组合优化与凸优化

第6章 约束最优化理论(Theory of constrained Optimization)

刘绍辉

计算机科学与技术学院 哈尔滨工业大学

shliu@hit.edu.cn

2023年春季

第6章约束最优化理论(Theory of Constrained Optimization)-回顾

- **无约束问题**
  - 一阶和二阶条件
- **有约束问题**
  - 不等式约束
  - 等式和不等式约束
  - 有约束问题的二阶充要条件
- **对偶问题**
- **总结**

◆牛顿法(Newton Methods)

- $g_k = \nabla f(x_k) \neq 0$,则$f(x) = f(x_k) + (x - x_k)^T \nabla f(x_k) + o(\|x - x_k\|)$,若记$x - x_k = \lambda d_k$,则满足$d_k^T g_k < 0$的方向为下降方向，$\lambda$取定后，$d_k^T g_k$的值越小，即$-d_k^T g_k$的值越大，函数下降越快，显然$|d_k^T g_k| \leq \|d_k^T\| \|g_k\|$，从而当且仅当$d_k^T = -g_k$时，$d_k^T g_k$最小，从而称$-g_k$是最速下降方向,此时迭代公式为$x_{k+1} = x_k - \lambda_k g_k$
- 最速下降法的收敛速度至少是线性的
- 由于一维搜索满足$g_{k+1}^T d_k = 0 = d_{k+1}^T d_k$,表明在相邻两个迭代点上函数的梯度方向是相互垂直的，因此会产生锯齿现象，接近极小点时，步长越来越小，前进愈慢

◆若用二次近似来逼近原目标函数，此时可得迭代公式为$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$,若令$G_k = \nabla^2 f(x_k), g_k = \nabla f(x_k)$, 则迭代公式为$x_{k+1} = x_k - G_k^{-1} g_k$

- 牛顿法在收敛点附近具有二阶收敛性，此时如果远离最优点，则G_k不正定，牛顿方向不一定是下降方向，其收敛性不能保证。说明采取步长恒为1的牛顿法是不合适的，因而需要采取一维搜索求取最优步长，但应强调的是，仅当步长因子$\{\lambda_k\}$收敛到1时，牛顿法才是二阶收敛的。此时，如果函数是严格凸函数，则带步长的牛顿法是整体收敛的。
- 牛顿法主要问题是Hessian矩阵G_k不正定，这时二次模型不一定有极小点，甚至没有平稳点，当G_k是不定时，二次模型无界，因此提出了很多修正的办法

◆ Goldstein和Price修正办法

➢ 当G_k不正定时，采用最速下降方向$-g_k$,结合角度策略$\theta \le \frac{\pi}{2} - \mu$,参数$\mu > 0$, 此时$\theta$是$-g_k$与$d_k$之间的某个角度，方向$d_k = \begin{cases} -G_k^{-1} g_k, \text{如果} cos\theta \ge \eta \\ -g_k, \text{其它} \end{cases}$,$\eta > 0$为给定的常数。此时搜索方向总满足$cos(d_k, -g_k) \ge \eta$

◆ Goldfeld方法

➢ 不采用最速下降方向替换牛顿方向，而是让牛顿方向$-G_k^{-1} g_k$偏向最速下降方向$-g_k$.明确来讲，就是将$Hessian$矩阵G_k改为$G_k + v_k I$,$v_k > 0$,使得$G_k + v_k I$正定，v_k的准则就是不要太大，使$G_k + v\, I$正定的最小v

➢ 其步骤一般如下：$\bar{G}_k = G_k + v_k I$,若G_k正定，则$v_k = 0$,否则$v_k > 0$,解$\bar{G}_k d = -g_k$,求得d_k,迭代公式:$x_{k+1} = x_k + d_k$

➢ 信任域方法或有限步长法都与这有关

◆ 有限差分牛顿方法(Finite-Difference Newton's Method)

➢ 使用有限差分作为牛顿法中导数和Hessian矩阵的近似

第6章约束最优化理论(Theory of Constrained Optimization)-回顾

- **负曲率方向法(Negative Curvature Direction Method)**
- **前面提到牛顿法和修正牛顿法远离最优点时，Hessian矩阵不正定，尤其在靠近鞍点处，$\nabla f(x_k)=0$,且$\nabla^2 f(x_k)$不是半正定的情况**
- **定义：设$f: R^n \to R$在开集D上二次连续可微，**
  - 如果$\nabla^2 f(x)$至少有一个负特征值，则$x \in D$叫做不定点；
  - 如果x是一个不定点，若方向d满足$d^T\nabla^2 f(x)d<0$,则称d为$f(x)$在x处的负曲率方向
  - 如果$s^T\nabla f(x)\le 0, d^T\nabla f(x)\le 0, d^T\nabla^2 f(x)d<0$,则向量对$(s,d)$称为不定点$x$处的下降对；若$x$不是一个不定点，则满足$s^T\nabla f(x)<0, d^T\nabla f(x)\le 0, d^T\nabla^2 f(x)d=0$称为在点$x$处的下降对
- **下降对的例子，如$s=-\nabla f(x), d=\begin{cases}0, 如果\nabla^2 f(x)\ge 0\\ -sign(u^T\nabla f(x))u\end{cases}$,其中$u$是对应于$\nabla^2 f(x)$的负特征值的单位特征向量**
- **显然，仅仅当$\nabla f(x)=0$和$\nabla^2 f(x)$半正定时，不存在下降对**
- **并且，从定义可以看出，在稳定点，负曲率方向必为下降方向。在一般点处，若负曲率方向d满足$d^T\nabla f(x)=0$,则d和$-d$均为下降方向；若$d^T\nabla f(x)\le 0$则d是下降方向，若$d^T\nabla f(x)>0$则$-d$是下降方向**
- **负曲率方向的牛顿法与普通牛顿法、修正牛顿法的主要区别：**
  - 对于非正定Hessian矩阵，先强迫正定，在接近稳定点时采用负曲率方向；或对于对称不定的Hessian矩阵，采用对称不定分解，直接产生负曲率方向
  - 当采用下降对时，在选择步长因子的一维搜索方法中，代替一阶步长准则，采用二阶步长准则

- **Gill-Murray稳定牛顿法**
- **基本思想：当G_k为不定矩阵时，采用修改的Cholesky分解强迫矩阵正定；当g_k趋于0时，采用负曲率方向使函数值下降**
- **令$\bar{G}_k = G_k + E_k = L_k D_k L_k^T, E_k = diag(e_{11}, \cdots, e_{nn}), D_k = diag(d_{11}, \cdots, d_{nn})$**
- **令$\psi_j = d_{jj} - e_{jj}, j = 1, \cdots, n$;求得$\psi_t = min\{\psi_j | j = 1, \cdots, n\}$;若$\psi_t \geq 0$,停止，否则解$L_k^T d = e_t$得到方向$d_k$,这里$e_t$是第$t$个元素为1的单位向量。注意：此时得到的方向$d_k$是$x_k$处的负曲率方向，并且$d_k, -d_k$中至少有一个是$x_k$处的下降方向**
- **1.给定初始点$x_0, \epsilon > 0, k = 1$;2.计算g_k, G_k;3.进行Cholesky分解$L_k D_k L_k^T = G_k + E_k$;4.若$\|g_k\| > \epsilon$,解方程$L_k D_k L_k^T d = -g_k$,求出搜索方向$d_k$，Goto 6;否则，Goto 5；5.求取负曲率方向，若不能求得方向$d_k(\psi_t \geq 0)$,则停止；否则，求出方向d_k,令$d_k = \begin{cases} -d_k, 若 g_k^T d_k > 0 \\ d_k, 其它 \end{cases}$；6.一维搜索求$\alpha_k$,使得$f(x_k + \lambda_k d_k) = min_{\lambda \geq 0} f(x_k + \lambda d_k)$,令$x_{k+1} = x_k + \lambda_k d_k$;7.若$f(x_{k+1}) \geq f(x_k)$,则停止计算；否则置$k = k + 1$,Goto2**
- **其它类似算法包括Fiacco-McCormick: $G_k = LDL^T$,Fletcher-Freeman：$P^T G_k P = LDL^T$等**

第6章约束最优化理论(Theory of Constrained Optimization)-回顾

◆不精确牛顿法

- 牛顿法计算量大，尤其是维数很大的问题
- $F(x)=0$的牛顿法求根迭代公式为$x_{k+1}=x_k-\frac{F(x_k)}{F'(x_k)}$,令$s_k=x_{k+1}-x_k$,则表示为$F'(x_k)s_k=-F(x_k)$,以此来求得优化中牛顿法的方向。
- 考虑不精确的情况，解$F'(x_k)s_k=-F(x_k)+r_k$,其中$\frac{\|r_k\|}{\|F(x_k)\|}\le\eta_k$,然后迭代公式为:$x_{k+1}=x_k+s_k$
- 当$\{\eta_k\}$一致小于1的情况下，不精确牛顿法局部收敛
- 这里的Inexact Newton法可以用于各种上述介绍的牛顿型方法，尤其对于大规模问题会收到很好的效果

◆共轭方向法

- 共轭方向法介于最速下降发和牛顿法之间的一个方法，仅仅需要利用一阶导数信息，克服了最速下降法收敛慢的特点，又避免了存储和计算牛顿法所需要的二阶导数信息
- 典型的共轭方向法包括共轭梯度法(Conjugate Gradient Method)和拟牛顿法
- 一般步骤
- 1.初始点x_0,计算$g_0=g(x_0)$;2. 计算d_0,使得$d_0^Tg_0<0$, 3. 令$k=0$;4.计算λ_k, x_{k+1},使得$f(x_k+\lambda_k d_k)=min_\lambda f(x_k+\lambda d_k), x_{k+1}=x_k+\lambda_k d_k$;5.计算共轭方向$d_{k+1}$: $d_{k+1}^T G d_j=0, j=0,1,\cdots,k$,6.令$k=k+1$,Goto 4
- 共轭方向法的性质：只要执行精确的一维搜索，就能得到二次终止性

- 例如对于二次函数$f(x)=\frac{1}{2}x^TGx+b^Tx+c$,显然$g(x)=Gx+b$,令$d_0=-g_0$,这时进行一维精确搜索得到$\lambda_0$,且有$g_1^Td_0=0$,令$d_1=-g_1+\beta_0d_0$,选择$\beta_0$: $d_1^TGd_0=0\Rightarrow\beta_0=\frac{g_1^TGd_0}{g_0^TGd_0}=\frac{g_1^T(g_1-g_0)}{\left(d_0^T(g_1-g_0)\right)}=\frac{g_1^Tg_1}{g_0^Tg_0}$,后续第$k$次迭代，令$d_k=-g_k+\sum_{i=0}^{k-1}\beta_id_i$,根据共轭性质，求得$\beta_{k-1}$.从而求得$d_k$，$x_{k+1}=x_k+\lambda_kd_k$再按照精确搜索求得$\lambda_k$.
- $d_{k+1}=-g_{k+1}+\beta_kd_k$,其中$\beta_k$的不同求法，对应不同的方法
- $\beta_k=\frac{||g_{k+1}||^2}{<g_{k+1}-g_k,d_k>}$(Dai-Yuan公式)
- $\beta_k=\frac{g_{k+1}^T(g_{k+1}-g_k)}{d_k^T(g_{k+1}-g_k)}$ $(Crowder-Wolfe$公式);
- $\beta_k=\frac{g_{k+1}^Tg_{k+1}}{g_k^Tg_k}$ $(Fletcher-Reeves$公式); +Wolfe-Powell改进条件2可保证下降
- $\beta_k=\frac{g_{k+1}^T(g_{k+1}-g_k)}{g_k^Tg_k}$ $(Polak-Ribiere-Polyak$公式);+$\beta_k=max\{\beta_k,0\}$自动重设可保证下降;
- $\beta_k=\frac{g_{k+1}^Tg_{k+1}}{d_k^Tg_k}$ $(Dixon$公式)

◆ **拟牛顿法:牛顿法成功的关键是利用了Hessian矩阵的曲率信息。但计算Hessian矩阵工作量大，或很难求得，这就导致仅利用目标函数一阶导数的方法，拟牛顿法就是这类方法，利用一阶导数信息构造目标函数曲率的近似，而不需要明显形成Hessian矩阵，同时具有收敛速度快的优点**

- 回忆$f: R^n \to R$的二次近似$f(x) \approx f(x_{k+1}) + g_{k+1}^T(x - x_{k+1}) + \frac{1}{2}(x - x_{k+1})^T G_{k+1}(x - x_{k+1})$,两边求导
- $g(x) \approx g_{k+1} + G_{k+1}(x - x_{k+1})$,同样令$x = x_k, s_k = x_{k+1} - x_k, y_k = g_{k+1} - g_k$,可得$G_{k+1}^{-1} y_k \approx s_k$,通过求解$y_k$获得下次迭代的方向
- 但如果Hessian不好计算，甚至不好求出，此时需要用近似的办法来满足上述的关系，即$H_{k+1} y_k = s_k$---(拟牛顿方程)，注意该式也可以用下述公式表示$B_{k+1} s_k = y_k$,相当于H_{k+1}是对Hessian逆矩阵的拟牛顿近似，而B_{k+1}是对Hessian矩阵的拟牛顿近似
- 一般步骤
- 1.给出$x_0 \in R^n, H_0 \in R^{n\times n}, 0 \le \epsilon, k = 0$;2. 如果$\|g_k\| \le \epsilon$,则停止，否则计算$d_k = -H_k g_k$;3. 沿方向$d_k$作线性搜索求得$\lambda_k > 0$:$x_{k+1} = x_k + \lambda_k d_k$;4.校正$H_k$产生$H_{k+1}$,使得$H_{k+1} y_k = s_k$条件成立；5.$k = k + 1$,Goto 2
- 与牛顿法比较，仅需一阶导数，H_k保持正定，是的方法具有下降性质，每次迭代需要$O(n^2)$次乘法，牛顿法需要$O(n^3)$次乘法

- **DFP法就是拟牛顿法中比较有名的一种方法**
- 采用$d_j=-G_j\nabla f(y)$, **来代替**牛顿法中的$-H^{-1}(y)\nabla f(y)$,
- $H_{j+1}=H_j+auu^T+bvv^T, u,v\in R^n, a,b$**为待定的标量,**
  - 由拟牛顿方程$H_{j+1}y_j=s_j\Rightarrow H_jy_j+auu^Ty_j+bvv^Ty_j=s_j$,注意a,b不唯一，但其显然可以取：$u=s_j, v=H_jy_j$,
  - 从而可得$a=\frac{1}{u^Ty_j}=\frac{1}{s_j^Ty_j}$,$b=-\frac{1}{y_j^TH_jy_j}$,
- **因此得秩2校正的DFP公式为**

$$H_{j+1}=H_j+\frac{s_js_j^T}{s_j^Ty_j}-\frac{H_jy_jy_j^TH_j}{y_j^TH_jy_j}$$

- **这种校正一般要求保持正定性**

◆DFP方法

- 初始化：给定$x_0 \in R^n, H_0 \in R^{n\times n}$为对称正定矩阵，$\epsilon > 0, k = 0$
- 第k步迭代：a) 若$||g_k|| \le \epsilon$,终止；b)计算$d_k = -H_k g_k$;c)计算步长α_k;d)$s_k = \alpha_k d_k, x_{k+1} = x_k + s_k, y_k = g_{k+1} - g_k, H_{k+1} = H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}$,e)k++.

◆DFP方法具有二次终止性:如果f是二次函数，G是其正定Hessian矩阵，那么当采用精确线性搜索时，DFP方法，具有遗传性质和方向共轭性质，即对于$i = 0, 1, \cdots, m$,有

- $H_{i+1} y_j = s_j, j = 0, 1, \cdots, i$(遗传性质);
- $s_i^T G s_j = 0, j = 0, 1 \cdots, i-1$(方向共轭性)
- 方法在$m + 1 \le n$步后迭代终止，如果$m = n - 1$,则$H_n = G^{-1}$
- 因此，DFP拟牛顿法是共轭方向法，如果初始近似H_0取作单位矩阵，则方法为共轭梯度法

◆ DFP方法

- 初始化：给定$x_0 \in R^n, H_0 \in R^{n\times n}$为对称正定矩阵，$\epsilon > 0, k = 0$
- 第k步迭代：a) 若$||g_k|| \le \epsilon$,终止；b)计算$d_k = -H_k g_k$;c)计算步长α_k;d)$s_k = \alpha_k d_k, x_{k+1} = x_k + s_k, y_k = g_{k+1} - g_k, H_{k+1} = H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}$,e)k++.

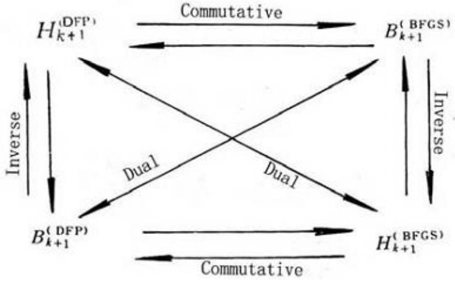
◆ DFP方法对一般的函数

- DFP更新保持对称正定
- 每次迭代需要$3n^2 + O(n)$次乘法
- **DFP拟牛顿法超线性收敛**
- 对于严格凸函数，精确线搜索下，DFP方法全局收敛

DFP：Davidon, Fletcher, Powell
BFGS：Broyden, Fletcher, Goldfarb, Shanno

- $H_{k+1}y_k = s_k$---(拟牛顿方程)，注意该式也可以用下述公式表示$B_{k+1}s_k = y_k$,相当于H_{k+1}是对Hessian逆矩阵的拟牛顿近似，而B_{k+1}是对Hessian矩阵的拟牛顿近似
  - 注意，如果要求更新中近似当前矩阵B_k,前满足对称性质和拟牛顿方程，则求B_{k+1}的过程可以形式化描述为：
  - $min_B||B - B_k||, s.t. B = B^T, Bs_k = y_k, s_k^T y_k > 0, B_k$对称正定
    - ✓范数任意，为Frobenius加权范数时对应DFP：$||A||_W = ||W^{\frac{1}{2}}AW^{\frac{1}{2}}||_F$，$||C||_F^2 = \sum_{i=1}^n \sum_{j=1}^n c_{ij}^2$，权矩阵只需要满足：$Wy_k = s_k$即可.例如$W = \overline{G_k}^{-1}$，$\overline{G_k} = \left[\int_0^1 \nabla^2 f(x_k + \tau\alpha_k p_k) d\tau\right], y_k = \overline{G_k}\,\alpha_k p_k = \overline{G_k} s_k$
  - $min_H||H - H_k||, s.t. H = H^T, Hy_k = s_k$, 范数为加权Frobenius范数，权矩阵$W$要求满足: $Ws_k = y_k$,则问题唯一解即为:BFGS

- DFP方法具有数值不稳定性，有时产生数值上奇异的Hessian近似，下面的BFGS校正了DFP校正的缺陷，BFGS是迄今为止最好的校正公式之一，具有DFP所有的性质
- 我们知道$H_{k+1}y_k = s_k$,$B_{k+1}s_k = y_k$都是拟牛顿方程($y_k = g_{k+1} - g_k$)，实际上通过交换$H_{k+1} \leftrightarrow B_{k+1}$, $y_k \leftrightarrow s_k$可以得到另一个,因此他们之间存在对偶性



- 从$H_{k+1}y_k = s_k$可得关于H_k 的DFP校正公式$H_{k+1} = H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}$,
- 从$B_{k+1}s_k = y_k$可得关于B_k的DFP校正公式$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$ (6-1a),
- 由于$B_k s_k = -\lambda_k g_k$, $B_k d_k = -g_k$,因此上式也可写为$B_{k+1} = B_k + \frac{g_k g_k^T}{g_k^T d_k} + \frac{y_k y_k^T}{\lambda y_k^T d_k}$ (6-1b)
- 式(6-1a),(6-1b)称为互补DFP公式，对此两次应用逆的秩1校正，就可得到关于H_k的BFGS校正

$$\begin{aligned}
H_{k+1}^{(BFGS)} &= H_k + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right)\frac{s_k s_k^T}{s_k^T y_k} - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k} \\
&= H_k + \frac{(s_k - H_k y_k) s_k^T + s_k (s_k - H_k y_k)^T}{s_k^T y_k} - \frac{(s_k - H_k y_k)^T y_k}{(s_k^T y_k)^2} s_k s_k^T \\
&= \left(I - \frac{s_k y_k^T}{s_k^T y_k}\right) H_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}
\end{aligned}$$

$$\begin{aligned}
B_{k+1}^{(DFP)} &= B_k + \left(1 + \frac{s_k^T B_k s_k}{y_k^T s_k}\right)\frac{y_k y_k^T}{y_k^T s_k} - \frac{y_k s_k^T B_k + B_k s_k y_k^T}{y_k^T s_k} \\
&= B_k + \frac{(y_k - B_k s_k) y_k^T + y_k (y_k - B_k s_k)^T}{y_k^T s_k} - \frac{(y_k - B_k s_k)^T s_k}{(y_k^T s_k)^2} y_k y_k^T \\
&= \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) B_k \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) + \frac{y_k y_k^T}{y_k^T s_k}.
\end{aligned}$$

- 注意DFP和BFGS校正都是对称秩2校正，都由$H_k y_k$和s_k构成，因此DFP校正和BFGS校正的加权组合也具有同样的类型，考虑如下校正族

$$H_{k+1}^{\phi}=(1-\phi)H_{k+1}^{DFP}+\phi H_{k+1}^{BFGS}\text{(6-2)},$$ 其中ϕ为参数，

- 上式(6-2)称称为Broyden校正族，显然，满足拟牛顿条件，上式也可以写为

$$\begin{aligned}H_{k+1}^{\phi} &= H_{k+1}^{DFP}+\phi v_k v_k^T\\ &= H_{k+1}^{BFGS}+(\phi-1)v_k v_k^T\\ &= H_k+\frac{s_k s_k^T}{s_k^T y_k}-\frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}+\phi v_k v_k^T,\end{aligned}$$

- 这里 $v_k=(y_k^T H_k y_k)^{\frac{1}{2}}\left[\frac{s_k}{s_k^T y_k}-\frac{H_k y_k}{y_k^T H_k y_k}\right]$

- 由参数的不同可得一系列校正方式，$\phi=0$,为DFP校正;$\phi=1$,为BFGS校正;$\phi=\frac{s_k^T y_k}{(s_k-H_k y_k)^T y_k}$,为SR1校正;$\phi=\frac{1}{1\mp\left(\frac{y_k^T H_k y_k}{(s_K^T y_k}\right)}$,为Hoshino校正

- Broyden族校正具有二次终止性，具有遗传性质和方向共轭性，且若$\phi\geq 0$时，$s_k^T y_k>0\Leftrightarrow$ *Broyden*族校正公式(6-2)保持正定。注意只要$\phi>\bar{\phi}$,则(7-2)正定，其中$\bar{\phi}=\frac{1}{1-\mu}=\frac{1}{1-\frac{y_k^T H_k y_k s_k^T B_k s_k}{(s_k^T y_k)^2}}$。且Broyden族中任何校正方法在某种程度上与参数ϕ无关，实际上它只改变搜索长度，不改变搜索方向。在精确线性搜索条件下，即使对非二次函数，所有Broyden族的校正公式都产生相同的迭代点列

- Broyden校正族中矩阵$\{H_k\}$都是对称的，且满足拟牛顿条件 $H_{k+1}y_k = s_k$
- Huang族中取消了对对称的限制，产生的矩阵$\{H_k\}$满足：$H_{k+1}y_k = \rho s_k$,ρ为参数.Huang族算法应用于正定二次函数时，产生共轭方向，具有二次终止性。所有的Huang族校正公式产生相同的迭代点序列。对于非二次函数，Huang族校正公式所生成的点列仅依赖于参数ρ。
- 推导略，这里直接给出公式
- Broyden校正族公式:$H_{k+1} = H_k + C_k A_k C_k^T, C_k = [s_k \;\; H_k y_k]_{n\times 2}, A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}_{2\times 2}$,
- 易见，如果$a_{12} = a_{21} = 0$,则为DFP公式；如果$a_{12} = a_{21}, a_{22} = 0$,则为BFGS公式，表明Broyden族是Huang族的子族
- 当极小化正定二次函数时，若取$H_0 = I$,则Huang族校正公式产生的搜索方向与Fletcher-Reeves共轭梯度法相同

◆**一阶和二阶条件：一阶考虑目标函数的导数满足的条件，二阶考虑目标函数的Hessian矩阵满足的条件**

- 必要条件
- (一阶)对无约束问题而言，假设$\boldsymbol{f}: \boldsymbol{R}^N \to \boldsymbol{R}$在$\boldsymbol{x}^*$可微,如果$\boldsymbol{x}^*$是局部最优点，则有$\nabla \boldsymbol{f}(\boldsymbol{x}^*) = \mathbf{0}$, 否则以$-\nabla \boldsymbol{f}(\boldsymbol{x}^*)$为方向必然存在下降方向
- (二阶)对无约束问题而言，假设$\boldsymbol{f}: \boldsymbol{R}^N \to \boldsymbol{R}$在$\boldsymbol{x}^*$二次可微,如果$\boldsymbol{x}^*$是局部最优点，则有$\nabla \boldsymbol{f}(\boldsymbol{x}^*) = \mathbf{0}$,且$\boldsymbol{H}(\boldsymbol{x}^*)$半正定
- 充分条件
- 对无约束问题而言，假设$\boldsymbol{f}: \boldsymbol{R}^N \to \boldsymbol{R}$在$\boldsymbol{x}^*$二次可微,如果$\nabla \boldsymbol{f}(\boldsymbol{x}^*) = \mathbf{0}$,且$\boldsymbol{H}(\boldsymbol{x}^*)$正定，则$\boldsymbol{x}^*$是严格局部最优点
- 如果目标函数$\boldsymbol{f}: \boldsymbol{R}^n \to \boldsymbol{R}$在$\boldsymbol{x}^*$是伪凸的，则$\boldsymbol{x}^*$是全局最优点当前仅当$\nabla \boldsymbol{f}(\boldsymbol{x}^*) = \mathbf{0}$

◆**一阶和二阶条件：一阶考虑目标函数的导数满足的条件，二阶考虑目标函数的Hessian矩阵满足的条件**

- 如果$f: R \to R$无穷次可微的单变量函数，则$x^* \in R$是局部最小点当前仅当要么$f^{(j)}(x^*) = 0, j = 1, 2, \cdots$,要么存在某个偶数$n \geq 2$,使得$f^{(n)}(x^*) > 0$,而其余的$f^{(j)}(x^*) = 0, 1 \leq j < n$,这里$f^{(j)}$表示第j阶导数(考虑$f\left(x^* + h\right) - f(x^*) = hf^{(1)}(x^*) + \frac{h^n}{n!} f^{(n)}(x^*) + \cdots \geq 0$)

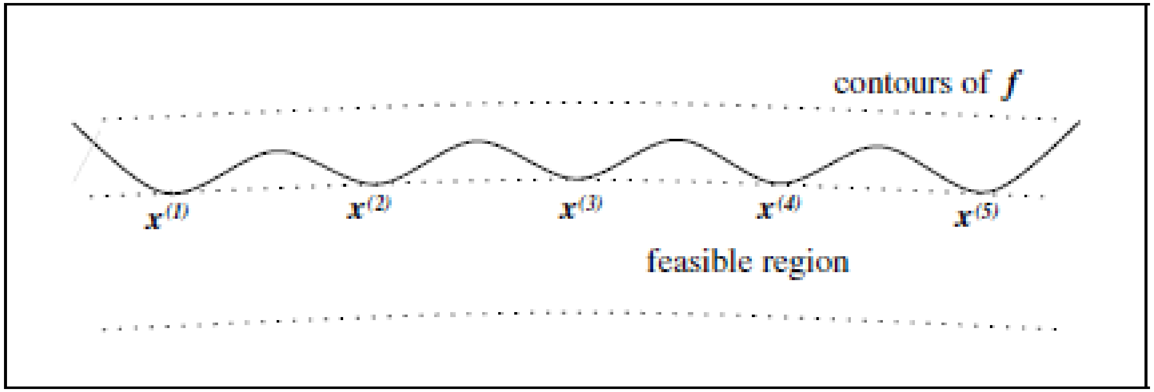
◆局部和全局解

- 思考：到底那个难？
- 如果添加约束呢？为什么？

◆约束可能使事情变得更为困难！

$$min(x_2+100)^2+0.01x_1^2, s.t. x_2-cosx_1\geq 0$$

- 没有约束的时候，显然唯一解:$(0,-100)^T$
- 有约束的时候，靠近$x^{(k)}=(k\pi,-1)^T, k=\pm 1,\pm 3,\pm 5,\cdots$ 都有局部极小点



◆光滑性(smoothness)

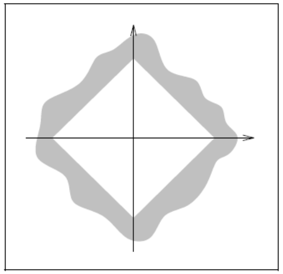
- 目标函数和约束的光滑性在刻画解的时候非常重要，为什么？
- 优化问题的可行域的光滑性与约束函数的光滑性有什么关系？(非光滑边界可由光滑函数集合来描述)
- 如图，可由单个非光滑约束描述，也可由光滑约束来描述

$$(a)\ \|x\|_1 = |x_1| + |x_2| \leq 1$$

$$(b)\ x_1 + x_2 \leq 1, x_1 - x_2 \leq 1, -x_1 + x_2 \leq 1, -x_1 - x_2 \leq 1$$

- 一般选择第二种方式来作为约束,并且将非光滑，无约束问题形式化为具有光滑约束的问题

$$f(x) = max(x^2, x)$$



- $min\ t, s.t.\ t \geq x, t \geq x^2$

◆注意：这种目标函数是$\|\cdot\|_1, \|\cdot\|_\infty$,以及函数集合的最大值的形式，一般都这么做

◆约束非线性优化问题P

$$Min\ f(x), s.t.\begin{cases} g_i(x) \geq 0, i = 1,2,\cdots,m \\ h_i(x) = 0, i = 1,2,\cdots,l \\ x \in X \end{cases} \quad (6\text{-}3)$$

◆如果约束g_i, h_i都是线性的，则称为线性约束优化问题，如果此时目标函数f是二次函数，则称为二次规划问题

◆与以前的线性规划类似，也有可行域等概念

◆如假设x^*是上述问题的一个局部极小点，如果存在$i_0 \in [1, m]$,使得$g_{i_0}(x^*) > 0$,则这时候可以将第i_0个不等式约束去掉，且x^*仍然是去掉这个约束后的问题的局部极小点，称第i_0个约束在x^*处是非积极的.

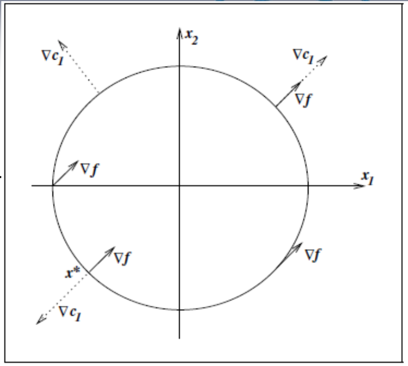
◆积极约束(Active Constraint),紧约束(tight),捆绑约束(binding)，起作用约束

◆定义$E=\{1,2,\cdots,l\}$,$I=\{1,2,\cdots,m\}$,$I(x)=\{i|g_i(x)=0,i\in[1,m]\}$, 对任意$x\in R^n$,集合$A(x)=E\cup I(x)$称为积极约束，$g_i(x)(i\notin A(x))$称为在$x$点的非积极约束，有时约束也用$c_i(x)$表示

◆若已知问题P在解处的积极约束$A(x^*)$,则只需求解如下的等式约束优化问题

◆$Min\ f(x), s.t. \begin{cases} h_i(x)=0, i\in A(x^*) \end{cases}$,而这个问题一般比原问题更容易求解

第6章约束最优化理论(Theory of Constrained Optimization)-例



- 单个等式约束问题

$$min\ x_1 + x_2, s.t.\ x_1^2 + x_2^2 - 2 = 0$$

- 显然$f(x) = x_1 + x_2, I = \emptyset, E = \{1\}, h_1(x) = x_1^2 + x_2^2 - 2$
- 从任意点沿可行域移动，可得解
- x^*处，约束在该点的梯度与目标函数在该点的梯度平行，即存在标量$\lambda_1^*(= -\frac{1}{2})$,使得

$$\nabla f(x^*) = \lambda_1^* \nabla h_1(x^*)$$

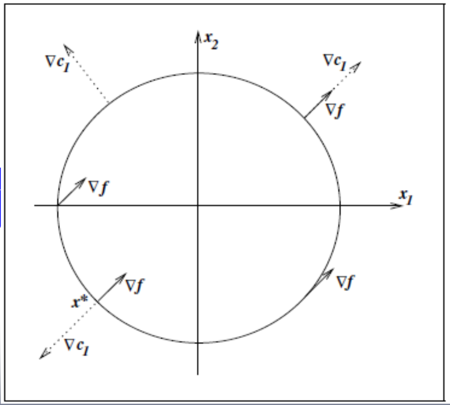
- 检查对目标函数和约束函数的一阶泰勒逼近，一样也可以获得类似的结果

- 为保持$h_1(x)=0$的可行性，任意移动步p都需要满足可行性条件

$$0=h_1(x+p)\approx h_1(x)+\nabla h_1(x)^T p=\nabla h_1(x)^T p$$

- 因此,要求$\nabla h_1(x)^T p=0$,与以前类似，要求目标函数下降:$0>f(x+p)-f(x)\approx \nabla f(x)^T p$
  - 若存在这样的迭代步，表明存在方向d,使得$\nabla h_1(x)^T d=0,\nabla f(x)^T \mathrm{d}<0$
  - 若不存在这样的迭代步，表明x^*就是一个局部极小点
- 若不平行，则
- $p=-(I-\frac{\{\nabla \mathrm{h}_1(\mathrm{x})\nabla \mathrm{h}_1(\mathrm{x})^{\mathrm{T}}\}}{\{||\nabla \mathrm{h}_1(\mathrm{x})||^2\}})\nabla f(x), d=\frac{p}{||p||}$,满

降方向



◆引入Lagrangian(拉格朗日)函数

$$L(x,\lambda_1)=f(x)-\lambda_1 h_1(x)$$

◆令$\nabla_x L(x,\lambda_1)=\nabla f(x)-\lambda_1\nabla h_1(x)$,则前述$\nabla f(x^*)=\lambda_1^*\nabla h_1(x^*)$条件等价于$\nabla_x L(x,\lambda_1)=0$

◆表明：求解等式约束问题的最优解等价于寻找拉格朗日函数的稳定点，其中的标量λ_1称为拉格朗日乘子

◆表明这种条件是必要条件，而不是充分条件，例如$x=(1,1)^T,\lambda_1=\frac{1}{2}$也满足条件，但不是问题的解

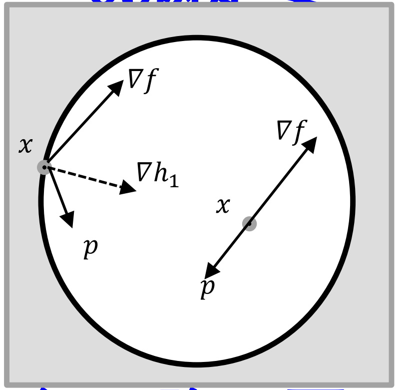
◆而且，等式约束问题中，上述$\nabla f(x^*)=\lambda_1^*\nabla h_1(x^*)$条件不能通过改变乘子的符号而变成充分条件

- 单个不等式约束

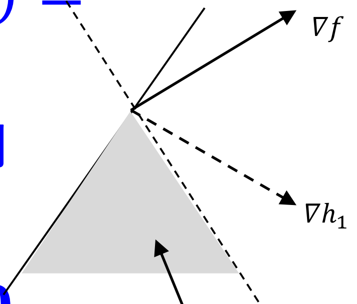
$$min\ x_1 + x_2, s.t. 2 - x_1^2 - x_2^2 \geq 0$$

- 解仍然不变，乘子也不变，但与等式约束不同的是这里乘子的符号扮演了重要的作用
- 与之前的猜想类似，如果x^*不是最优点，意味着存在可行下降方向
- 因此如果存在迭代步p,函数值下降，则 (**a**)$\nabla f(x)^T p < 0$且可行(**b**)$0 \leq h_1(x+p) \approx h_1(x) + \nabla h_1(x)^T p$
- 确定是否迭代步p满足上述两个条件，可考虑两种情况
    - x严格在圆的内部
    - x正在圆上

- **1:在圆内**，则表明$h_1(x) > 0$成立,此时任意p都满足条件(b),只要步长足够小就可以。实际上，当$\nabla f(x) \neq 0$,设定$p = -\alpha\nabla f(x),\alpha > 0$即可以满足条件(a)(b)。如果$\nabla f(x) = 0$,满足要求性质的迭代步没有要求



- **2:在圆上**，此时约束为起作用约束
- 此时条件为$\nabla f(x)^T p < 0, \nabla h_1(x)^T p \geq 0$
- 第一个定义了开半空间，第二个定义了闭半空间，从右图可以看出，两个空间交中的任意方向都满足可行下降，要想其为空集，必须$\nabla f(x), \nabla h_1(x)$为同一方向，$\nabla f(x) = \lambda_1 \nabla h_1(x), \lambda_1 \geq 0$,

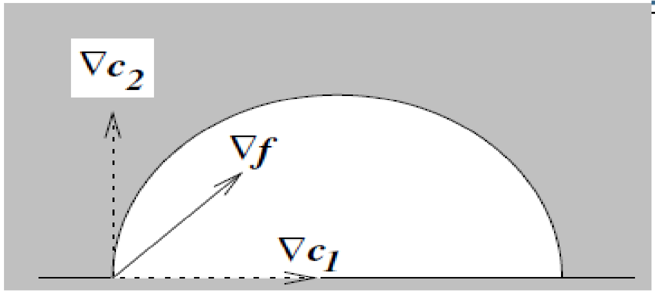


- 注意符号的正负性。原来约束为等式的时候，符号的正负性没有限制
- 这两种情况的最优性条件可用拉格朗日函数表达为：

$$\nabla_x L(x^*, \lambda_1^*) = 0, \lambda_1^* \geq 0, (c): \lambda_1^* h_1(x^*) = 0$$

- 条件(c)称为互补性条件，意味着拉格朗日乘子当约束为起作用约束时，对应乘子严格为正

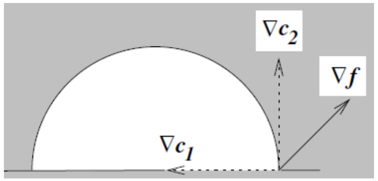
第6章约束最优化理论(Theory of Constrained Optimization)-例



- 两个不等式约束
- $min\ x_1 + x_2, s.t. 2 - x_1^2 - x_2^2 \geq 0, x_2 \geq 0$
- 可行域如图，$x^* = \left(-\sqrt{2}, 0\right)^T$同时满足两个约束
- 同前述一样，满足一阶下降条件要求

$$(*) h_i(x)^T p \geq 0, i \in I = \{1,2\}, \nabla f(x)^T p < 0$$

- 从图中可见x^*不存在这样的p
- $L(x,\lambda) = f(x) - \lambda_1 h_1(x) - \lambda_2 h_2(x)$,$\lambda = (\lambda_1, \lambda_2)^T$,此时扩展前述结论
  - $L_x = 0, \lambda^* \geq 0$
  - 应用上页的互补性条件(c)得$\lambda_1^* h_1(x^*) = 0, \lambda_2^* h_2(x^*) = 0$
- 代入得 $\nabla f(x^*) = (1,1)^T, \nabla h_1(x^*) = \left(2\sqrt{2}, 0\right)^T, \nabla h_2(x^*) = (0,1)^T$



- 易验证若$\lambda^* = \begin{pmatrix} \frac{1}{2\sqrt{2}} \\ 1 \end{pmatrix}$,$\nabla_x L(x^*, \lambda^*) = 0$
- 若看看别的不是问题解的可行点，是否也满足拉格朗日和其梯度方程
- 点$x = \left(\sqrt{2}, 0\right)^T$,显然这点也是起作用约束点,但$p = (-1,0)^T$就满足$(*)$式，此时拉格朗日方程梯度为0得:$\lambda = \left(-\frac{1}{2\sqrt{2}}, 1\right)^T$,但$\lambda_1 < 0$,此时互补性条件(c)不满足

◆两个不等式约束

$$\min x_1 + x_2, s.t. 2 - x_1^2 - x_2^2 \geq 0, x_2 \geq 0$$

◆**练习:**验证点$x = (1,0)^T$是否满足问题的拉格朗日函数及其梯度和约束所构成的条件

$$\nabla_x L(x, \lambda) = 0, \text{某个} \lambda \geq 0$$
$$\lambda^T h(x) = 0$$

◆寻找p满足$(*) h_i(x)^T p \geq 0, i \in I = \{2\}, \nabla f(x)^T p < 0$

◆例如方向$p = \left(-\frac{1}{2}, \frac{1}{4}\right)^T$

- ◆**有约束问题求解也是从下降方向开始，在可行域中前进**
  - ➢并且从几何上看，给出了最优点的一些直观的印象
  - ➢可行方向集合不包含使得目标函数值下降的方向
- ◆**如何从直觉转化为可计算的形式呢？**
  - ➢上述直觉用切锥来精确描述（定义见后），但不容易计算
  - ➢为了易于计算，线性化可行方向
    - ✓等式约束：满足梯度与可行方向垂直
    - ✓起作用约束：满足可行方向
  - ➢但这与切锥不一样，切锥包含于线性化可行方向集合
  - ➢因此，引入一些约束，使得这两个一样，最基本的称为线性无关约束规范！要求起作用约束的梯度是线性无关的！

- 显然，问题P的局部极小点x^*如果直接由定义来判别，则需要判别无穷多个点，有必要给出只依赖x^*处的目标函数和约束函数信息，且与定义等价的条件，这些条件称为最优性条件(Optimality Conditions)
- 约束规范(Constraint Qualifications): 前面根据目标函数$f(x)$和约束函数$h(x), g(x)$的梯度来判断是否从给定的可行点x可以找到可行下降方向.实际上采用一阶泰勒级数来形成一个逼近问题，其中目标和约束函数都是线性的，但这种方法只有当线性逼近捕获了x附近可行集的本质几何特征的时候，方法才有意义。因此必须对x附近起作用约束的本质做一些假设，来确保这种线性化处理在x附近与可行集相似
- 可行方向(Feasible Direction):设$x^* \in X, 0 \neq d \in R^n$,如果存在$\delta > 0$,使得$x^* + td \in X, \forall t \in [0, \delta]$,则称$d$是$X$在$x^*$处的可行方向，所有可行方向的集合定义为$FD(x^*, X)$,这是一个锥，称为$X$在$x^*$处的可行方向锥
- $x^* \in X, d \in R^n$,若$d^T \nabla h_i(x^*) = 0, d^T \nabla g_i(x^*) \geq 0, i \in I$,则称$d$是$X$在$x^*$处的线性化可行方向(Linearized feasible direction)。所有在x^*处的线性化可行方向记为$LFD(x^*, X)$(注意这里$\geq$是针对$g \geq 0$的约束，如果约束为≤ 0,则这里也是$\leq$)

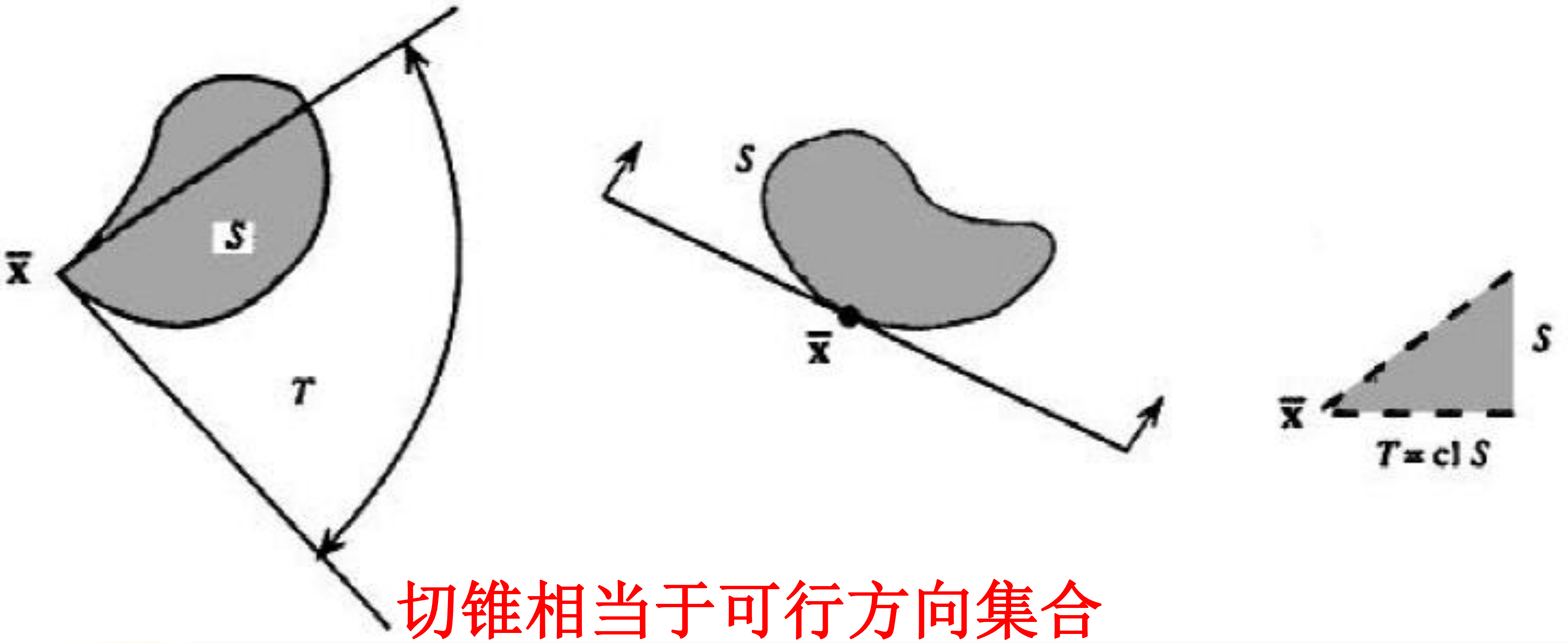
第6章约束最优化理论(Theory of Constrained Optimization)

- 序列可行方向SFD(Sequential feasible direction):存在$d_k(k=1,2,\cdots)$和$\delta_k(k=1,2,\cdots)$,使得$x^*+\delta_k d_k\in X,\forall k,d_k\to d,\delta_k\to 0$,则极限方向$d$称为SFD，所有SFD的集合称为$SFD(x^*,X)$, 若SFD还包含0向量，则称$T_X(x^*)=SFD(x^*,X)\cup\{0\}$为$X$在$x^*$点的切锥(Tangent cone)
- 引理：令$x^*\in X$,若所有的约束函数在x^*处都可微，则$FD(x^*,X)\subseteq SFD(x^*,X)\subseteq LFD(x^*,X)$
  - 证明：只需证明$\forall d\in FD(x^*,X)\Rightarrow d\in SFD\ (x^*,X)\Rightarrow d\in LFD\ (x^*,X)$
- 引理：设$x^*\in X$是问题P的局部极小点，如果目标函数和约束函数在x^*点都可微，则必有$d^T\nabla f(x^*)\ge 0,\forall d\in SFD(x^*,X)$
  - 证明:由定义，$\forall d\in SFD(x^*,X),\exists\delta_k>0,d_k,x^*+\delta_k d_k\in X$,且$\delta_k\to 0,d_k\to d$,因而对充分大的k,$f(x^*)\le f(x^*+\delta_k d_k)=f(x^*)+\delta_k d_k^T\nabla f(x^*)+O(\delta_k)$,从而立即可得$d_k^T\nabla f(x^*)\ge 0$

◆切锥

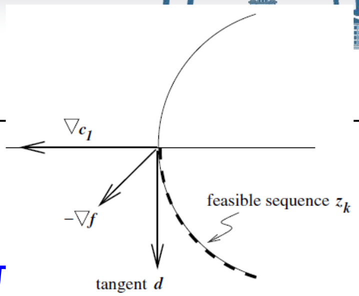
➢是一个集合，包含所有方向$\boldsymbol{d}$,这个方向定义为：有一个收敛到$\boldsymbol{x}$的可行序列$\{\boldsymbol{x}_k\}$,使得$\boldsymbol{x}_k - \boldsymbol{x}$收敛到$\boldsymbol{d}$，则所有这种方向构成了在点$\boldsymbol{x}$处的切锥，即$\boldsymbol{lim}_{k\to\infty} \frac{\boldsymbol{x}_k - \boldsymbol{x}}{\boldsymbol{t}_k} = \boldsymbol{d}$,

正标量序列$\boldsymbol{t}_k \to \boldsymbol{0}$



切锥相当于可行方向集合

- 序列可行方向SFD(Sequential feasible direction):存在$d_k(k=1,2,\cdots)$和$\delta_k(k=1,2,\cdots)$,使得$x^*+\delta_k d_k\in X,\forall k, d_k\to d,\delta_k\to 0$,则极限方向$d$称为SFD，所有SFD的集合称为$SFD(x^*,X)$, 若SFD还包含0向量，则称$T_X(x^*)=SFD(x^*,X)\cup\{0\}$为$X$在$x^*$点的切锥(Tangent cone)
- 引理：令$x^*\in X$,若所有的约束函数在x^*处都可微，则$FD(x^*,X)\subseteq SFD(x^*,X)\subseteq LFD(x^*,X)$
  - 证明：只需证明$\forall d\in FD(x^*,X)\Rightarrow d\in SFD(x^*,X)\Rightarrow d\in LFD(x^*,X)$
- 引理：设$x^*\in X$是问题P的局部极小点，如果目标函数和约束函数在x^*点都可微，则必有$d^T\nabla f(x^*)\geq 0,\forall d\in SFD(x^*,X)$
  - 证明:由定义，$\forall d\in SFD(x^*,X),\exists\delta_k>0,d_k,x^*+\delta_k d_k\in X$, 且$\delta_k\to 0,d_k\to d$, 因而对充分大的$k,f(x^*)\leq f(x^*+\delta_k d_k)=f(x^*)+\delta_k d_k^T\nabla f(x^*)+O(\delta_k)$,从而立即可得$d_k^T\nabla f(x^*)\geq 0$



- $\min x_1+x_2, s.t. x_1^2+x_2^2-2=0$
- 图中显示的序列$\{z_k\}$即为逼近$x=\left(-\sqrt{2},0\right)^T$的可行序列$z_k=(-\sqrt{2-\frac{1}{k^2}},-\frac{1}{k})$，显然$\{z_k\}$也不是优化问题的解，从另外一个方向，也可以构造一个可行序列，在x的切锥为$\{(0,d_2)^T|d_2\in R\}$
- 实际上，可行方向要求$0=\nabla h_1(x)^T d\Rightarrow 0=-2\sqrt{2}d_1$,从而$LFD(x,X)=\{(0,d_2)^T,d_2\in R\}$与其切锥一样
- 如$h_1(x)=\left(x_1^2+x_2^2-2\right)^2$,方向$d\in LFD(x,X)$,则可知$LFD(x,X)=R^2$
- 分析问题$\min x_1+x_2, s.t. x_1^2+x_2^2-2\le 0$可行域上的点$x=\left(-\sqrt{2},0\right)^T$,则存在无穷多个可行序列收敛到该点，从圆内任意一点可与该点连成直线，直线上的点形成序列$z_k=\left(-\sqrt{2},0\right)^T+\frac{1}{k}w$,$w=(w_1,w_2)^T,w_1>0$,此时，只要$||z_k||\le\sqrt{2}$,则序列$\{z_k\}$就是可行的。可得$k\ge(w_1^2+w_2^2)/(2\sqrt{2}w_1)$,此外，还有无穷多个圆内的点列满足该性质，因此该点处的切锥为$\{(w_1,w_2)^T|w_1\ge 0\}$
- 同样根据可行集的定义有$0\le\nabla h_1(x)^T d=2\sqrt{2}d_1$,可得$d_1\ge 0$,这与该点的切锥一样

◆约束规范(Constraint Qualifications)

- ➢**LFD**与切锥相似的条件
- ➢大多数约束规范确保这两个集合相等

◆前面的例子说明

- ➢通过线性化代数约束条件构造的**LFD**能够捕获可行集在x的邻域内的几何特征(由该点的切锥来表示)
- ➢如果所有的约束函数都是线性的，则显然**LFD**与切锥一样

◆根据之前的论述，最优点x^*满足该点无下降方向！用线性可行方向来表达则为：

$$\left\{ d \,\middle|\, \begin{array}{c} d^T\nabla f(x^*) < 0 \\ d^T\nabla c_i(x^*) = 0, i \in E \\ d^T\nabla c_i(x^*) \le 0, i \in A(x^*) \cap I \end{array} \right\} \text{是空集}$$

$$\left\{ d \middle| \begin{cases} d^T\nabla f(x^*) < 0 \\ d^T\nabla c_i(x^*) = 0, i \in E \\ d^T\nabla c_i(x^*) \le 0, i \in A(x^*) \cap I \end{cases} \right\}$$

- **Farkas引理(择一性引理)** 设l, l'是两个非负整数，$a_0, a_i(i = 1, 2, \cdots, l)$和$b_i(i = 1, 2, \cdots, l') \in R^n$,则线性方程组和不等式组：

$$\begin{cases} d^T a_i = 0, i = 1, 2, \cdots, l & (6-4) \\ d^T b_i \le 0, i = 1, 2, \cdots, l' & (6-5) \\ d^T a_0 < 0 & (6-6) \end{cases}$$ 无解，

当且仅当存在实数$\lambda_i(i = 1, 2, \cdots, l)$和非负实数$\mu_i(i = 1, 2, \cdots, l')$使得$a_0 + \sum_{i=1}^{l} \lambda_i a_i + \sum_{i=1}^{l'} \mu_i b_i = 0$ (6-7)

- 证明：假设(6-7)成立，则存在非负实数$\mu_i \ge 0(i = 1, 2, \cdots, l')$，且$a_0 = -\sum_{i=1}^{l} \lambda_i a_i - \sum_{i=1}^{l'} \mu_i b_i$ 从而根据(6-4)(6-5)可知$d^T a_0 = d^T\left(-\sum_{i=1}^{l} \lambda_i a_i - \sum_{i=1}^{l'} \mu_i b_i\right)$= -$\sum_{i=1}^{l'} \mu_i d^T b_i \ge 0$,因此(6-6)无解
- 假设不存在实数$\lambda_i(i = 1, \cdots, l)$和非负实数$\mu_i(i = 1, 2, \cdots, l')$使得(6-7)成立。定义集合$S = \{a | a = -\sum_{i=1}^{l} \lambda_i a_i - \sum_{i=1}^{l'} \mu_i b_i, \lambda_i \in R, \mu_i \ge 0\}$,显然$S$是$R^n$中的闭凸锥，由于$a_0 \notin S$,因此必存在$d \in R^n$,使得$d^T a_0 < \alpha < d^T a, \forall a \in S$,其中$\alpha$是某一常数。由于$0 \in S$,因此$d^T a_0 < 0$.对$\forall \lambda > 0$, 都有$-\lambda b_i \in S$,因此$-\lambda d^T b_i > \alpha, \forall \lambda > 0$,从而可推出$d^T b_i \le 0, i = 1, 2, \cdots, l'$, 同理对$\forall \lambda > 0, \lambda a_i \in S, (-\lambda) a_i \in S$, 都有$d^T a_i \le 0, d^T(-a_i) \le 0, i = 1, 2, \cdots, l$, 因此$d^T a_i = 0, i = 1, 2, \cdots, l$ 。证毕!

式(6-7)称为互补条件，他要求$\mu_i^*, g_i(x^*)$必须有一个为0

- **(一阶优化条件Kuhn-Tucker条件)定理:设x^*是问题P的一个局部极小点，如果$SFD(x^*, X) = LFD(x^*, X)$，则必存在$\lambda_i^*, \mu_i^* (i = 1, 2, \cdots, m + l)$, 使得$\nabla f(x^*) + \sum_{i=1}^{l} \lambda_i^* \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) = 0$ (6 − 8); $\mu_i^* \geq 0, \mu_i^* g_i(x^*) = 0$ (6-9)**
- **证明：由SFD=LFD和Farkas引理，可知如下线性系统无解**

$$\begin{cases} d^T \nabla h_i(x^*) = 0, i = 1, 2, \cdots, l \\ d^T \nabla g_i(x^*) \leq 0, i = 1, 2, \cdots, m, i \in I \\ d^T \nabla f(x^*) < 0 \end{cases}$$

,利用Farkas引理，可知存在实数$\lambda_i (i = 1, 2, \cdots, l)$和非负实数$\mu_i (i = 1, 2, \cdots, l')$使得$\nabla f(x^*) + \sum_{i=1}^{l} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0$ 证毕

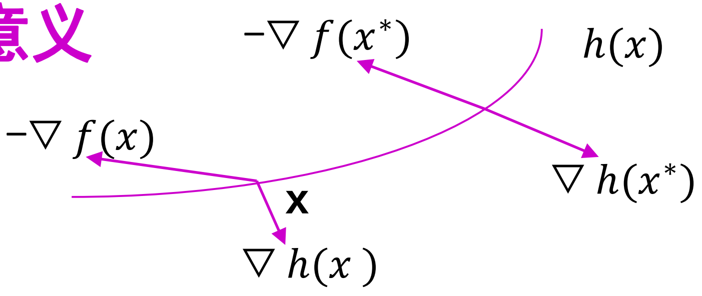
- **函数$L(x, \lambda, \mu) = f(x) - \lambda^T h(x) + \mu^T g(x)$称为拉格朗日函数，其系数$\lambda, \mu$称为拉格朗日乘子**
- **如果$x^* \in X$,且存在λ, μ使得(6-8)(6-9)式成立，则称点x^*为问题P的Kuhn-Tucker点，简称KT点,有时也称为KKT点**

约束规范条件(Constraint Qualifications)：条件成立时，问题P的局部极小点才是KKT点，否则可能不一定是KKT点

- 例:$min\ x_1, s.t. x_1^3 - x_2 \geq 0, x_2 \geq 0$
- 显然，$x^* = (0,0)^T$是问题的全局极小点，在x^*处，我们可以得到$SFD(x^*,X) = \{d|d = \binom{\alpha}{0}, \alpha \geq 0\}$,且$LFD(x^*,X) = \{d|d = \binom{\alpha}{0}, \alpha \in R^1\}$,因此上述一阶最优性条件定理中的约束规范条件不成立。直接计算可知$\nabla f(x^*) = (1\ 0)^T, \nabla g_1(x^*) = (0\ \ -1)^T, \nabla g_2(x^*) = (0\ 1)^T$,因此不可能存在$\mu_1, \mu_2$使得$\nabla f(x^*) = \mu_1 \nabla g_1(x^*) + \mu_2 \nabla g_x(x^*)$
- 但实际计算中，约束规范条件 $SFD(x^*,X) = LFD(x^*,X)$不容易直接验证，因此采用更强的，但容易验证的约束规范条件
  - 如约束条件都是线性函数
  - $\nabla h_i(x^*)$线性无关，且$S^* = \{d|d^T h(x^*) = 0; d^T g(x^*) > 0\}$非空
  - $\nabla h_i(x^*)$，$\nabla g_i(x^*)$线性无关 ⟵ 最常用的约束规范条件，且易于检验
- 上面的都是必要条件
- 实际上，针对等式约束和不等式约束，分别有上述定理的表达形式
- 等式约束：若x^*是P的只带等式约束的局部最优点 ,则存在$\lambda^* \in R^l$使 $\nabla f(x^*) + {\lambda^*}^T h(x^*) = 0$。表明目标函数在最优点的梯度值是约束函数在该点梯度值的组合

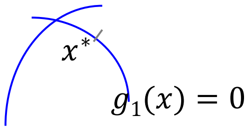
第6章约束最优化理论(Theory of Constrained Optimization)

- 等式约束的几何意义,另外可分析拉格朗日函数对右端项扰动的几何意义



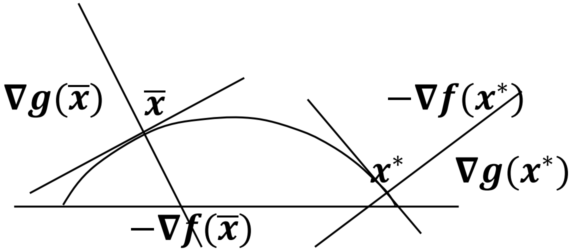
这里 $\boldsymbol{x}^*$为局部最优点. $\nabla f(x^*)$与$\nabla h(\boldsymbol{x}^*)$ 共线
，而x**非局部最优点**$\nabla f(x)$与$\nabla h(x)$不共线,
说明存在某个新方向可以使得函数值进一步下降。

- 不等式约束：此时只考虑积极约束



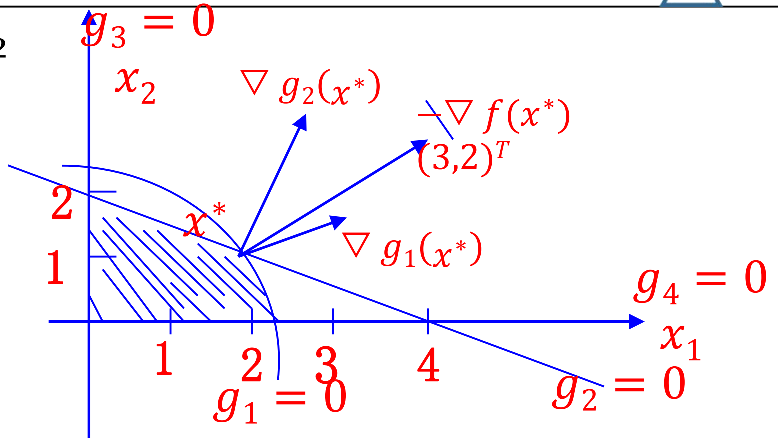
$g_1(x^*) = 0, g_1$为起作用约束

- 此时考虑x^*是否是最优点时，只有那些使$g_i(x)$变为大于0的方向上有限制，而对其它变化方向上没有限制,如下图所示。



- 几何意义类似：目标函数的负梯度可表示成紧约束(或起作用约束)函数梯度的非负组合，或者说负梯度属于紧约束函数梯度所张成的锥.

◆ **例**
$$\begin{cases}\min\ f(x_1,x_2)=(x_1-3)^2+(x_2-2)^2\\ s.t.\, g_1(x_1,x_2)=x_1^2+x_2^2-5\le 0\\ g_2(x_1,x_2)=x_1+2x_2-4\le 0\\ g_3(x_1,x_2)=-x_1\le 0\\ g_4(x_1,x_2)=-x_2\le 0\end{cases}$$



◆ $x^*=(2,1)^T, I=\{1,2\}$(积极约束),沿函数下降方向，会移出约束集，因此x^*是局部极小点，这里可以看到$\nabla f(x^*)=(-2\ \ -2)^T,\nabla g_1(x^*)=(4,2)^2,\nabla g_2(x^*)=(1,2)^T$, 且有

$$\nabla f(x^*)+\frac{1}{3}\nabla g_1(x^*)+\frac{2}{3}\nabla g_2(x^*)=0$$

◆ 不等式约束:若x^*是P的只带不等式约束的局部最优点 ,I为紧约束集，在$x^*,f,g_i,i\in I$可微，$g_i,i\notin I$连续，$\{\nabla g_i(x^*)|i\in I\}$线性无关。则存$\mu^*\in R^{|I|}$使$\nabla f(x^*)+\mu^{*^T}\nabla g_I(x^*)=0$。表明目标函数在最优点的梯度值是起作用约束函数在该点梯度值的组合。如果$g_i,i\notin I$在x^*点也可微,则下列关系式成立

- $\nabla f(x^*)+\frac{\partial g(x^*)}{\partial x}\mu^*=0$
- $\mu^*\ge 0$
- $\mu^{*^T}g(x^*)=0$：互补松弛性条件

约束规范

包含等式和不等式约束的最优性条件见前述，实际可以可以通过上述的分解再合成来描述

$$\begin{cases}\min f(x_1,x_2)=(x_1-3)^2+(x_2-2)^2 \\ s.t.\ g_1(x_1,x_2)=x_1^2+x_2^2-5\le 0 \\ g_2(x_1,x_2)=x_1+2x_2-4\le 0 \\ g_3(x_1,x_2)=-x_1\le 0 \\ g_4(x_1,x_2)=-x_2\le 0\end{cases}$$

- **例子，可能的K-T点出现在下列情况：**

①两约束曲线的交点：g1与g2,g1与g3,g1与g4,g2与g3,g2与g4,g3与g4。

②目标函数与一条曲线相交的情况： g1，g2, g3, g4

对每一种情况求得满足以下条件的点$(x_1,x_2)^T$及乘子μ_i,验证当满足可得，且$\mu_i>0$时，即为一个K-T点。

$$\begin{cases}2(x_1-3)+\mu_1 2x_1+\mu_2+\mu_3(-1)+\mu_4\cdot 0=0 \\ 2(x_2-2)+\mu_1 2x_2+\mu_2 2+\mu_3\cdot 0+\mu_4\cdot(-1)=0 \\ \mu_i\ge 0 \\ \mu_1(x_1^2+x_2^2-5)=0 \\ \mu_2(x_1+2x_2-4)=0 \\ \mu_3 x_1=0 \\ \mu_4 x_2=0\end{cases}$$

g_1与g_3交点：$\begin{cases}x_1^2+x_2^2-5=0 \\ x_1=0\end{cases}$ 得$x=(0,\pm\sqrt{5})^T$

$(0,-\sqrt{5})^T\notin S$,故不是$K-T$点；

$(0,\sqrt{5})^T\notin S$,不满足$g_2\le 0$,故不是$K-T$点。

g_3,g_4交点：$x=(0,0)^T\in S$ $I=\{3,4\}$故$u_1=u_2=0$

解$\begin{cases}2(0-3)-u_3=0 \\ 2(0-2)-u_4=0\end{cases}$得$u_3=-6<0,u_4=-4<0$

故非$K-T$点.

目标函数$f(x)$与$g_1(x)=0$相切的情况：

$I=\{1\}$,则$u_2=u_3=u_4=0$

解$\begin{cases}2(x_1-3)+2x_1u_1=0 \\ 2(x_2-2)+2x_2u_1=0 \\ x_1^2+x_2^2-5=0\end{cases}$ 得$(\pm\sqrt{\frac{45}{13}},\pm\sqrt{\frac{20}{13}})\notin S$

故均不是$K-T$点 $g_2(x_1,x_2)=\sqrt{\frac{45}{13}}+2\sqrt{\frac{20}{13}}-4=7\sqrt{\frac{5}{13}}-4=0.34>0$

- 继续考虑上述的例子，验证$x^* = (2,1)^T$是KT点
- $\nabla f(x) = \begin{pmatrix} 2(x_1-3) \\ 2(x_2-2) \end{pmatrix}$,$\nabla g_1(x) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$, $\nabla g_2(x) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\nabla g_3(x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, $\nabla g_4(x) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$，从而得到$KT$条件

式:$\begin{cases} 2(x_1-3)+\mu_1 2x_1+\mu_2+\mu_3(-1)+\mu_4\cdot 0=0 \\ 2(x_2-2)+\mu_1 2x_2+\mu_2 2+\mu_3\cdot 0+\mu_4\cdot(-1)=0 \\ \mu_i \ge 0 \\ \mu_1(x_1^2+x_2^2-5)=0 \\ \mu_2(x_1+2x_2-4)=0 \\ \mu_3 x_1=0 \\ \mu_4 x_2=0 \end{cases}$ $\Rightarrow I=\{1,2\}, \mu_3=\mu_4=0 \Rightarrow$

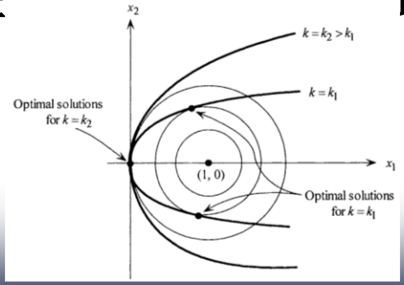
$\begin{cases} 4\mu_1+\mu_2=2 \\ 2\mu_1+2\mu_2=2 \end{cases}$,从而得$\mu_1^* = \frac{1}{3}, \mu_2^* = \frac{2}{3}$,均非负，因而$x^* = (2,1)^T$是KT点

- 注解:K-T条件的地位相当于无约束最优化问题中的驻点条件。因此，许多约束最优化算法的目标设定为求得K-T点
- 前面讨论的均为充分条件，但如果问题P是凸规划问题，则上述讨论同时也是必要条件
- 定理：设$f: R^n \to R$为可微凸函数，$g: R^n \to R^m$为可微凸函数组，$h(x) = Ax - b, A$为$l \times n$矩阵,$b \in R^l$. $x^* \in S = \{x | g(x) \le 0, h(x) = 0\}$,且满足$\nabla f(x^*) + h(x^*) + \mu_I^T g_I(x^*) = 0$, 则$x^*$为问题P的全局最优解。这里$I$表示起作用约束集
- 证明：根据凸函数的性质，$\forall x \in S, f(x) \ge f(x^*) + \nabla f^T(x^*)(x - x^*), g(x) \ge g(x^*) + \nabla g^T(x^*)(x - x^*) \Rightarrow \nabla g^T(x^*)(x - x^*) \le 0$,且有$h(x) - h(x^*) = A(x - x^*) = 0$,将$\nabla f(x^*) + h(x^*) + \mu_I^T g_I(x^*) = 0$两边同乘$(x - x^*) \Rightarrow \nabla f^T(x^*)(x - x^*) \ge 0$,得证

- 线性规划的KT条件点，必然是最优解
- 考虑$max\ z = c^T x, s.t. Ax = b, x \geq 0 \Rightarrow min\ f = -c^T x, s.t. -x \leq 0, Ax - b = 0$
- 其KT条件式为:$\begin{cases} -c + (-E)\mu + A^T\lambda = 0 \\ \mu^T x = 0 \\ \mu \geq 0 \end{cases}$,$E$为单位矩阵，$\mu \in R^n, \lambda \in R^m$,可得

$\mu = -c + A^T\lambda \geq 0, \left(c - A^T\lambda\right)^T x = 0 \Rightarrow c^T x = (Ax)^T\lambda = b^T\lambda$,若$x = (x_B\ x_N)^T, x_B > 0, x_N = 0$,且$c, \mu$均这样分解，有互补松弛性条件$\mu_B^T x_B = 0 \Rightarrow \mu_B = -c_B + B^T\lambda = 0, \lambda^T = c_B^T B^{-1}$(即对偶问题的解)，代入$\mu_N^T = -c_N^T + \lambda^T N = -c_N^T + c_B^T B^{-1} N \geq 0$,于是$-\mu_N^T = c_N^T - c_B^T B^{-1} N \leq 0$,这就是线性规划的最优性条件,$-\mu_N$即为检验数

- 约束最优化问题的二阶条件
  - 无约束问题的二阶必要条件和充分条件是要求二阶Hessian矩阵半正定和正定
  - 对于约束问题只考虑其Hessian矩阵是不够的，即使$H(f) = \nabla^2 f(x^*)$正定，但不能保证x^*是最优点
  - 例如非凸规划问题$P: min\ \{(x_1 - 1)^2 + x_2^2 : g_1(x) = 2kx_1 - x_2^2 \leq 0\}$



- 下面讨论充分条件
- 设$x^* \in X$，如果目标函数和约束函数在x^*点都可微，且$d^T\nabla f(x^*) > 0, \forall 0 \neq d \in SFD(x^*, X)$,则$x^*$是问题P的严格局部极小点.注意这里$LFD \leftarrow SFD$也成立。
- 定理：设$f(x), h(x), g(x)$在可行域的某一开集上连续可微，如果x^*是问题P(6-1)的局部极小点，则必存在不全为0的$\lambda_0^* \geq 0, \lambda \in R^l, \mu \in R^m$,使得$\lambda_0^*\nabla f(x^*) - \lambda^T h(x^*) + \mu^T g(x^*) = 0, \mu \geq 0, \mu^T g(x^*) = 0$
- 称满足上述条件的点为Fritz John点，称下面的加权Lagrange函数为Fritz John函数$L(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x) - \lambda^T h(x) + \mu^T g(x)$
- 显然Fritz John点是Fritz John函数的稳定点，注意到$\lambda_0 \geq 0$,如果$\lambda_0 > 0$,则Fritz John函数可以看成是Lagrange函数的λ_0倍，如果$\lambda_0 = 0$，则Fritz John函数与目标函数无关，仅仅描述了约束函数，这就是Fritz John点没受到重视的原因

◆给定非线性规划问题，存在一个与之对应的非线性规划问题，前者称之为原始问题或者主要问题(Primal Problem),后者称之为 (Lagrangian dual problem)

◆主问题P：$Min\ f(x), s.t. \begin{cases} g_i(x) \le 0, i = 1,2,\cdots,m \\ h_i(x) = 0, i = 1,2,\cdots,l \\ x \in X \end{cases}$

◆拉格朗日对偶问题D:

$$Max\ \theta(u,v), s.t. u \ge 0,$$

$$\theta(u,v) = inf\{f(x) + \Sigma_{i=1}^{m} u_i g_i(x) + \Sigma_{i=1}^{l} v_i h_i(x), \mathrm{x} \in X\}$$

◆拉格朗日函数

➢ $L(x,u,v) = f(x) + \Sigma_{i=1}^{m} u_i g_i(x) + \Sigma_{i=1}^{l} v_i h_i(x)$

- 注意函数$L(x,u,v)=f(x)+\Sigma_{i=1}^{m}u_i g_i(x)+\Sigma_{i=1}^{l}v_i h_i(x)$，其输入为三变量，一般称为Lagrangian函数，u,v分别为对应不等式约束和等式约束的拉格朗日乘子向量或者对偶变量
- 注意θ可取$-\infty$，有时也称$\theta(u,v)$为拉格朗日对偶函数，此时约束都通过拉格朗日乘子或者对偶变量u_i,v_i整合到目标函数当中了。θ是一族u,v的仿射变换的逐点下确界，因而是凹函数，即使主问题不是凸的也成立。为什么？如何证明？
- 这种将约束采用拉格朗日乘子或对偶办法结合到目标函数中的办法称之为对偶化(Dualization)，需要注意：与不等式相关的乘子$u_i \geq 0$，而与等式约束相关的乘子v_i则没有限制

- 拉格朗日函数给原优化问题的最优值提供了下界，且该下界依赖于参数λ, μ的选取
- 弱对偶原理： 对于任意的$u \geq 0, v$,拉格朗日函数给出了优化问题的最优值p^*的一个下界，即
- $\theta(u, v) \leq p^*, u \geq 0$
- 证明（略）：只需要按照定义，代入拉格朗日函数的定义中，再根据θ函数的inf定义即可。

- **二阶必要条件**
- 问题$min\ f(x), s.t.\ g(x) \le 0, h(x) = 0$,设$x^* \in X, g_i(x)(i \in I), h_j(x), j = 1,2,\cdots,l$在$x^*$二次可微，$g_i(x), i \notin I$在$x^*$连续。约束规范$\{\nabla g_i(x^*), i \in I; \nabla h_j(x^*), j = 1,2,\cdots,l\}$线性无关。如果$x^*$是局部最优解，那么存在$u_i \ge 0 (i \in I)$,$v \in R^l$,使得KKT条件式成立。由乘子构造拉格朗日函数：

$$L(x,u,v) = f(x) + \sum_{i \in I} u_i g_i(x) + \sum_{j=1}^{l} v_j h_j(x)$$

- （临界锥）令$S_h = \{d | \nabla h_j^T(x^*) d = 0, j = 1,2,\cdots,l\}$，$S_{g_I} = \{d | \nabla g_i^T(x^*) d = 0, i \in I\}$
- 则$d \in S_{g_I} \cap S_h$,有$d^T \nabla_x^2 L(x^*,u,v) d \ge 0$

◆ 二阶充分条件

◆ 问题$min\ f(x), s.t. g(x) \le 0, h(x) = 0$,设$x^* \in X$是KT点，即满足拉格朗日函数在该点梯度为0,且满足互补条件。如果$d \in S_h \cap \{d|\nabla f^T(x^*)d = 0, \nabla g_i^T(x^*)d \le 0, i \in I\}$,且$d \ne 0$, 均有$d^T \nabla_x^2 L(x^*, u, v)d > 0$,则$x^*$是最优化问题的严格局部最优解

◆ 例子：$min\ x_1^2 + x_2^2, s.t. \frac{x_1^2}{4} + x_2^2 - 1 = 0$

- $L(x, \lambda) = x_1^2 + x_2^2 + \lambda(\frac{x_1^2}{4} + x_2^2 - 1)$，则任意点$(x_1, x_2)^T$处的LFD为$\{(d_1, d_2)|d \cdot \nabla c_i = 0\} \Rightarrow \frac{x_1}{4} d_1 + x_2 d_2 = 0$,线性约束规范成立，因此临界锥与LFD一样。求出其四个KKT点对分别为$(2, 0, -4), (-2, 0, -4)$和$(0, 1, -1), (0, -1, -1)$
- 考虑第1和第3个点，计算$\nabla_{xx}^2 L(\cdot) = \begin{pmatrix} 0 & 0 \\ 0 & -6 \end{pmatrix}$,临界锥为$\{(d_1, d_2)|d_1 = 0\}$
- 取$d = (0, 1)^T \Rightarrow d^T \nabla_{xx}^2 L(\cdot) d = -6 < 0$
- 而对第3个点，$\nabla_{xx}^2 L(\cdot) = \begin{pmatrix} \frac{3}{2} & 0 \\ 0 & 0 \end{pmatrix}$,临界锥为$\{(d_1, d_2)|d_2 = 0\}$, $d^T \nabla_{xx}^2 L(\cdot) d = \frac{3}{2} d_1^2 > 0$,因此为严格局部最优点

◆以只有不等式约束为例

$$min_{x_1,x_2}\ 0.5\left(x_1^2+x_2^2\right), s.t.\, x_1-1\geq 0$$

◆其拉格朗日对偶$L(x_1,x_2,\lambda_1)=?$

◆如何求解？练习求解该问题！

◆解：$\mathcal{L}(x_1,x_2,\lambda_1)=0.5\left(x_1^2+x_2^2\right)-\lambda_1(x_1-1)$

◆$\theta(\lambda_1)=inf_{x_1,x_2}\{\mathcal{L}(x_1,x_2,\lambda_1)\}\Rightarrow x_1-\lambda_1=0, x_2=0$

◆因此：$\theta(\lambda_1)=0.5\lambda_1^2-\lambda_1(\lambda_1-1)=-0.5\lambda_1^2+\lambda_1$

◆对偶问题为：$max_{\lambda_1\geq 0}\theta(\lambda_1)$,从而得出其最优解为：

◆$\lambda_1=1$, 最优目标函数值：0.5

- 由于对偶问题包含最大化函数$f(x)+\Sigma_{i=1}^{m}u_i g_i(x)+\Sigma_{i=1}^{l}v_i h_i(x)$的下确界(Infimum)，因而有时也称之为*Max-min Dual Problem*
- 因此严格来讲，对偶问题应该为$D: sup\{\theta(u,v): u\geq 0\}$而不是$max\{\theta(u,v): u\geq 0\}$，因为最大值可能不存在
- 以后一般可以将约束简记为$g(x)\leq 0, h(x)=0, x\in X$，此时$g, h$均为向量函数，包括多个约束
- 相应的对偶问题：$max\ \theta(u,v), \theta(u,v)=inf\{f(x)+u^T g(x)+v^T h(x), x\in X$

练习:线性规划问题的对偶，即拉格朗日对偶

◆ **线性规划问题的对偶**

◆ **考虑如下线性规划问题：$min_x c^T x, s.t. Ax = b, x \geq 0$**

➢ 对等式约束，引入拉格朗日乘子v;对非负约束，引入拉格朗日乘子$u \geq 0$

➢ 构造如下拉格朗日函数

➢ $L(x,u,v) = c^T x + v^T(Ax - b) - u^T x = -b^T v + \left(A^T v - u + c\right)^T x$,其拉格朗日对偶函数为：

➢ $\theta(u,v) = inf_x L(x,u,v) = \begin{cases} -b^T v & A^T v - u + c = 0 \\ -\infty & \text{其他} \end{cases}$

➢ 只需考虑$A^T v - u + c = 0$的情形，其余对应不可行解。因此线性规划问题的对偶问题为：

➢ $max_{u,v} - b^T v, s.t. A^T v - u + c = 0, u \geq 0$,变量替换$y = -v$,则转换为：

➢ $max_{u,y} b^T y, s.t. A^T y + u = c, u \geq 0$

◆ **问题：这种形式好像与第一章的情况不一样？**

➢ 如果保留$x \geq 0$,直接写等式约束的乘子为y,拉格朗日函数应该如何写？

◆线性规划问题的对偶

- 如果保留$x \geq 0$,直接写等式约束的乘子为y,拉格朗日函数应该如何写?
- $L(x,y) = c^T x - y^T(Ax - b) = b^T y + \left(c - A^T y\right)^T x$
- 则对偶问题需要将$x \geq 0$添加到约束中
- $max_y \left\{inf_x b^T y + \left(c - A^T y\right)^T x, s.t. x \geq 0\right\}$,
- 从而得:
- $max_y b^T y, s.t. A^T y \leq c \quad (*)$

◆问题:上述(*)式的对偶问题是什么?

- 等价地:$min_y - b^T y, s.t. A^T y \leq c$
- 对不等式约束引入拉格朗日乘子$x \geq 0$,则拉格朗日函数为$L(y,x) = -b^T y + x^T\left(A^T y - c\right) = -c^T x + (Ax - b)^T y$
- 因此对偶函数$\theta(x) = inf_y L(y,x) = \begin{cases} -c^T x & Ax = b \\ -\infty & \text{其他} \end{cases}$
- 相应的对偶问题是:$max_x - c^T x, s.t. Ax = b, x \geq 0$,这与原问题等价!

◆ l_1正则化问题的对偶

- $min_{x\in R^n}\frac{1}{2}||Ax-b||^2+\mu||x||_1, A\in R^{m\times n}, b\in R^m, \mu$为正则化参数
- 引入$Ax-b=r$,则问题转化为
- $min_{x\in R^n}\frac{1}{2}||r||^2+\mu||x||_1, s.t. Ax-b=r$
- 此时，其拉格朗日函数为$L(x,r,\lambda)=\frac{1}{2}||r||^2+\mu||x||_1-\lambda^T(Ax-b-r)=\frac{1}{2}||r||^2+\lambda^T r+\mu||x||_1-\left(A^T\lambda\right)^T x+b^T\lambda,$
- 这是一个二次函数，从而有

对偶范数：$||x||_*=sup\{z^T x, ||x||\le 1\}$

- $\theta(\lambda)=inf_{x,r}L(x,r,\lambda)=\begin{cases} b^T\lambda-\frac{1}{2}||\lambda||^2 & ||A^T\lambda||_\infty\le\mu \\ -\infty & \text{其他} \end{cases}$
- 因此对偶问题为：
- $max\ b^T\lambda-\frac{1}{2}||\lambda||^2, s.t. ||A^T\lambda||_\infty\le\mu$

练习

- 1.$min\ c^Tx, s.t. Ax-b\geq 0$的对偶问题
- 2. $min\frac{1}{2}x^TGx+c^Tx, s.t. Ax-b\geq 0$的对偶问题，$G$对称正定
- 解1：$\theta(\lambda)=inf_x\{\mathcal{L}(x,\lambda)\}=inf_x\{c^Tx-\lambda^T(Ax-b)\}=inf_x\{(c^T-\lambda^TA)x+\lambda^Tb\}$
  - 显然，若$c^T-\lambda^TA\neq 0$,则$\theta(\lambda)=-\infty$
  - 若$c^T-\lambda^TA=0$,对偶目标函数即为：λ^Tb.
- 最大化$\theta(\lambda)$时，可以排除$c^T-\lambda A\neq 0$的情况，因为不可能在$q(\lambda)=-\infty$的地方取得最大值.因此问题转化为：

$$max_\lambda b^T\lambda, s.t. A^T\lambda=c, \lambda>0$$

- 解2：$\theta(\lambda)=inf_x\{\mathcal{L}(x,\lambda)\}=inf_x\left\{\frac{1}{2}x^TGx+c^Tx-\lambda^T(Ax-b)\right\}$,凸函数，因此必定有：$\nabla_x\mathcal{L}(x,\lambda)=0\Rightarrow Gx+c-A^T\lambda=0, \lambda\geq 0$

$$q(\lambda)=-\frac{1}{2}(A^T\lambda-c)^TG^{-1}(A^T\lambda-c)+b^T\lambda$$

- 进一步：$(c-A^T\lambda)^Tx=-x^TGx$,可将其化为Wolfe对偶形式(包含x,λ作为变量)：

$$max_{\lambda,x}\left\{\frac{1}{2}x^TGx+c^Tx-\lambda^T(Ax-b)\right\}, s.t. Gx+c-A^T\lambda=0, \lambda\geq 0$$

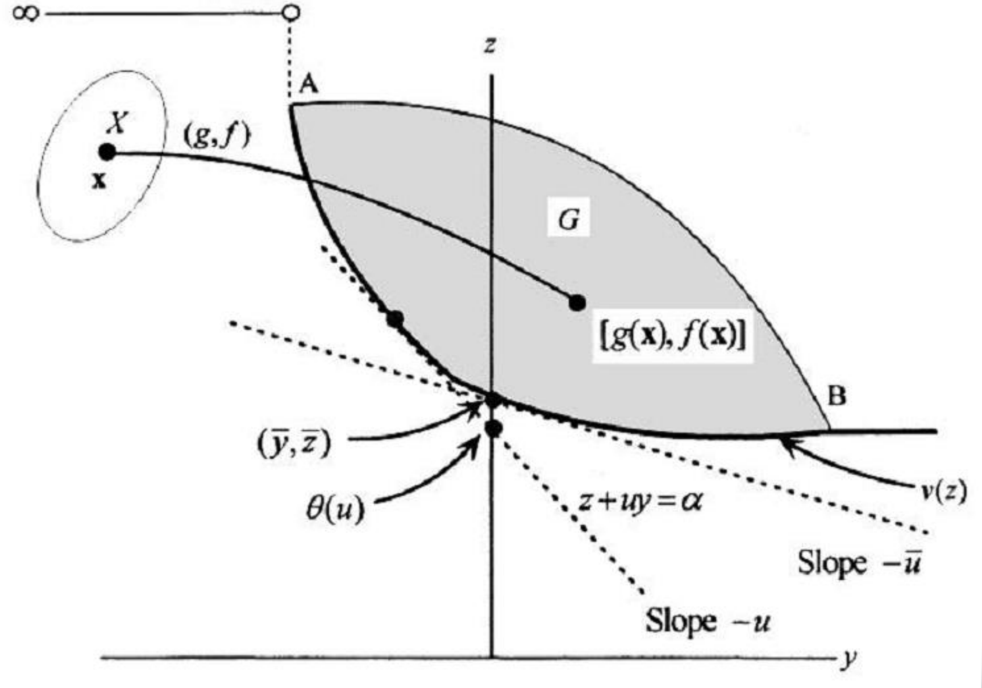
- 为使目标函数为凹函数看的更清楚, $(c-A^T\lambda)^Tx=-x^TGx$,得

$$max_{\lambda,x}\left\{-\frac{1}{2}x^TGx+\lambda^Tb\right\}, s.t. Gx+c-A^T\lambda=0, \lambda\geq 0$$

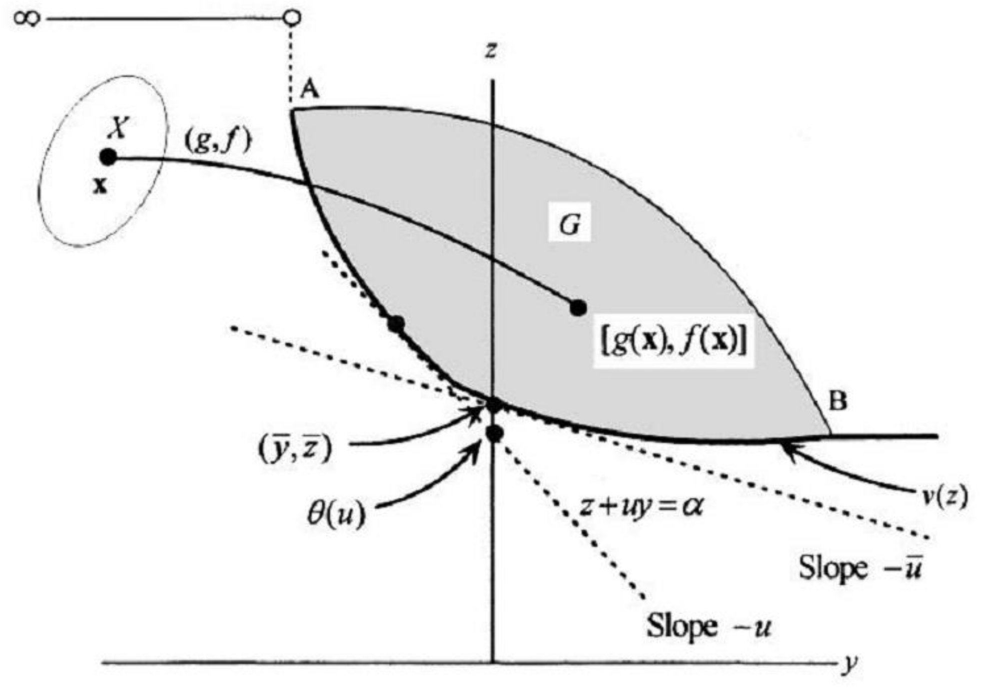
◆对偶问题的几何解释

- 为简化讨论，只讨论一个不等式约束$g(x) \le 0$，没有等式约束的情况
- 在(y, z)平面内，集合$\{(y, z): y = g(x), z = f(x), x \in X\}$用$G$表示,则$G$是$(g, f)$映射下$X$的像

原问题：就是在$y \le 0$的G中找到一个最小的点，使得其有最小的纵坐标，该点为$(\bar{y}, \bar{z})$



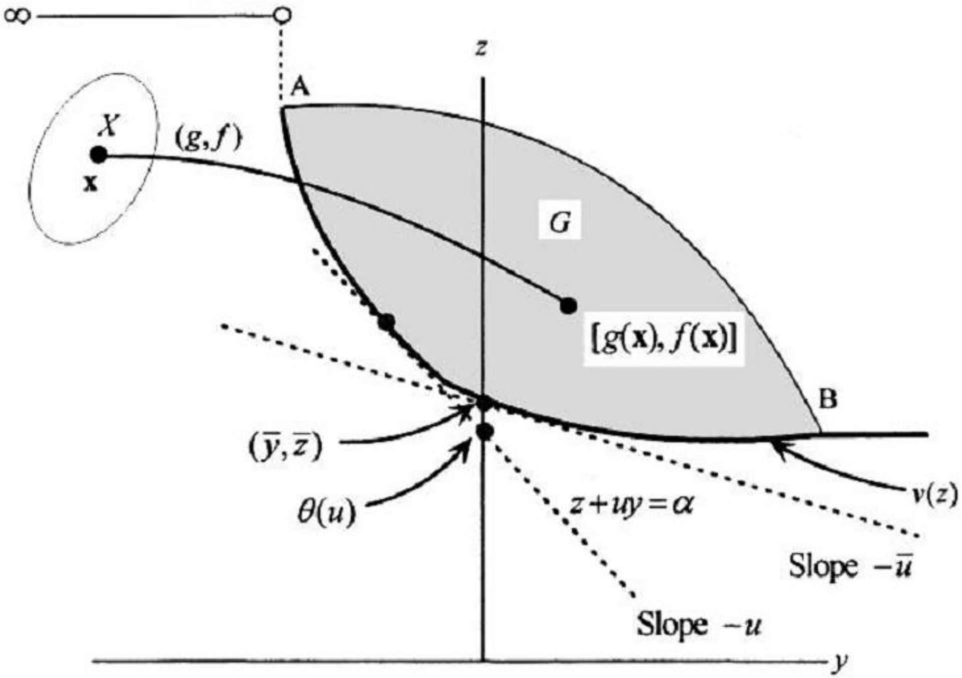
- 现在假设$u \geq 0$,为了确定$\theta(u)$,需要$min\ f(x) + ug(x), x \in X$.令$y = g(x), z = f(x), x \in X$,我们需要在$G$中最小化$z + uy$,显然$z + uy = \alpha$是一条斜率为$-u$且在z轴上截距为$\alpha$的直线，为了最小化$z + uy$,需要沿负梯度方向来移动直线直到与$G$相切为止，此时得到的z轴的截距为$\theta(u)$
- 因此，对偶问题等价于找到一个支撑超平面的斜率并在z轴上的截距最大，从图中可以看出，这样的超平面斜率为$-\bar{u}$, 支撑点为$(\bar{y}, \bar{z})$, 因此最优对偶解为$\bar{u}$,最优对偶目标函数值为$\bar{z}$. 此时，原问题与对偶问题的最优值相等



- **如果此时令$v(y) = min\{f(x): g(x) \le y, x \in X\}$，则称这个函数为扰动函数(Perturbation function),从函数可以看出，只是原问题不等式约束的右边项由0变成了y,如图所示，$v(y)$为y的非递增函数。从中看到v相应于G的AB点之间的下包络，注意其单调递减的性质。**

- 且v在B点对于大于y的值都为取值对应B点的常数值，而在A点左边，由于其不可行性，此时v取值为∞，注意v在0点可微，且$v'(0) = -\bar{u}$, 因此当约束右端从0增加时，目标函数值的变化率等价$-\bar{u}$,即最优点的对偶拉格朗日乘子的负数，如果v是凸但不可微，则$-\bar{u}$为在0点的次梯度，任何情况都满足$v(y) \ge v(0) - \bar{u}y, y \in R$

- 如v不是凸的或不可微，则上述条件仍然成立等价于$\bar{u}$是相应于最优解$\bar{x}$的KKT拉格朗日乘子



◆例子：$Min\ x_1^2+x_2^2, s.t. \begin{cases} -x_1-x_2+4\le 0 \\ x_1,x_2\ge 0 \end{cases}$,显然最优解为$(x_1,x_2)=(2,2)$,目标函数值为8　练习:如何用对偶方法来求解?

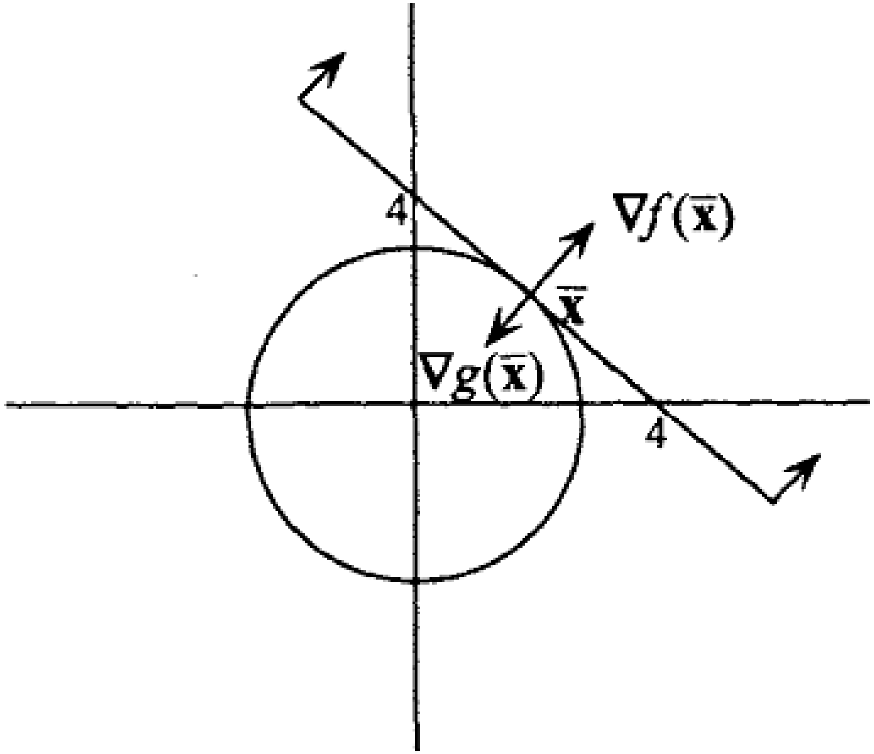
◆令$g(x)=-x_1-x_2+4, X=\{(x_1,x_2): x_1,x_2\ge 0\}$,对偶函数为$\theta(u)=inf\{x_1^2+x_2^2+u(-x_1-x_2+4): x_1,x_2\ge 0\}$

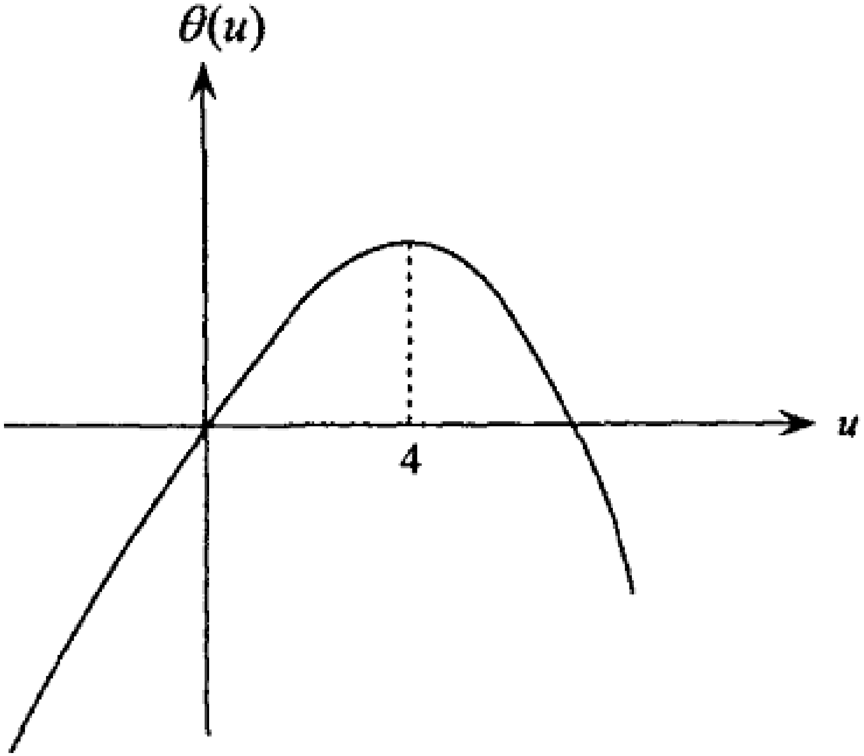
$$=inf\{x_1^2-ux_1: x_1\ge 0\}+inf\{x_2^2-ux_2\ge 0\}+4u$$

◆注意，上述下确界如果$u\ge 0$则在$x_1=x_2=\frac{u}{2}$达到,如果$u<0$,则在$x_1=x_2=0$达到，因此$\theta(u)=$

$$\begin{cases} -\frac{1}{2}u^2+4u, u\ge 0 \\ 4u, u<0 \end{cases}$$

◆注意θ是一个凹函数，$u \geq 0$时其最大点发生在$\bar{u} = 4$,注意最优原目标函数和对偶目标函数值都等于8

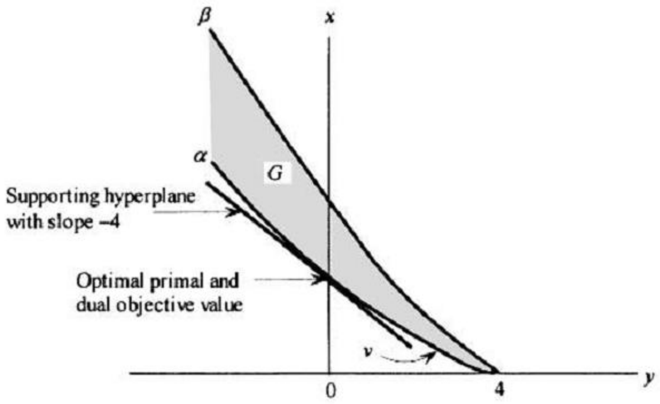




第6章约束最优化理论(Theory of Constrained Optimization)-对偶问题

- 现在考虑(y,z)平面，这里$y=g(x), z=f(x)$,$X=\{(x_1,x_2), x_1\geq 0, x_2\geq 0\}$的在映射$(g,f)$的像$G$是什么？
- 给定y,用$\alpha(y),\beta(y)$表示下列问题P_1,P_2的最优目标函数值
- P_1: $min\ x_1^2+x_2^2, s.t. -x_1-x_2+4=y, x_1,x_2\geq 0$
- P_2: $max\ x_1^2+x_2^2, s.t. -x_1-x_2+4=y, x_1,x_2\geq 0$

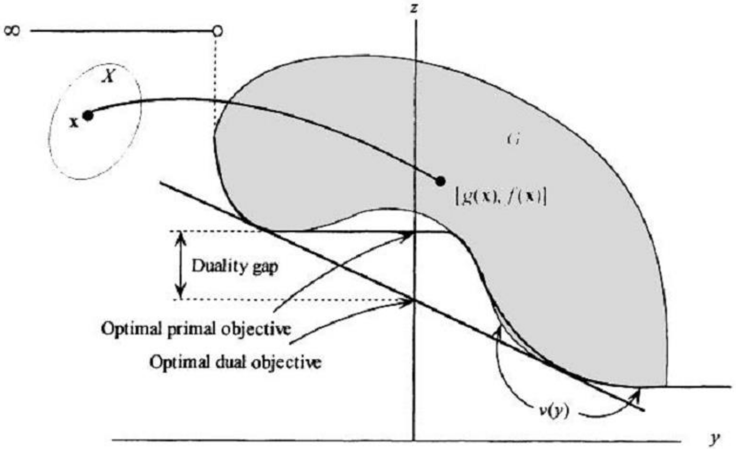
$$Min\ x_1^2+x_2^2, s.t. \begin{cases} -x_1-x_2+4\leq 0 \\ x_1,x_2\geq 0 \end{cases}$$



- 易知$\alpha(y)=\frac{(4-y)^2}{2}, \beta(y)=(4-y)^2, y\leq 4$
- 注意最优对偶解$\bar{u}=4$,为图中支持超平面的负斜率。最优对偶目标函数值$\alpha(0)=8$等于原问题的最优值
- 大家可以自行验证前面提到的解释，例如，v对于$y\leq 4$相应于$\alpha(y)$的下包络，对于$y\geq 4$,$v(y)$保持常数值0，斜率$v'(0)=-4$，相应于最优的对偶拉格朗日乘子值的负数，且$v(y)\geq v(0)-4y, y\in R$,注意，这是原问题和对偶问题具有相等最优解的充要条件

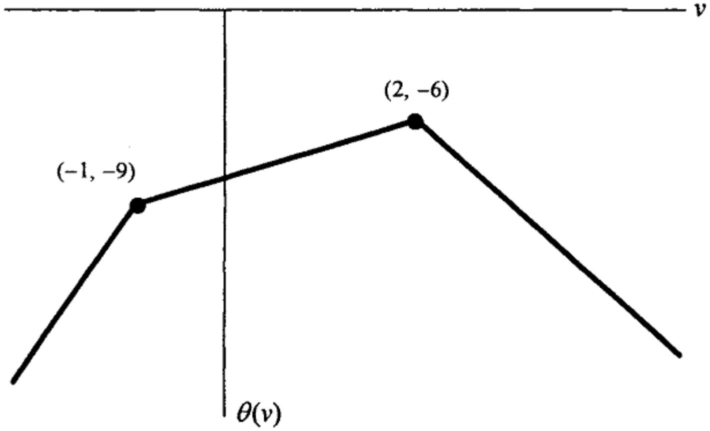
第6章约束最优化理论(Theory of Constrained Optimization)-对偶问题

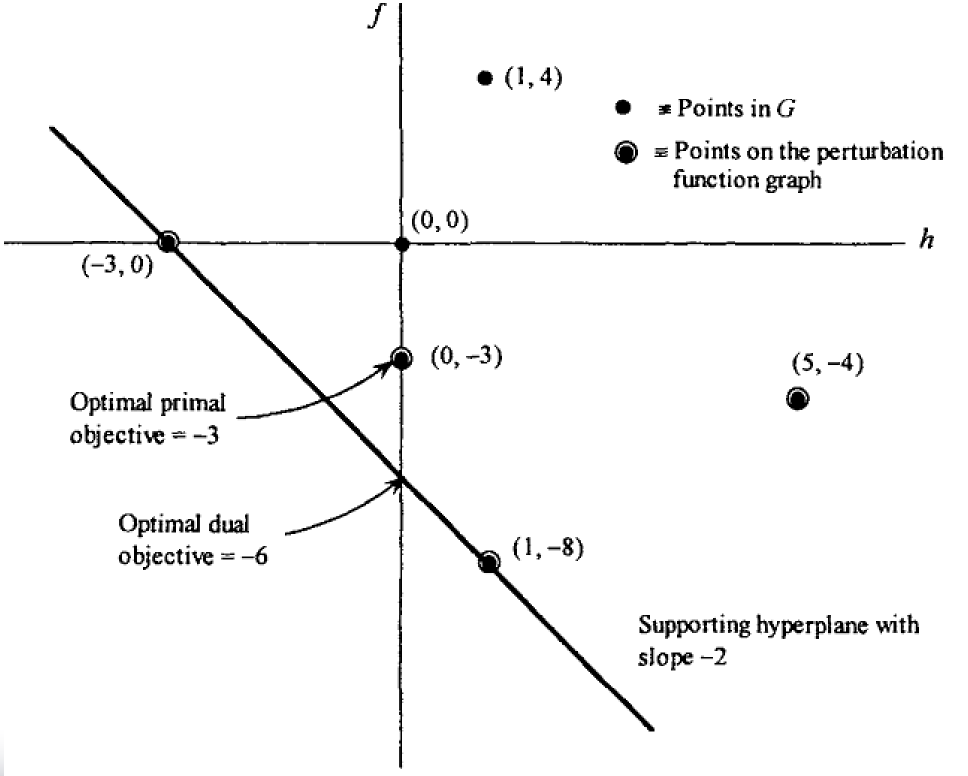
- **定理(弱对偶定理)：令x是问题P的可行解，即$x \in X, g(x) \le 0, h(x) = 0$.并令$(u, v)$是问题$D$的可行解，即$u \ge 0$,则有$f(x) \ge \theta(u, v)$。对偶问题估计下界**
- **证明：$\theta(u, v) = inf\{f(y) + u^T g(y) + v^T h(y): y \in X\} \le f(x) + u^T g(x) + v^T h(x) \le f(x)$证毕！**



- **推论1(Corollary):**
- **$inf\{f(x): x \in X, g(x) \le 0, h(x) = 0\} \ge sup\{\theta(u, v): u \ge 0\}$**
- **推论2：$f(\overline{x}) = \theta(\overline{u}, \overline{v}), \overline{u} \ge 0$, 且$\overline{x} \in \{x \in X: g(x) \le 0, h(x) = 0\}$,则$\overline{x}$和$(\overline{u}, \overline{v})$分别是原问题和对偶问题的解**
- **推论3:如果$inf\{f(x): x \in X, g(x) \le 0, h(x) = 0\} = -\infty$,则对每个u ≥ 0, 都有$\theta(u, v) = -\infty$**
- **推论4：如果$sup\{\theta(u, v): u \ge 0\} = \infty$,则原问题没有可行解**
- **对偶间隙(Duality Gap):若上述对偶定理中原问题大于对偶问题的不等式严格成立，则说明存在对偶间隙**

- 例：$min\ f(x)=-2x_1+x_2, s.t.\ h(x)=x_1+x_2-3=0, (x_1,x_2)\in X, X=\{(0,0),(0,4),(4,4),(4,0),(1,2),(2,1)\}$
- 易验证(2, 1)是原问题的最优解，目标函数值为-3，对偶目标函数$\theta(v)=min\{(-2x_1+x_2)+v(x_1+x_2-3):(x_1,x_2)\in X\}$,此时$\theta(v)=\begin{cases}-4+5v, v\le -1\\ -8+v, -1\le v\le 2\\ -3v, v\ge 2\end{cases}$,最优解$\overline{v}=2$,目标函数值-6. 存在对偶间隙
- 集合解释如下右图所示
- 此时不存在$v(y)\ge v(0)-\overline{v}y$





第6章约束最优化理论(Theory of Constrained Optimization)-对偶问题

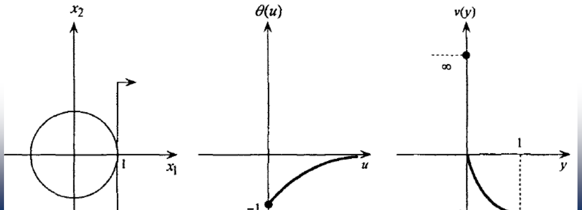
- 引理:令X为R^n非空凸集,$\alpha: R^n \to R, g: R^n \to R^m$为凸，令$h: R^n \to R^l$为仿射变换，即$h(x) = Ax - b$,则下述系统1无解，则系统2有解$(u_0, u, v)$,如果$u_0 > 0$,反之也成立
  - 系统1：对某$x \in X, \alpha(x) < 0, g(x) \le 0, h(x) = 0$
  - 系统2：对所有$x \in X, u_0\alpha(x) + u^T g(x) + v^T h(x) \ge 0$
- 在某些约束条件和凸性假设下，对偶问题和原问题的目标函数值相等
- 强对偶定理(Strong Duality Theorem)：令X为R^n非空凸集,$\alpha: R^n \to R, g: R^n \to R^m$为凸，令$h: R^n \to R^l$为仿射变换，即$h(x) = Ax - b$,假设下列条件成立：存在某个$\hat{x} \in X$,使得$g(\hat{x}) < 0, h(\hat{x}) = 0, 0 \in int\ h(X)$, 这里$h(X) = \{h(x): x \in X\}$,则$inf\{f(x): x \in X, g(x) \le 0, h(x) = 0\} = sup(\theta(u, v): u \ge 0\}$,而且如果下确界是有限值，则在$(\bar{u} \ge 0, \bar{v})$点达到$sup\{\theta(u, v): u \ge 0\}$,如果下确界在$\bar{x}$点达到，则$\bar{u}^T g(\bar{x}) = 0$

- 上述定理表明，原问题和对偶问题在某些条件下其目标函数值是一样的，这等价于鞍点问题的存在性
- 给定原问题P,定义拉格朗日函数$\phi(x,u,v)=f(x)+u^Tg(x)+v^Th(x)$,其解$(\bar{x},\bar{u},\bar{v})$称为**拉格朗日函数的鞍点**，如果$\bar{x}\in X,\bar{u}\geq 0$,且$\phi(\bar{x},u,v)\leq\phi(\bar{x},\bar{u},\bar{v})\leq\phi(x,\bar{u},\bar{v})$对所有$x\in X,(u,v),u\geq 0$都成立($\bar{x}=min_x\,\phi(x,\bar{u},\bar{v}),(\bar{u},\bar{v})=max_{u,v}\,\phi(\bar{x},u,v)$)
- 鞍点最优性和不存在对偶间隙定理：$(\bar{x},\bar{u},\bar{v})$为拉格朗日函数$\phi(x,u,v)=f(x)+u^Tg(x)+v^Th(x)$的鞍点当且仅当
  - $\phi(\bar{x},\bar{u},\bar{v})=min\{\phi(x,\bar{u},\bar{v}):x\in X\}$
  - $g(\bar{x})\leq 0,h(\bar{x})=0$
  - $\bar{u}^Tg(\bar{x})=0$
- 而且，$(\bar{x},\bar{u},\bar{v})$为鞍点当且仅当$\bar{x},(\bar{u},\bar{v})$分别是原问题P和对偶问题D的没有对偶间隙的最优解，即$f(\bar{x})=\theta(\bar{u},\bar{v})$

- 鞍点最优性条件和KKT最优性条件之间的关系
- 令$S=\{x\in X: g(x)\le 0, h(x)=0\}$,问题 P:$min\ f(x), s.t.\ x\in S$. 假设$\bar{x}\in S$满足KKT条件；也即，存在$\bar{u}\ge 0,\bar{v}$使得$\nabla f(\bar{x})+\nabla g(\bar{x})^T\bar{u}+\nabla h(\bar{x})^T\bar{v}=0, \bar{u}^T g(\bar{x})=0$. 假设$f, g_i, i\in I$在$\bar{x}$是凸的，这里$I=\{i: g_i(\bar{x})=0\}$表示积极或者有效约束集。而且如果$\bar{v}_i\ne 0$,则$h_i$是仿射的。则有$(\bar{x},\bar{u},\bar{v})$为拉格朗日函数$\phi(x,u,v)=f(x)+u^T g(x)+v^T h(x)$的鞍点
- 反之，若假设$(\bar{x},\bar{u},\bar{v})$，$\bar{x}\in int\ X$，$\bar{u}\ge 0$为拉格朗日函数$\phi(x,u,v)=f(x)+u^T g(x)+v^T h(x)$的鞍点.则$\bar{x}$为问题P的可行解，而且$(\bar{x},\bar{u},\bar{v})$满足KKT条件

第6章约束最优化理论(Theory of Constrained Optimization)-对偶问题-例子

- $P: min\,\{x_2: x_1 \geq 1,, x_1^2 + x_2^2 \leq 1, (x_1, x_2) \in R^2\}$
- 解为$(\bar{x}_1, \bar{x}_2) = (1, 0)$,目标函数值为0.但该问题虽然是凸优化问题，但最优点不是KKT点，也不存在鞍点
- 若$g(x) = 1 - x_1 \leq 0, X = \{(x_1, x_2): x_1^2 + x_2^2 \leq 1\}$,此时对偶问题为：$sup\{\theta(u): u \geq 0\}, \theta(u) = inf\left\{x_2 + u(1 - x_1): x_1^2 + x_2^2 \leq 1\right\}$
- 易验证最优值在点$x_1 = \frac{u}{\sqrt{1+u^2}}, x_2 = -1/\sqrt{1 + u^2}$达到，因此 $\theta(u) = u - \sqrt{1 + u^2}, u \to \infty, \theta(u) \to 0$,因此$sup\{\theta(u): u \geq 0\} = 0$,但这点不能取到，也就是说max并不存在
- 扰动函数$v(y) = min\left\{x_2: 1 - x_1 \leq y, x_1^2 + x_2^2 \leq 1\right\}$,从而$v(y) = \infty, y < 0; v(y) = -\sqrt{y(2 - y)}, 0 \leq y \leq 1$,$v(y) = -1, y \geq 1$,说明在$v(y), y \in R$的上镜图中在$(0, 0)$不存在任何支撑超平面，$v$在0点的导数为$-\infty$。



- **考虑优化问题$min_{x\in R^n}\frac{1}{2}||x-y||_2^2, s.t. Ax=b$**
- **思考如何求解?**
  - 假设A是行满秩的
  - 构造拉格朗日函数
  - $L(x,\lambda)=\frac{1}{2}||x-y||^2+\lambda^T(Ax-b)$
  - x^*为全局最优解，等价于存在λ^*使得：
  - $\begin{cases} x^*-y+A^T\lambda=0 \\ Ax^*=b \end{cases} \Rightarrow Ax^*-Ay+AA^T\lambda=0 \Rightarrow \lambda=(AA^T)^{-1}(Ay-b)$,将其代入条件式第一式得
  - $x^*=y-A^T(AA^T)^{-1}(Ay-b)$,因此点$y$到集合$\{x|Ax=b\}$的投影为$y-A^T(AA^T)^{-1}(Ay-b)$

◆练习，思考SVM的求解问题

http://www.dataguru.cn/forum.php?mod=viewthread&tid=371987

第6章约束最优化理论(Theory of Constrained Optimization)-总结

- **无约束问题**
  - 一阶和二阶条件
- **有约束问题**
  - 不等式约束
  - 等式和不等式约束
  - 有约束问题的二阶充要条件
- **对偶问题**
- **总结**