离。

我们要做的是推广这种标准，首先是一个系统a x=b受约束x≥0，其次是一组不等式约束ax≤b和x≥0。确实有这样的标准以法卡斯·莱玛的名义存在。

关键是涉及多面体锥的分离结果，称为Farkas–Minkowski命题。我们有下面的基本分离引理。

提案46.1。设c rn为闭合的非空圆锥体。对于任意点a∈rn，如果a/∈c，则存在一个线性超平面h（到0），这样

1. c位于由h确定的两个半空间之一。
2. A/∈H
3. a位于由h决定的另一半空间。

我们说H严格地把C和A分开。

命题46.1是另一个分离定理的一个简单结果，该分离定理断言，给定任意两个非空闭凸集a和b，存在严格分离a和b的超平面h（这意味着a h=？，b h=？，a位于两个半空间中的一个确定由h决定，b位于由

一千四百六十九

h）；见Gallier[73]（第7章，推论7.4和命题7.3）。这个证明是不平凡的，涉及哈恩-巴纳赫定理的几何版本。

Farkas–Minkowski命题是46.1号命题，适用于多面体圆锥。

c=λ1a1+····+λnanλi≥0，i=1，…，n

其中a1，…，an是有限数量的向量ai∈rn。根据43.2，任何多面体锥都是闭的，所以46.1命题适用，我们得到了以下分离引理。

提案46.2.（farkas–minkowski）让c rn是非空多面体圆锥体c=圆锥体（a1，…，an）。对于任何点b∈rn，如果b/∈c，那么有一个线性超平面h（到0），这样

1. c位于由h确定的两个半空间之一。
2. B/∈H
3. B位于H所确定的另一半空间。

等价地，存在一个非零线性形式y∈（rn）这样

1. 当i=1，…，n时，yai≥0。
2. Yb<0.

本节末尾给出了不涉及46.1号提案的Farkas–Minkowski提案的直接证明。

注：应用于无限维实希尔伯特空间的Farkas–Minkowski命题有一个推广；见定理47.11（或Ciarlet[41]，第9章）。

命题46.2暗示了法卡斯引理的第一个版本。

提案46.3。（farkas-lemma，版本i）让a是m×n矩阵，让b∈rm

是任何向量。线性系统ax=b没有解x≥0，如果有一些非零线性形式y∈（rm）这样。

证据。首先，假设存在一个非零线性形式y∈（rm），使得yA≥0，yB<0。如果x≥0是ax=b的解，那么我们得到

Yax=Yb，

但如果yA≥0且x≥0，那么yAX≥0，但根据假设yB<0，这是一个矛盾。

接下来假设ax=b没有溶液x≥0。这意味着b不属于多面体圆锥体c=圆锥体（a1，…，an），由a.的列按命题跨越。

46.2，存在一个非零线性形式y∈（rm）这样：

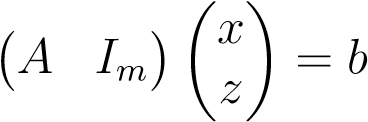
46.1。法卡斯引理的变体

1. j=1，…，n时，yaj≥0。
2. Yb<0，也就是说Yb<0。

然后考虑形式a x≤b和x≥0的不等式系统的可解性。

提案46.4.（Farkas引理，第二版）让a是m×n矩阵，让b∈rm是任意向量。不等式系统ax≤b没有解x≥0，如果存在一些非零线性形式y∈（rm），那么，和yb<0。

证据。我们使用线性规划的技巧，它包括添加“松弛变量”zi，将不等式aix≤bi转换为等式aix+zi=bi，其中zi≥0在定义43.5之前已经讨论过。如果我们让z=（z1，…，zm），很明显系统a x≤b有一个解x≥0，如果方程



具有x≥0且z≥0的溶液。现在由Farkas I，上述系统没有x≥0和z≥0的解，如果存在一些非零线性形式y∈（rm），那么



Yb<0，即，和Yb<0。

在下一节中，我们使用FarkasII来证明线性规划中的对偶定理。观察到，通过对Farkas II中等价物的否定，我们得到了一个可解性的标准，即：

不等式系统a x≤b对于每一个非零线性形式y∈（rm）有一个解x≥0 iff，那么yb≥0。

我们现在证明了Farkas–Minkowski命题，而不使用46.1号命题。这种方法使用了距离函数从一个点到一个闭集的基本属性。

设x rn为任意非空集，设a∈rn为任意点。从A到X的距离d（a，x）定义为

d（a，x）=inf kA−xk。

这里x∈x，k k表示欧几里得范数。

提案46.5。设x rn为任意非空集，设a∈rn为任意点。如果x是封闭的，那么有一些z∈x，这样ka−zk=d（a，x）。

BallProof.br（sincea）=x x是非空的，选择任意∈rn kx−ak≤r，然后明确地选择x0∈x，并让r=ka−x0k。如果br（a）是关闭的

d（a，x）=xinfx ka−xk=x∈xinf br（a）ka−xk。γ

自x 7→k br（a）是紧凑的，x是封闭的，k=x br（a）也是紧凑的。但是在紧集k上定义的函数a−xk是连续的，而由连续函数定义的紧集的图像是紧凑的，因此由Heine–Borel定义的它有一个最小值，由一些z∈k x实现。

注：如果u是希尔伯特空间v的非空闭凸子集，则希尔伯特空间理论（投影定理）的标准结果断言，对于任何v∈v，都存在一个

唯一的p∈u，使得kv−pk=uinf∈u kv−uk=d（v，u），

和

HP−V，U−Pi≥0，适用于所有U∈U。

这里kwk=phw，wi，其中h−，−i是希尔伯特空间v的内积。

我们现在可以证明法卡—明可夫斯基命题（46.2号命题）。

Farkas–Minkowski命题的证明。设c=cone（a1，…，am）为多面体

假设线性超平面是一个圆锥点（非空），并假设that最接近b，因为b/b/∈∈c c and。根据命题43.2，多面体锥是z z∈cc，因此我们得到ud=（b，cz−）=b=06 kb−，我们要求zk；也就是说，z闭合，根据命题46.5，有一些

h与u正交，如图46.1所示。

首先让我们展示一下

hu，zi=hz−b，zi=0.（1）

如果z=0，这是微不足道的，所以假设z 6=0。如果hu，zi 6z=00∈，那么eitherc更接近bhtanu，zi z>是，一个矛盾。0或hu，zi<0。在这两种情况下，我们都可以找到一些点，案例1:hu，zi>0。

对于任何α，设z0=（1−α）z，使0<α<1。然后z0∈c，既然u=z−b，

z0−b=（1−α）z−（z−u）=u−αz，

所以

kz0−bk2=ku−αzk2=kuk2−2αhu，zi+α2 kzk2。

如果我们选取k 2α>20，使得α<2hu，zi/kzk2z，那么是−2αhu，zci+最接近α2 kzk2 b<。0，所以kz0−bk2<uk=kz−bk，与以下事实相矛盾：

病例2：胡，子<0.

46.1。法卡斯引理的变体

a

1

a

2

3

a

b

z

H

C

图46.1：垂直于z-b的超平面h将点b与c=圆锥体（a1，a2，a3）分开。

z0−letb=（1+z0=（1+α）zα−z（zfor any−u）=a使用+αz soα≥−1。那么z0∈c，既然u=z-b，我们有

kz0−bk2=ku+αzk2=kuk2+2αhu，zi+α2 kzk2，

如果

0<α<−2hu，zi/kzk2，

那么2αhu，zi+α2 kzk2<0，所以kz0−bk2<kuk2=kz−bk2，一个矛盾如上所述。

因此，hu，zi=0。我们有hu，ui=hu，z−bi=hu，zi−hu，bi−hu，bi，

既然u 6=0，我们有hu，ui>0，所以hu，ui=−hu，bi意味着

hu，bi<0.（2）

仍然需要证明，hu，aii≥0，对于i=1，…，m。选取任意x∈c，使x 6=z。我们声称

hb−z，x−zi≤0.（3）

否则，hbon的直线段[−z，x−zi>0，也就是说，z，xh]z相对于−b，x−bzithan<0，我们表明我们可以找到一些z。点Z0∈C

对于0≤α≤1的任何α，我们得到z0=（1−α）z+αx=z+α（x−z）∈c，由于z0−b=z−b+α（x−z），我们得到kz0−b k2=kz−b+α（x−z）k2=kz−bk2+2αhz−b，x−zi+α2 kx−zk2，因此对于任何α>0，这样

α<−2Hz−b，x−zi/kx−zk2，

假设我们有2αhzz−是b，x−zi+αc2 K的一个点，关闭tox−zk2<b.0，这意味着kz0−bk2<kz−bk2，相反-

由于hb−z，x−zi≤0，u=z−b，并且（1），hu，zi=0，我们有

0±Hb- z，x＝Zi= h，u，x＝Zi＝Hu，X+Hu，Zi= Hu，Xi，

也就是说

所有X的C，Hi，0，（3）

如要求。特别地，

hu，aii≥0，i=1，…，m.（4）

Yathen，by（i≥i2）=1 and，…，m（4），the linear form defined by，which expense the farkas–minkowski proposition.y=u>satives the properties Yb<0 and

0用于

还有其他方法来证明Farkas–Minkowski命题，例如使用最小不可行系统或Fourier–Motzkin消除；见Matousek和Gardner[120]（第6章，第6.6和6.7节）。

## 46.2线性规划中的对偶定理

设（p）为线性规划

最大化cx，以ax≤b和x≥0为准，

假设目标函数为m×n矩阵，假设（xp7→）有一个可行解，且在其上有界，cx有界于onax≤b，对于任意p（a，bx），它可能是有用的∈p（a，b）。我们可以从不等式中推导出cx的上界，如下所示：对于每个不等式

Ax≤Bi 1≤I≤m，

选取一个非负的标量yi，将上述不等式的两边乘以yi得到

yiaix≤yibi 1≤i≤m，

（由于yi≥0，不等式的方向保持不变），然后将这些m方程相加，得出

（y1a1+······+ymam）x≤y1b1+·····+ymbm。

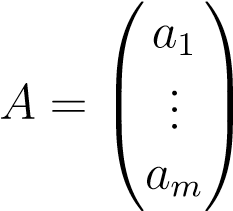
如果我们可以选择yi≥0这样

c≤y1a1+·····+ymm，

那么，既然xj≥0，我们有

cx≤（y1a1+·······+ymm）x≤y1b1+····+ymbm，

也就是说，对于任意可行解x∈p（a，b），我们找到（p）目标函数的值cx的上界。如果我们让y是线性形式y=（y1，…，ym），那么



y1a1+·····+ymm=ya，y1b1+·····+ymbm=yb，我们所做的是寻找一些y∈（rm）这样

c≤ya，y≥0，

所以我们有

cx≤yb，对于所有x∈p（a，b）。（）

然后自然地寻找yb的“最佳”值，即最小值，这导致了线性规划（p）对偶的定义，这是约翰·冯·诺依曼提出的概念。定义46.1.给定任意线性规划（P）

最大化cx，以ax≤b和x≥0为准，

对于m×n矩阵，（p）的对偶（d）是以下优化问题：

根据Ya≥c和Y≥0将Yb最小化，

其中y∈（rm），原线性规划（p）称为原线性规划。

下面是一个线性程序及其对偶的显式示例。

例46.1。考虑图46.2所示的线性程序。

最大化2x1+3x2

4x1+8x2≤12

2x1+x2≤3 3x1+2x2≤4 x1≥0，x2≥0。

其双线性程序如图46.3所示。

最小化12y1+3y2+4y3，以

4y1+2y2+3y3≥2 8y1+y2+2y3≥3 y1≥0，y2≥0，y3≥0。

可以检验：（x1，x2）=（1/2,5/4）是原线性规划的最优解，目标函数2x1+3x2的最大值等于19/4，（y1，y2，y3）=（5/16,0,1/4）是双线性规划的最优解，目标函数的最小值为第12y1+3y2+4y3节也等于19/4。

y

1

2

3

4

x

+

8

y

=

1

2

2

x

+

y

=

3

3

x

+

2

y

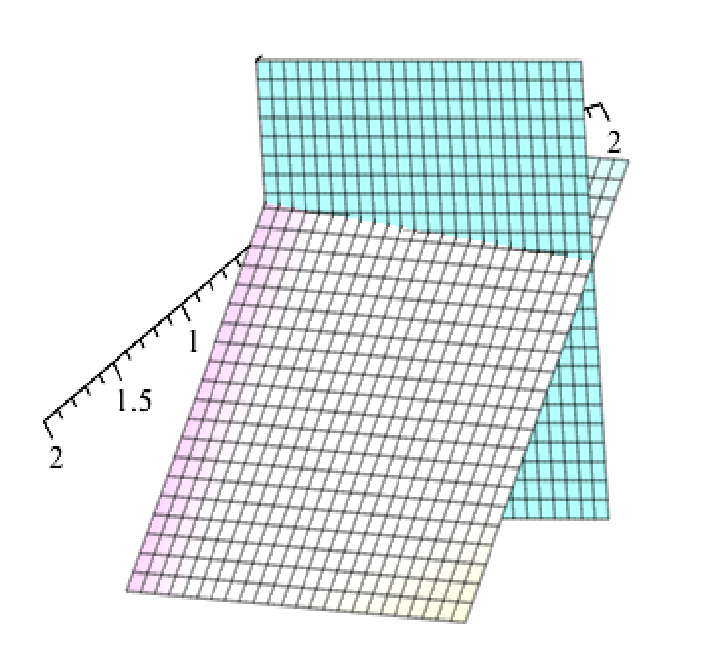
=

4

0 0 0.5 1 1.5 2 x

图46.2：实施例46.1线性程序的H-多面体。注x1→x和x2→y。

观察到在初等线性规划（P）中，我们寻找一个向量x∈rn，使形式cx最大化，并且约束由矩阵a的行在x上的作用决定。另一方面，在对偶线性规划（D）中，我们正在寻找



x

y

4

x + 2y + 3z =

2

8

x + y + 2z =

3

图46.3：实施例46.1的双线程序的H多面体是“粉色平面上方”和“蓝色平面前方”的空间区域。注y1→x、y2→y和y3→z。

对于线性形式y（r）m，最小化形式yb，约束由y在a列上的作用决定，这是（d）是对偶（p）的意义。在大多数演示中，（p）和（d）在相互对偶的空间中搜索解决方案的事实被过度使用换位所掩盖。

为了将双程序（d）转换为标准最大化问题，我们将目标函数y b更改为−b>y>并且将不等式y a≥c更改为−a>y>≤−c>。双线性程序（d）现在表示为（d0）

最大化−b>y>受−a>y>的影响≤−c>和y>大于等于0，

式中y∈（rm）。观察双程序（D0）的最大化形式（D00）的双程序返回原始程序（P）。

上述讨论建立了以下不等式，即弱对偶性。提案46.6.（弱对偶）给定任何线性规划（P）

最大化cx，以ax≤b和x≥0为准，

利用m×n矩阵，对于初等问题（p）的任何可行解x∈rn和对偶问题（d）的每一可行解y∈（rm），我们得到

Cx≤Yb。

我们说双线性规划（d）是有界的如果是有界的下面。

如果x是（p）的最优解，如果y是（d）的最优解，会发生什么？我们有cx≤y b，但是否存在“二元性差距”，也就是说，cx<y b可能存在？

答案是否定的，这是强对偶定理。事实上，强对偶定理比这更能说明问题。

定理46.7。（线性规划的强对偶性）让（p）成为任何线性规划

最大化cx，以ax≤b和x≥0为准，

具有m×n矩阵。初值问题（P）有一个可行解，在iff上有界，对偶问题（D）有一个可行解，在iff下有界。此外，如果（p）有一个可行解并且在其上有界，那么对于（p）的每一个最优解x和（d）的每一个最优解y，我们有

cx=y b.

证据。如果（p）有一个可行解，并且在其上有界，那么我们从44.1命题中知道（p）有一些最优解。设x为（p）的任何最优解。首先，我们将证明（d）有一个可行的解v。

=cx是目标函数x 7→cx的最大值。然后对于任何大于0的

不平等制度



没有溶液，因为否则，μ不会是目标函数cx的最大值。我们要应用Farkas II，所以首先我们将上述不平等系统转化为系统

.

根据命题46.4（Farkas II），存在一些线性形式（λ，z）∈（Rm+1）使得λ≥0，z≥0，

和

也就是说

，

也就是说，

.

另一方面，由于x≥0是系统ax≤b的最优解，由farkas ii再次（取等价的否定），由于λa≥0（与之前相同的λ），我们必须

λb≥0.（1）

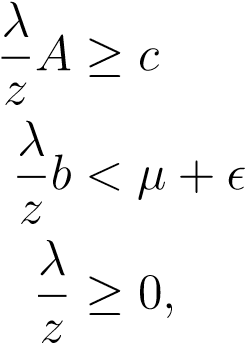
我们声称z>0。否则，因为z≥0，我们必须有z=0，但是



暗示

λb<0，（2）

由于λb≥0乘（1），我们有一个矛盾。因此，我们可以除以z>0而不改变不等式的方向，我们得到



结果表明，v=λ/z是对偶问题（d）的一个可行解。然而，弱对偶性（命题46.6）意味着，对于对偶程序（d）的任何可行解y≥0，cx==yb，因此（d）在下面是有界的，应用于（d）的一个最大化问题，我们得出（d）有一些最优解。对于（d）的任何最优解y，因为v是（d）的可行解，因此我们必须



由于我们的推理对任何>0都有效，因此我们得出结论：cx==y b。

如果我们假设对偶规划（d）有一个可行解，并且在下面有界，因为（d）的对偶是（p），我们得出（p）也是可行的，并且在上面有界。

强对偶定理也可以用单纯形方法证明，因为当它以（p）的最优解终止时，最后的表也产生（d）的最优解y，通过翻转它们的符号可以读出n+1，…，n+m列的降低成本。我们遵循CIARLET[41]中的证据（第10章）。定理46.8。考虑线性程序（P）

最大化cx，以ax≤b和x≥0为准，其等效版本（p2）为标准格式，

最大化C x BB

服从且x≥0，

乙

其中ab是m×（n+m）矩阵，bc是（rn+m）中的线性形式，xb∈rn+m，由

，

和（p）的对偶（d）由

根据Ya≥c和Y≥0将Yb最小化，

式中y∈（rm）。如果应用于线性程序（p2）的单纯形算法以最优解（u，k）终止，其中u是基本可行解，k是b的基

u，则是（d）的最优解，使得cu=y b。此外，y bbb是根据y=-（（ck）n+1…（ck）n+m的降低成本给出的。

证据。我们知道k是1，…，n+m的子集，由m个索引组成，因此ab的相应列是线性独立的。设N=1，…，N+M−K。单纯形法在（a）种情况下以最优解终止，即当

0表示所有j∈n，

其中abj=pk∈kγkjabk，或使用第45.3节的符号，

0表示所有j∈n。

上述不平等可以写成

，

或等同于

.（1）

目标函数的最佳解的值，由于满足方程b，目标函数的值为

（2）

那么，如果我们让，显然我们有，那么如果我们能证明y是对偶线性规划（d）的可行解，通过弱对偶，y是（d）的最优解。我们有

，（3）

到（1）我们得到

.（4）

设p为（n+m）×（n+m）置换矩阵，定义如下：

.

那么我们也有

利用方程（3）和（4），我们得出

，

也就是说，

相当于

，

那就是

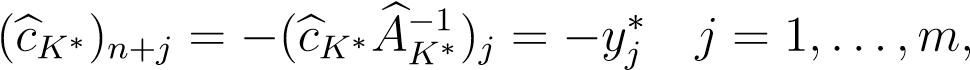
y a≥c，y≥0，

正是这些条件说明Y是双方案（D）的可行解。

降低的成本由（c，i=1，…，n+m.给出，但是

乙

i=n+1，…，n+m每列abn+j是单位矩阵im的jth向量，所以



如要求。

上面的证明相当简短这一事实是有欺骗性的，因为这个证明依赖于这样一个事实，即使用透视规则的simplex算法版本可以防止循环，但是这种透视规则正确工作的证明是相当长的。其他证据见Matousek和Gardner[120]（第6章第6.3节）、Chvatal[40]（第5章）和Papadimitriou和Steiglitz[130]（第2.7节）。

观察到，由于最终表格的最后m行实际上是通过乘以b b得到的，因此由最后m列和最后m行组成的m×m矩阵是a−k1（基本上，simplex算法执行了高斯-乔丹约简的步骤）。这个事实允许在原始对偶方法中保存一些步骤。

将弱对偶性和强对偶性结合起来，得到了如下定理，证明了正是四种情况出现。

定理46.9。（线性规划对偶定理）设（p）为任意线性规划

最大化cx，以ax≤b和x≥0为准，

让（d）成为它的双重计划

根据Ya≥c和Y≥0将Yb最小化，

具有m×n矩阵。然后就出现了以下可能性之一：

1. （p）和（d）都没有可行的解决方案。
2. （p）是无界的，（d）没有可行的解。
3. （p）没有可行解，（d）没有边界。
4. （p）和（d）都有一个可行的解决方案。然后两者都有一个最优解，对于（p）的每一个最优解x和（d）的每一个最优解y，我们有

cx=y b.

定理46.9的一个有趣的推论是，存在一个确定线性程序（P）是否具有最优解的测试。事实上，（p）有一个最优解，如果满足以下约束集：

ax≤b ya≥c cx≥yb

x≥0，y≥0>m。

实际上，对于上述系统的任何可行解（x，y），x是（p）的最优解，y是（d）的最优解。

## 46.3补充松弛条件

强对偶定理的另一个有用的推论是下面的结果称为平衡定理。

定理46.10。（平衡定理）对于任何线性规划（p）及其对偶线性规划（d）（具有一组不等式ax≤b，其中a是m×n矩阵，目标

46.3。补充松弛条件

函数x 7→cx），对于（p）的任意可行解x和（d）的任意可行解y，x和y是iff的最优解。

yi=0，其中（d）

和

xj=0，对于所有j。（p）

证据。首先，假设（d）和（p）保持不变。（d）中的方程表示yi=0，除非，因此

.

同样地，（p）中的方程表示xj=0，除非，因此

.

因此，我们得到cx=yb。

根据弱二元性（命题46.6），我们有

cx≤yb=cx

对于（p）的所有可行解x，所以x是（p）的最优解。同样地，

Yb=cx≤Yb

对于（d）的所有可行解y，所以y是（d）的最优解。

现在假设x是（p）的最优解，y是（d）的最优解。那么，正如在第46.6号提案的证明中，

.

由于强对偶性，因为x和y是最优解，所以上述不等式实际上是相等的，因此我们特别有

.

由于X和Y是可行的，Xi为0，YJ为0，因此，如果我们必须具有XJ＝0。

同样，我们也有

，

如果，那么yi=0。

（d）和（p）中的方程通常称为互补松弛条件。在对偶问题的帮助下，这些条件可以用来求解原问题的最优解，反之亦然。实际上，如果我们猜测一个问题的解决方案，那么我们可以使用互补松弛条件来求解对偶的解决方案，然后检查我们的猜测是否正确。这就是原始对偶方法的本质。为了呈现这种方法，首先我们需要更仔细地观察已经在标准形式中的线性程序的对偶。

## 46.4标准形式线性程序的对偶性

设（p）为标准形式的线性规划，其中，对于某些m×n秩矩阵（p），a x=b，we m和一些目标函数x 7→cx（当然，x≥0）。为了得到m）×n的对偶，将方程ax=b转换为包含（2矩阵）的以下不等式系统。

.

然后，如果用（y0，y00）表示2 m双变量，用y0，y00∈（rm），则上述程序的双变量为

最小化Y0B−Y00B

以和y0为准，y00≥0，

其中y0，y00∈（rm），相当于

根据（y0−y00）a≥c和y0，y00≥0，将（y0−y00）b最小化，

其中y0，y00∈（rm）。如果我们写y=y0−d）：y00，我们发现上面的线性程序等价于下面的线性程序（

根据Ya≥C将Yb最小化，

式中y∈（rm）。注意，y不必是非负的；它是任意的。

接下来，我们想知道对于已经是标准形式的线性程序，定理46.8的版本是什么。这很简单。

定理46.11。以标准形式考虑线性程序（p2）

最大化cx，以ax=b且x≥0为准，

46.4。标准形式线性规划的对偶性

它的对偶（d）由

根据Ya≥C将Yb最小化，

式中y∈（rm）。如果应用于线性程序（p2）的单纯形算法终止于最优解（u，k），其中u是基本可行解，k是u的基础，则y=ck a k1是（d）的最优解，因此cu=y b。此外，如果假设simplex算法从一个基本可行解（u0，k0）开始，其中

klast0=（n−m+1，…，n a的m列构成单位矩阵）（m最后一个微柱的指数），然后是最优解a）和a（n−m+1，…，ny）==cikm a（对于（d）的−k1是根据减少的成本

y=c（n−m+1，…，n）−（ck）（n−m+1，…，n）

由最后m列和最后m行组成的m×m矩阵为−k1。

证据。定理46.8的证明适用于a而不是ab，我们可以证明

CK A−K1 AN≥CN，

y=ck a−k1满足，cu=y b，和

y ak=ck a−k1 ak=ck，y an=ck a−k1 an≥cn。设p为n×n排列矩阵，定义如下：

.

那么我们也有

，

利用上述方程和不等式，我们得到

，

即y ap≥cp，相当于

y a≥c，

这表明y是（d）的可行解（记住，y是任意的，因此不需要约束y≥0）。

降低的成本由

（ck）i=ci−ck a−k1 ai，

因为对于j=n−m+1，…，n列aj是单位矩阵im的第（j+m−n）列，我们有

（ck）j=cj−（ck ak）j+m−n j=n−m+1，…，n，

也就是说，

y=c（n−m+1，…，n）−（ck）（n−m+1，…，n）

如要求。由于最终表格的最后m行是通过将[u0 a]乘以−k1得到的，A的最后m列构成im，因此最终表格的最后m行和最后m列构成−k1。

现在让我们来看一下定理46.10的互补松弛条件。如果我们回到（p）的版本

最大化cx

服从且x≥0，

以及（d）的版本

最小化Y0B−Y00B

以和y0为准，y00≥0，

式中，y0，y00∈（rm），由于不等式ax≤b和−ax≤−b一起意味着ax=b，我们对所有这些不等式约束都是相等的，因此条件（d）对y0和y00根本没有约束，而条件（p）则断言

xj=0表示所有j。

如果我们写y=y0−y00，上述条件等于

xj=0表示所有j。

因此，我们得到了定理46.10的以下版本。

定理46.12。（平衡定理，第2版）对于任何标准形式的线性规划（p2）（等式组a x≤b，其中a是m×n矩阵，目标函数x 7→cx）及其双线性规划（d），对于（p）的任何可行解x和（d）的任何可行解y，x和y都是最优解。使用IFF

xj=0，对于所有j。（p）

因此，应用于标准形式的线性规划（p2）及其对偶（d）的松弛条件仅对原始问题的变量xj施加松弛条件。

上述事实在原始对偶方法中起着至关重要的作用。

## 46.5双单纯形算法

以标准形式给出线性程序（p2）

最大化cx，以ax=b且x≥0为准，

其中a是m阶m的m×n矩阵，如果没有明显的可行解，但c≤0，那么我们可以使用一种称为双单纯形算法的方法，而不是使用第45.2节中描述的寻找可行解的方法。该方法使用基本解（u，k），其中au=b和uj=0表示所有uj∈/k，但不要求u≥0，因此u可能不可行。但是，Y=CKA−K1对于双方案是可行的。

根据Ya≥C将Yb最小化，

式中y∈（r）m.由于c≤0，观察到这是对偶的一个可行解。

如果发现（p2）的基本解u使u≥0，那么y=cka−k1时cu=yb，我们发现（p2）的最佳解u和（d）的最佳解y。对偶单纯形法是通过尝试使U零的负分量和减少对偶程序的目标函数来实现的。

双单纯形法从ax=b的基本解（u，k）开始，这是不可行的，但y=cka-k1是双可行的。在许多情况下，原线性规划是由一组不等式ax≤b和一些bi<0所规定的，因此通过加上松弛变量很容易找到这样的基本解u，如果加上c≤0，那么由于松弛变量的相关成本为0，我们认为y=0是可行的。双重解决方案。

给定ax=b（可行与否）的基本解（u，k），y=c k a−k1是双可行的iff cka−k1a≥c，由于cka−k1ak=ck，不等式cka−k1a≥c等于cka−kaan≥cn，即：

CN−CKA−K1an≤0，（1）

式中n=1，…，n−k.方程（1）等于

Cj−ckγkj≤0，所有j∈n，（2）

式中γkj=a−k1aj。回想一下，符号cj用于表示cj−ckγkj，这被称为变量xj的降低成本。

在simplex算法中，我们需要确定哪个列a k离开基k，哪个列aj进入新的基k+，以（d）的可行解，即0，其中n+=1，…，n−k+。我们使用

第45.2号提案决定第k列应离开基础。

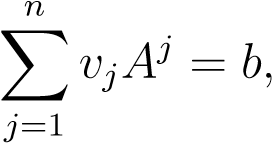
假设（u，k）是ax=b的解，其中y=cka-k1是双可行的。

案例（a）。如果u≥0，则u是（p2）的最优解。

案例（b）。有一些k∈k，使得uk<0。在这种情况下，选择一些k−k，使uk−<0（根据一些轴规则）。

案例（B1）。假设0代表所有j/∈k（实际上，代表所有j，因为代表所有j∈k）。如果是这样，我们认为（p2）是不可行的。

实际上，让v是一些基本可行的解。我们有v≥0和av=b，也就是说，



因此，通过将两边乘以−k1，并根据定义γkj=a−k1aj，我们得到

.

但回想一下，假设Uk−<0，而Vj≥0和γk j−≥0，所有j的指数k−的分量在左边为零或正，在右边为负，这是一个矛盾。因此，（p2）确实不可行。

案例（B2）。我们有些J有0。

我们选择AJ列，输入其中0的基础。由于我们假设所有j∈n的cj−ckγkj≤0（2），考虑

，

还有那套

.

我们选取一些指数j+∈n（礹+）作为进入基的列的指数（使用一些支点规则）。

回想一下，假设所有j/∈k的ci−ckγk i≤0，所有i∈k的ci−ckγki=0。

从0开始，对于任何索引0，我们都有0，从0开始

提案45.2

，

我们有0。对于任何指数i，如0，通过选择j+k，

，

所以

，

同样，Ci−ck+γki+≤0。因此，如果我们让k+=（k−k−）j+，那么是双重可行的。与单纯形算法一样，θ+由

，

而u+也可以用单纯形算法计算

u−θj+γij+如果i∈k

+=\_θj i+，如果i=j+。

用户界面

0，如果i/∈k j+

主程序和双程序的目标函数的变化（这是相同的，因为选择了Uk=A−K1b和Y=CKA−K1，这样cu=ckuk=yb）与单纯形算法中的变化相同，即

.

我们有θ+>0和0，所以如果为0，则双程序的目标函数将严格减小。

案例（B3）。礹+=0.

可能出现礹+=0，即，=0的可能性。在这种情况下，目标函数不变。这是退化的情况，类似于单纯形算法中出现的退化。我们仍然选择J+∈N（μ+），但是我们需要一个防止循环的轴规则。存在此类规则；见Bertsimas和Tsitiklis[21]（第4.5节）和Papadimitriou和Steiglitz[130]（第3.6节）。

读者肯定注意到，双单纯形算法与单纯形算法非常相似，除了单纯形算法保留了（u，k）可行的性质，而双单纯形算法保留了y=cka-k1是双可行的性质。人们可能会怀疑，双单纯形算法是否等同于应用于双问题的单纯形算法。事实上，在对偶问题上，对偶单纯形算法和单纯形算法之间存在一对一的对应关系。Papadimitriou和Steiglitz【130】（第3.7节）中描述了这种对应关系。

如果我们使用（完整的）tableaux来描述这些方法，那么最好地说明单纯形算法和双单纯形算法之间的比较。

回想一下，（完整）表格是一个（m+1）×n+1）矩阵，其组织如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 | 网络错误 | 网络错误  网络错误 |
| 网络错误 |  |  |  |  |  |

······

顶行包含目标函数的当前值和降低的成本，第一列（除了顶行）包含当前基本解决方案UK的组件，其余列（除了顶行）包含向量。观察到与k中指数j对应的γkj构成单位矩阵im的置换。表格和新的基础k+=（k-k-）j+包含计算新的英国+、新的和新的降低成本所需的所有数据。

在执行单纯形算法时，所有k∈k都有Uk≥0（所有j/∈k都有Uj=0），输入列j+通过选择其中一个列索引来确定，这样cj>0。然后，通过查看其中0（沿J+列）的最小比率来确定离开列的索引k−。

另一方面，在执行双单纯形算法时，所有j/∈k的Cj≤0（所有k∈k的Ck=0），输出列k−通过选择其中一个行索引来确定，这样uk<0。进入列的索引j+通过查看比率−cj/γkj−的最大值来确定，其中0（沿行k−）。

关于单纯形算法和双单纯形算法的比较，可以在Bertsimas和Tsitiklis[21]以及Papadimitriou和Steiglitz中找到更多细节。

〔130〕。

下面是一个双单纯形方法的例子。

例46.2。考虑以下标准形式的线性程序：

最大化−4x1−2x2−x3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 |  |  | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误  网络错误  网络错误 |

0\_

0\_

0\_\_

我们用（u，k）初始化双单纯形程序，其中u=−3和k=（4,5,6）。

 

−4\_

1.在明确计算降低成本之前，初始表是

.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
|  |  |  | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 |

-

因为u有负坐标，所以情况（b）适用，我们将设置k−=4。我们现在必须确定案例（b1）或案例（b2）是否适用。通过扫描表格中的前三列，观察每一列都有一个负的条目，可以完成这一确定。因此，情况（b2）是适用的，我们需要确定降低的成本。观察C=−4、−2、−1,0,0,0），这反过来意味着C（4,5,6）=（0,0,0）。方程式（2）表明，非零削减成本为

，

我们的舞台变成了

.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
|  |  |  | 网络错误  网络错误 | 网络错误  网络错误 | 网络错误  网络错误 | 网络错误  网络错误 |

-

由于k−=4，我们的轴列是表的第一行。为了确定J+的候选项，我们扫描此行，定位负条目并计算

.

因为μ+出现在j=2时，我们设置j+=2。我们的新基础是K+=（2,5,6）。我们必须规范化tableau的第一行，即乘以−1，然后将此规范化行的两倍添加到第二行，并从第三行减去规范化行，以获得更新的tableau。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误  网络错误 | 网络错误 | 网络错误 |  |  | 网络错误  网络错误 | 网络错误  网络错误 |

−−

它仍然需要更新降低的成本和目标函数的值，方法是将规范化行的两倍添加到顶层。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误  网络错误 |  |  | 网络错误  网络错误 | 网络错误  网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

我们现在重复案例（b2）的过程，并设置k−=6（因为这是唯一一个u+的负输入）。我们的Pivot行现在是更新后的TableAux的第三行，新的祄+变成了

，

这意味着j+=3。因此，新的基础是k+=（2,5,3），我们更新了表格，取第3行，将第3行标准化后的两倍添加到第1行，将第3行标准化后的三倍添加到第2行。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误  网络错误 | 网络错误  网络错误 |  | 网络错误  网络错误 |  |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

它仍然需要更新目标函数，并通过将标准化行的五倍添加到顶层来降低成本。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误  网络错误 | 网络错误  网络错误 |  | 网络错误  网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

由于u+没有负项，双单纯形方法终止，目标函数

4x1−2x2−x3最大化为at（0，

## 46.6原始对偶算法

设（p2）为标准形式的线性程序

最大化cx，以ax=b且x≥0为准，

其中a是m×n矩阵的秩m，（d）是由

根据Ya≥C将Yb最小化，

式中y∈（rm）。

首先，我们可以通过改变每一个bi<0的方程来假设b≥0。

. 如果我们恰好有对偶程序（d）的一些可行解y，我们从定理46.12知道，（p2）的可行解x是（p）中方程的最优解。如果我们用j表示1，…，n的子集，其中的等式

yaj=cj

那么根据定理46.12，（p2）的可行解x是iff的最优解。

xj=0表示所有j/∈j。

设j=p和n 1，…，n−j。上面建议寻找x∈rn，这样

xxjaj=b

J·J·J

所有j的xj≥0∈j xj=0所有j/∈j，

或同等

ajxj=b，xj≥0，（1）

且xn=0n−p。

为了搜索这样一个x，只需要寻找一个可行的xj，为此，我们可以使用受限的原始线性程序（rp），定义如下：

最大化−（ξ1+····+ξm）

服从和x，ξ≥0。

由于假设b≥0，且目标函数在0上有界，该线性规划有一个最优解（x j，ξ）。

如果ξ=0，则由u j=x j和un=0n p给出的向量u rn是（p）的最优解。

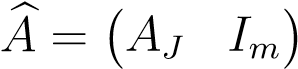
否则，ξ>0，我们无法解决（1）。然而，我们可以尝试使用ξ来改进y。为此，考虑（rp）的双（drp）：

最小化zb，以zaj≥0 z≥−1>m为准。

观察程序（DRP）与原始双程序（D）具有相同的目标功能。我们通过定理46.11知道（rp）的最优解（x j，ξ）产生（drp）的最优解z，这样

.

事实上，如果k是与（）关联的基础，并且如果我们



然后根据定理46.11我们得到

，

其中（ck）（p+1，…，p+m）表示与最后m列对应的最终表格中降低成本的行向量。

如果我们写y（θ）=y+θz，

则（d）的目标函数的新值为

Y（θ）b=Yb+θz\_b，（2）

由于z b<0，我们有机会改进（d）的目标函数，也就是说，如果y（θ）对（d）是可行的，那么减小θ>0的值就足够小。这是iff y（θ）a≥c iff的情况。

ya+θz a≥c.（3）

既然y是（d）的可行解，我们有yA≥c，所以如果z a≥0，那么（3）满足，y（θ）是（d）的解，对于所有θ>0，这意味着（d）是无界的。但这意味着（p）是不可行的。

让我们仔细看看不等式z a≥0。对于j j，由于z是（drp）的最优解，我们知道z aj≥0，所以如果z aj≥0对于所有j n，则（p）是不可行的。

否则，有一些j∈n=1，…，n−j这样

Z AJ<0，

然后根据j的定义，我们得到所有j∈n的yaj>cj，如果我们选取θ>0，那么

，

-

然后降低（d）的目标函数y（θ）b=yb+θz\_b（因为z\_b<0）。因此我们选择最好的θ，即

.（4）

接下来，我们将y更新为y+=y（θ+）=y+θ+z，并使用新的子集创建新的受限primal

j+=j∈1，…，n y+aj=cj，

然后重复这个过程。下面是原始对偶算法的步骤。

第1步。找到双程序（D）的一些可行解Y。稍后我们将展示这是永远可能的。

第2步。计算

j+=j∈1，…，n yaj=cj。

第3步。设置j=j+并使用simplex算法解决问题（rp），从上一轮确定的最优解开始，以k为基得到最优解（x j，ξ）。

第4步。

如果ξ=0，则停止（p）的最优解u，使u j=x j和u的其他组分为零。

否则就让

，

是（x j，ξ）与基k对应的（drp）的最优解。

如果所有j/∈j的z aj≥0，则停止；程序（p）没有可行的解。Else计算

，

和

j+=j∈1，…，n y+aj=cj。

回到步骤3。

下面的命题表明，在每次迭代中，我们都可以用在前一次迭代中获得的最优解来启动程序（RP）。

提案46.13。每个j j，使aj在最优解的基础上，ξ属于下一个指标集j+。

证据。这样一个指数j∈j对应于一个变量ξj，使得ξj>0，因此通过互补松弛度，双程序（drp）的约束z aj≥0必须是一个等式，即z aj=0。但是，我们有

y+aj=yaj+θ+z aj=cj，

说明J∈J+。

如果（u，ξ）以k为基础是程序（rp）的最优解，命题

46.13加上定理46.11的最后一个性质，我们可以用（u，ξ）k作为初始解（以k为基）重新启动步骤3中的（rp）。对于每个j∈j−j+，删除j列，对于每个j∈j+−j，新列aj通过将ab−k1和aj相乘来计算，但它是矩阵[1:m；p+1:p+m]，由最后表格中的最后m列组成，新的减少的cj由cj−z aj给出。重复使用以前的最优解（RP）可以显著提高效率。

另一个重要的观察是，对于任何指数j0∈n，比如θ+=（yaj0−cj0）/（−z aj0），我们都有

y+aj0=yaj0+θ+z aj0=cj0，

所以j0∈j+。这一事实用于确保原始对偶算法以有限的步骤终止（使用防止循环的轴规则）；见Papadimitriou和Steiglitz[130]（定理5.4）。

如何选取对偶规划（D）的一些初始可行解Y仍有待讨论。如果j=1，…，n的cj≤0，那么我们可以选择y=0。

我们应该注意，在许多应用中，自然的原始优化问题实际上是最小化一些目标函数cx=c1x1+·····+cnxn，而不是最大化。例如，Papadimitriou和Steiglitz[130]中考虑的许多优化问题都是最小化问题。

当然，最小化cx等同于最大化−cx，所以我们的演示也包括最小化。但是，如果我们处理的是最小化问题，那么权重Cj通常是非负的，因此从最大化的角度来看，我们将所有j的−Cj≤0，并且我们可以使用y=0作为起点。

回到最大化形式的原问题和最小化形式的对偶问题，我们仍然需要处理一些j的cj>0的情况，在这种情况下，（d）可能没有任何明显的y可行。最好是我们想找到这样一个非常便宜的Y。

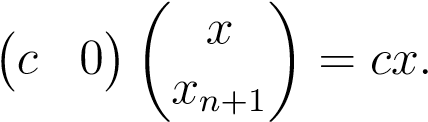
处理这种情况有一个诀窍。我们选取一些非常大的正数m，把新的方程加在方程组ax=b上。

x1+·····+xn+xn+1=m，

新变量xn+1被约束为非负。如果程序（P）有一个可行的解，则存在这样一个M。事实上，它可以证明，对于任何基本可行解u=（u1，…，un），每个ui\_都被一些仅依赖于a和b的表达式所限定；参见papadimitriou和steiglitz[130]（引理2.1）。证明并不困难，它依赖于这样一个事实：矩阵的逆矩阵可以用某些行列式（调节器）来表示。不幸的是，这个绑定包含m！作为一个因素，这使得它非常不切实际。在添加了上面的新方程之后，我们得到了新的方程组。

，

x≥0，xn+1≥0，新的目标函数由



上述线性程序的对偶是

根据yaj+ym+1≥cj j j=1，…，n ym+1≥0，将yb+ym+1m最小化。

如果一些j的cj>0，观察y给出的线性形式

（yei=max1≤j≤n cj>0≤

0如果1米

是一个可行的解决方案，新的双方案。实际上，我们可以选择m是一个接近所用计算机上可表示的最大整数的数字。

下面是T.Molla的数学588课堂笔记中给出的原始对偶算法的一个例子。

例46.3。考虑以下标准形式的线性程序：最大化−x1−3x2−3x3−x4

以且x1、x2、x3、x4≥0为准。

相关的双程序（D）是

最小化2

从属于。

我们用双可行点y=（-1/3 0 0）初始化原始对偶算法。注意，只有（d）的第一个不等式实际上是一个等式，因此j=1。我们形成了限制性原始程序（RP1）最大化−（ξ1+ξ2+ξ3）

服从且x1，ξ1，ξ2，ξ3≥0。

我们现在通过单纯形算法求解（rp1）。K=（2,3,4）和J=1的初始表格为

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

.

对于（rp1），c=（0、−1、−1、−1），（x1，ξ1，ξ2，ξ3）=（0,2,1,4），并且非零折减成本由下式给出：

.

因为只有一个非零的降低成本，我们必须设置j+=1。由于minξ1/3，ξ2/3，ξ3/6=1/3，我们看到k−=3和k=（2,1,4）。因此，我们以红色圆圈3为轴（即我们将第2行除以3，然后从第1行减去3×（第2行），从第3行减去6×（第2行），从第0行减去12×（第2行），得到表格

.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误  网络错误 | 网络错误  网络错误 |  | 网络错误  网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

-

在此阶段，（RP1）的单纯形算法终止，因为没有正的降低成本。由于最后一个表格的左上角不是零，所以我们继续执行原始对偶算法的步骤4并计算

，

所以

我们得出结论，（d）的新可行解是

.

当我们将y+代入（d）时，我们发现前两个约束是相等的，新的j是j=1,2。新减少的原始值（Rp2）最大化−（ξ1+ξ2+ξ3）

服从且x1，x2，ξ1，ξ2，ξ3≥0。

再次，我们通过simplex算法求解（rp2），其中c=（0,0、-1、-1、-1）、（x1，x2，ξ1，ξ2，ξ3）=（1/3,0,1,0,2）和k=（3,1,5）。初始表au是通过添加与变量x2对应的列，即

|  |  |
| --- | --- |
| 网络错误 | 网络错误 |

我们得到

.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误  网络错误 | 网络错误  网络错误  网络错误 |  | 网络错误  网络错误 | 网络错误  网络错误  网络错误 | 网络错误  网络错误  网络错误 |

-

注意，J+=2，因为第2列中只有正的成本减少。还观察到，由于minξ1/6，ξ3/8=ξ1/6=1/6，我们设置k−=3，k=（2,1,5）并沿红色6旋转以获得表

.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误  网络错误 | 网络错误  网络错误 | 网络错误  网络错误 |  | 网络错误  网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |

−−

由于降低的成本为零或负，单纯形算法终止，我们计算

Z=（−1−1−1）−（−7/3−5/3 0）=（4/3 2/3−1），

，

所以

.

当我们将y+插入（d）时，我们发现第一、第二和第四个约束是相等的，这意味着j=1,2,4。因此，新的限制初生体（Rp3）最大化−（ξ1+ξ2+ξ3）

服从且x1、x2、x4、ξ1、ξ2、ξ3≥0。

（rp3）的初始表au，其中c=（0,0,0，−1、−1、−1、−1）、（x1、x2、x4、ξ1、ξ2、ξ3）=（4/9,1/6、0,0,0,2/3）和k=（2,1,6），通过添加与变量x4对应的列（即

具有

，

我们得到

.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 | 网络错误 |
| 网络错误 | 网络错误 | 网络错误 | 1/3 | 7/3 | 5/3 | 零 |
| x2=1/6 | 零 | 一 | 1  /  3 | 1/6 | 1/6 | 零 |
| x1=4/9  ξ3=2/3 | 一  零 | 零  零 |  | 1/9  4/3 | 2/9  2/3 | 零  一 |

−−

由于第3列中只有正成本减少，所以我们设置j+=3。此外，由于min x2/（1/3），ξ3/（1/3）=x2/（1/3）=1/2，我们让k−=2，k=（3,1,6），并围绕红色圆圈1/3旋转以获得

.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X4 | 蔡1 | 蔡2 | 蔡3 |
| 1/2 | 零 | 1 | 零 | 5/2 | 3/2 | 零 |
| x4=1/2 x1=1/2 | 零  一 | 三  1/3 | 一  零 | 1/2  1/6 |  | 零  零 |
| ξ3=1/2 | 零 | 一 | 零 | 3/2 | 1/2 | 一 |

−−−

在这个阶段，没有正的降低成本，我们必须计算

Z=（−1−1−1）−（−5/2−3/2 0）=（3/2 1/2−1），

−3−3

（−1/6 1/2−1/3）6+3=13/2，−（3/2 1/2−1）6=3/2，

0 0

所以

.

我们将y+插入（d）中，发现第一、第三和第四个约束是相等的。

因此，j=1,3,4和受限的原始值（rp4）是

最大化−（ξ1+ξ2+ξ3）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 从属于 |  | 63  零 | 1-11个 | 一  零  零 | 0 1 1 | x1\_  0 x3 2 x4 1和x1，x3，x4，ξ1，ξ2，ξ3≥0。  0\_\_ξ1\_=  1 4  \_ξ2\_ξ3 |

（rp4）（c=（0,0,0，−1、−1、−1、−1）、（x1，x3，x4，ξ1，ξ2，ξ3）=（1/2，0,1/2，0,0,0,1/2）和k=（3,1,6）的初始表au是通过将变量x2对应的列替换为变量x3对应的列，即用

，

我们得到

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X3 | X4 | 蔡1 | 蔡2 | 蔡3 |
| 1/2 | 零 | 3/2 | 零 | 5/2 | 3/2 | 零 |
| x4=1/2 x1=1/2 | 零  一 |  | 一  零 | 1/2 1/6 |  | 零  零 |
| ξ3=1/2 | 零 | 3  /  2 | 零 | 3/2 | 1/2 | 一 |

.

通过分析降低成本的顶行，我们发现j+=2。此外，由于min x1/（1/2），ξ3/（3/2）=ξ3/（3/2）=1/3，我们让k−=6，k=（3,1,2），并沿着红色圆圈3/2旋转以获得

.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X3 | X4 | 蔡1 | 蔡2 | 蔡3 |
| 零 | 零 | 零 | 零 | 1 | 1 | 1 |
| x4=2 x1=1/3 x3=1/3 | 零  一  零 | 零  零  一 | 一  零  零 | 4  2/3  一 | 2  1/3  1/3 | 三  -21//33 |

−−

因为最后一个画面的左上角是零，并且降低的成本都小于等于0，所以我们最终完成了。那么y=（19/3 8/3−14/3）是（d）的最优解，但更重要的是（x1，x2，x3，x4）=（1/3,0,1/3,2）是我们最初线性程序的最优解，并提供−10/3的最优值。

线性规划的原对偶算法似乎不是目前求解线性规划的常用方法。但重要的是，它的基本原理是，使用一个限制（简单）的初等问题，涉及一个具有固定权重的目标函数，即1，以及对偶问题，通过改进对偶的目标函数来向初等问题提供反馈，从而产生了一类组合子。基于原始对偶范式的所有算法（通常是近似算法）。读者将通过参考Papadimitriou和Steiglitz[130]来了解这种算法，其中解释了Dijkstra的最短路径问题算法以及Ford和Fulkerson的最大流量算法是如何从原始双P中推导出来的。阿拉迪格姆。

第八部分

非线性优化

一千五百零三

第四十七章

# 希尔伯特空间基础

第48章中证明的关于实值函数极小值存在的大多数“深层”结果依赖于希尔伯特空间理论的两个基本结果：

1. 投影引理是希尔伯特空间v的非空闭凸子集的结果。
2. Riesz表示定理，它允许我们用V中的向量和V上的内积来表示希尔伯特空间V上的连续线性形式。

在拉格朗日对偶中出现的卡鲁什-库恩-塔克条件的正确性来自于一个版本的farkas–minkowski命题，也来自投影引理。

因此，我们认为回顾希尔伯特空间理论的一些基本结果是必不可少的，尽管在这里讨论的大多数应用中，希尔伯特空间是有限维的。然而，在优化理论中，有许多问题，我们寻求一个函数来最小化某种类型的能量泛函（通常由双线性形式给出），在这种情况下，我们要处理无限维希尔伯特空间，因此有必要开发工具来处理更通用的无限维希尔伯特空间的情形。

## 47.1投影引理，对偶性

假设Hermitian空间定义为kuk=phe，\_（u，ui，我们在第13.1节中显示函数），是e上的范数。因此，e是赋范向量空间。ifk k k:e→er也是完整的，这是一个非常有趣的空间。

还记得完整性与柯西序列的收敛性有关。赋范向量空间He，Ukk Ki（关于赋范向量空间和度量的定义，见第36章），是度量D下的度量空间，定义如下：

d（u，v）=kv−space，或lang[108，109]或dixmier[52]）。给定一个度量空间e和度量d，一个序列

一千五百零五

（a n）n≥1的元素an∈e是一个柯西序列iff，对于每>0，有一些n≥1这样

对于所有m，n≥n。

我们说e是完全的，如果每个柯西序列收敛到一个极限（这是唯一的，因为度量空间是hausdorff）。

R或C上的每个有限维向量空间都是完整的。例如，可以通过归纳法证明，给定e的任何基（e1，…，en），线性映射h:cn→e定义为h（（z1，…，zn））=z1e1+······+znen

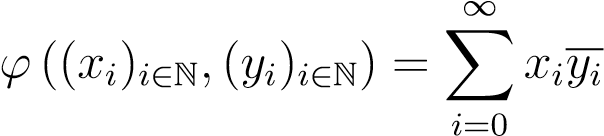
是同态（使用cn上的sup-norm）。我们还可以利用这样一个事实，即R或C上有限维向量空间上的任何两个范数都是等效的（见第8章，或lang[109]、Dixmier[52]、Schwartz[146]）。

但是，如果e有无穷大的维数，它可能不完整。当厄米空间完成时，有限维厄米空间的一些性质也适用于无限维空间。例如，任何封闭子空间都有一个正交补，特别是有限维子空间有一个正交补。在分析中，完备的埃尔米特空间也起着重要作用。由于希尔伯特首先研究了它们，所以它们被称为希尔伯特空间。

定义47.1.一个（复杂）厄米空间He，i是由\_引起的范数k k k下的完全范数向量空间，称为希尔伯特空间。一个真正的欧几里得空间He，i，在由\_诱导的k k范数下完成，称为真正的希尔伯特空间。

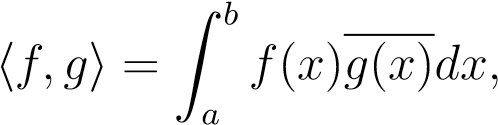
本节中的所有结果适用于复杂的希尔伯特空间以及真实的希尔伯特空间。我们只陈述复杂案例的所有结果，因为它们也适用于真实案例，而且复杂案例中的证明需要更多的关注。

例47.1。所有可数无穷序列的空间L2，复数的x=（Xi）i~（n），即希尔伯特空间。稍后将显示，图\_：l2×l2→c定义如下：



定义得很好，并且l2是\_下的希尔伯特空间。事实上，我们将证明一个更普遍的结果（提案A.3）。

例47.2。光滑函数f[a，b]→c的集合c∞[a，b]是厄米形式下的厄米空间。



但它不是希尔伯特空间，因为它不完整。可以构造它的完备l2（[a，b]），它是上勒贝格可积函数的空间。

[a，b]。

定理36.63给出了一个事实的快速证明，即任何厄米空间e（带有厄米积H−、−I）都可以嵌入到希尔伯特空间e h中。

希尔伯特空间定理47.1.（e h，给出了赫米特空间h−、−i）和线性映射（e，ω：h−e，→−ie）h（，因此相应地。欧几里得空间），有一个

H

h u，vi=h\_（u），（v）ih

对于所有的u，v∈e，和（e）在eh中是稠密的。此外，eh在同构方面是独一无二的。

证据。设（e，b k k eb）为Banach空间，Letp瓒：e→eb为线性等距，给定

根据定理36.63。让kuk=hu、ui和e−hi=可以用normeb表示。如果e是一个实向量空间，我们知道

根据第11.1节，极性方程得出的内积h−

，

如果e是一个复杂的向量空间，我们从13.1节知道，我们有极性方程。

.

根据柯西-施瓦兹不等式，−i:e×e→r）是连续的。但是，它不是均匀连续的，但是我们可以hu，vi≤kukkkvk，地图H−，−i:e×e→c（resp.

h-通过使用极性方程将问题扩展到一个连续的映射来解决这个问题。

由一个正定埃尔米特内积（resp.欧几里得内积）通过连续性，极性方程也包含油墨keb i eh，这表明h−、−i−在eh h−上延伸到h。

延伸H−，−。

备注：我们按照Schwartz[145]中的方法（第二十三章第42节）。定理2）。其他方法见Munkres[127]（第7章第43节）和Bourbaki[27]。

有限维厄米特（和欧几里得）空间最重要的事实之一是它们有正交基。这意味着，在同构之前，每个有限维厄米空间都同构于cn（对于某些n∈n），并且内积由

.

此外，每个子空间w都有一个正交补码w，并且内积

归纳出e和e（实际上，e和e）之间的自然对偶性，其中e是e上线性形式的空间。

当e是希尔伯特空间时，e可能是无限维，通常是不可数维。因此，我们不能期望E总是有一个正交基。但是，如果我们修改基的概念，使“希尔伯特基”是一个在e中也很密集的正交族，也就是说，每个v∈e都是希尔伯特基向量的有限组合序列的极限，那么我们就可以恢复有限维的大部分“好”性质。艾尔·赫米特的同事空间。例如，如果（ck=hv，uki/kuukk）k k∈，那么k是希尔伯特基，因为everyv是其傅立叶级数v∈e的“和”，我们可以定义fourierpk∈k ckuk。然而，索引集k的基数可能非常大，因此有必要定义由k索引的一系列向量的可和性意味着什么。我们将在A.1节中进行此操作。结果表明，每个希尔伯特空间都同构于形式为l2（k）的空间，其中l2（k）是示例47.1空间的推广（见定理A.8，通常称为Riesz-Fischer定理）。

我们的第一个目标是证明希尔伯特空间的闭子空间具有正交补。我们还表明，如果我们将e的对偶e0重新定义为e上连续线性映射的空间，则对偶性是成立的。我们的演示与Bourbaki[27]密切相关。我们也受到了鲁丁[136]、朗[108，109]、施瓦茨[146，145]和迪克斯米尔[52]的启发。事实上，我们强烈推荐Dixmier[52]作为一个关于拓扑和分析基础的清晰而简单的文本。我们首先证明了所谓的投影引理。

回想一下，在度量空间e中，e的子集x对于每个收敛序列都是闭的iff

（xn）点xn∈x的极限x=limn→∞xn也属于x。x的闭包x是点xn∈x的收敛序列（xn）的所有极限的集合。显然，x x。我们

假设e的子集x在e iff e=x中是稠密的，x的闭包，这意味着每一个a∈e都是点xn∈x的某些序列（xn）的极限，凸集将再次起到至关重要的作用。

首先，我们陈述下面简单的“平行四边形不等式”，其证明留作练习。

提案47.2.如果e是厄米空间，对于任意两个向量u，v∈e，我们有

ku+vk2+ku−vk2=2（kuk2+kvk2）。

从上面我们得到了以下建议：

提案47.3.如果e是厄米空间，给定任意d，δ∈r，使0≤δ<d，设

b=u∈e kuk<d和c=u∈e kuk≤d+δ。

对于任何凸集，如a c−b，我们有

Kv−Uk≤√12dδ，

对于所有u，v∈a（见图47.1）。

*C*

*B*

*A*

*u*

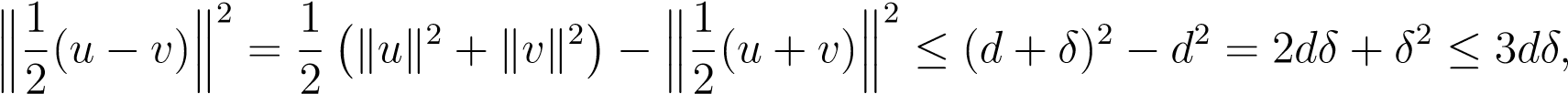
*v*

图47.1：命题47.3的不等式

证据。因为a是凸的，所以…从形式上写的平行四边形不等式

，

因为δ<d，我们得到



从哪来的

Kv−Uk≤√12dδ。

如果x是度量空间（e，d）的非空子集，对于任何a∈e，回想一下，我们将a到x的距离d（a，x）定义为

d（a，x）=inf d（a，b）。BX

此外，x的直径δ（x）定义为

δ（x）=sup d（a，b）a，b∈x。

可能δ（x）=∞。我们将以下标准的两个事实作为练习（见Dixmier[52]）：

**Proposition 47.4.** *Let E be a metric space.*

1. *For every subset X* ⊆ *E, δ*(*X*) = *δ*(*X*)*.*
2. *If E is a complete metric space, for every sequence* (*Fn*) *of closed nonempty subsets of E such that Fn*+1 ⊆ *Fn, if* lim*n*→∞ *δ*(*Fn*) = 0*, then*  *consists of a single point.*

We are now ready to prove the crucial projection lemma.

**Proposition 47.5.** *(Projection lemma) Let E be a Hilbert space.*

1. *vectorFor any nonempty convex and closed subsetpX*(*u*) ∈ *X such that X* ⊆ *E, for any u* ∈ *E, there is a unique*

k*u* − *pX*(*u*)k = *v*inf∈*X* k*u* − *v*k = *d*(*u,X*)*.*

*See Figure 47.2.*

1. *The vector pX*(*u*) *is the unique vector w* ∈ *E satisfying the following property (see Figure 47.3):*

*w* ∈ *X and* <h*u* − *w,z* − *w*i ≤ 0 *for all z* ∈ *X.* (∗)

1. *If X is a nonempty closed subspace of E then the vector pX*(*u*) *is the unique vector w* ∈ *E satisfying the following property:*

*w* ∈ *X and* h*u* − *w,z*i = 0 *for all z* ∈ *X.* (∗∗)

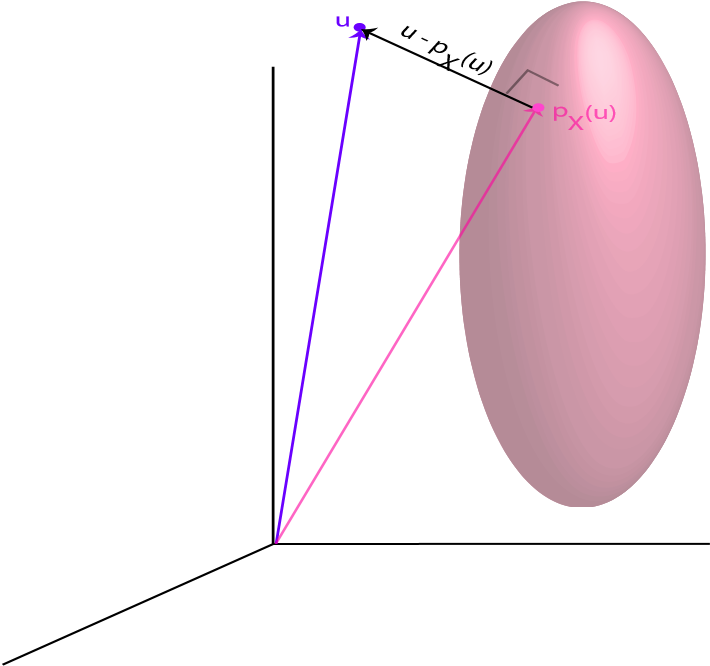
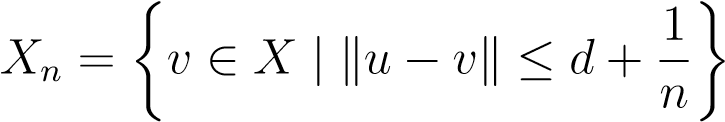


Figure 47.2: Let *X* be the solid pink ellipsoid. The projection of the purple point *u* onto *X* is the magenta point *pX*(*u*).

*Proof.*follows: for every(1) Let *d* =*n*inf≥ 1*v*∈,*X* k*u* − *v*k = *d*(*u,X*). We define a sequence *Xn* of subsets of *X* as

*.*

*X*

*w*

*u*

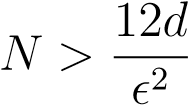
*z*

Figure 47.3: Inequality of Proposition 47.5

It is immediately verified that each *Xn* is nonempty (by definition of *d*), convex, and that *Xn*+1 ⊆ *Xn*. Also, by Proposition 47.3, we have

sup{k*w* − *v*k | *v,w* ∈ *Xn*} ≤ p12*d/n,*

and thus, T*n*≥1 *Xn* contains at most one point. We will prove that T*n*≥1 *Xn* contains exactly one point, namely, *pX*(*u*). For this, define a sequence (*wn*)*n*≥1 by picking some *wn* ∈ *Xn* for every *n* ≥ 1. We claim that (*wn*)*n*≥1 is a Cauchy sequence. Given any  *>* 0, if we pick *N* such that

*,*

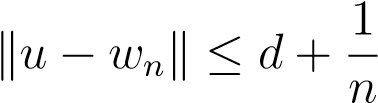
since (*Xn*)*n*≥1 is a monotonic decreasing sequence, which means that *Xn*+1 ⊆ *Xn* for all *n* ≥ 1, for all *m,n* ≥ *N*, we have



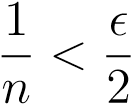
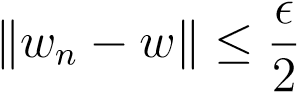
as desired. Since *E* is complete, the sequence (*wn*)*n*≥1 has a limit *w*, and since *wn* ∈ *X* and *X* is closed, we must have *w* ∈ *X*. Also observe that

k*u* − *w*k ≤ k*u* − *wn*k + k*wn* − *w*k*,*

and since *w* is the limit of (*wn*)*n*≥1 and

*,*

given any  *>* 0, there is some *n* large enough so that

 and *,*

and thus



Since the above holds for everyT  *>* 0, we have k*zu*∈− *Xw*ksuch that= *d*. Thus,k*uww*−, which we denote as∈*z*k*Xn*=for all*d*(*u,Xn*) =≥ 1*d*, which proves that *n*≥1 *Xn* = {*w*}*z*. Now, any= *w*, proving the uniqueness of also belongs to every *Xn*, and thus *pX*(*u*). See Figure 47.4.

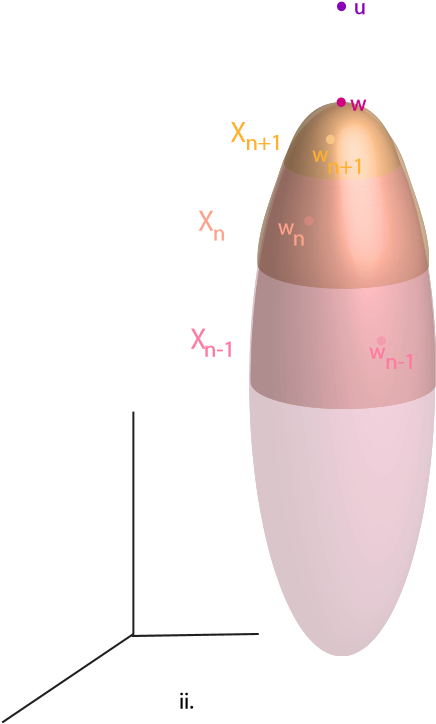
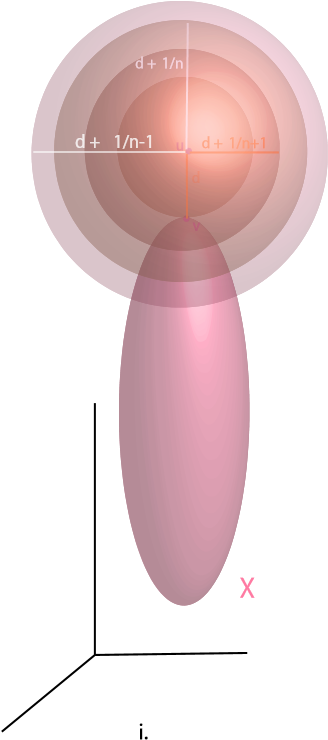
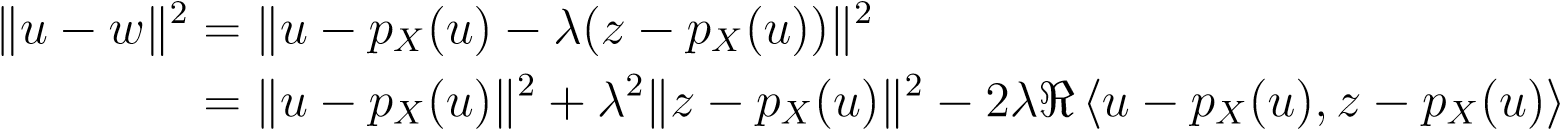


Figure 47.4: Let *X* be the solid pink ellipsoid with *pX*(*u*) = *w* at its apex. Each *Xn* is the intersection of *X* and a solid sphere centered at *u* with radius *d* + 1*/n*. These intersections are the colored “caps” of Figure ii. The Cauchy sequence (*wn*)*n*≥1 is obtained by selecting a point in each colored *Xn*.

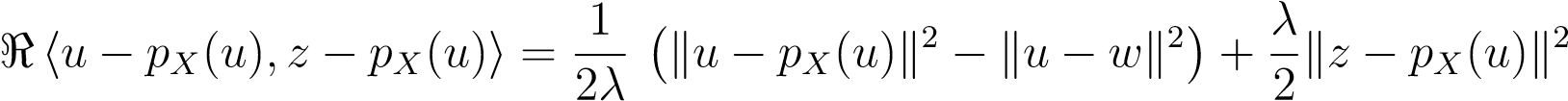
(2) Let *z* ∈ *X*. Since *X* is convex, *w* = (1 − *λ*)*pX*(*u*) + *λz* ∈ *X* for every *λ*, 0 ≤ *λ* ≤ 1. Then, we have

k*u* − *w*k ≥ k*u* − *pX*(*u*)k

for all *λ*, 0 ≤ *λ* ≤ 1, and since

*,*

for all *λ*, 0 *< λ* ≤ 1, we get

*,*

and since this holds for every *λ*, 0 *< λ* ≤ 1 and

k*u* − *w*k ≥ k*u* − *pX*(*u*)k*,*

we have

<h*u* − *pX*(*u*)*,z* − *pX*(*u*)i ≤ 0*.*