组合优化与凸优化

第7章 约束问题的最优化方法(Methods 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

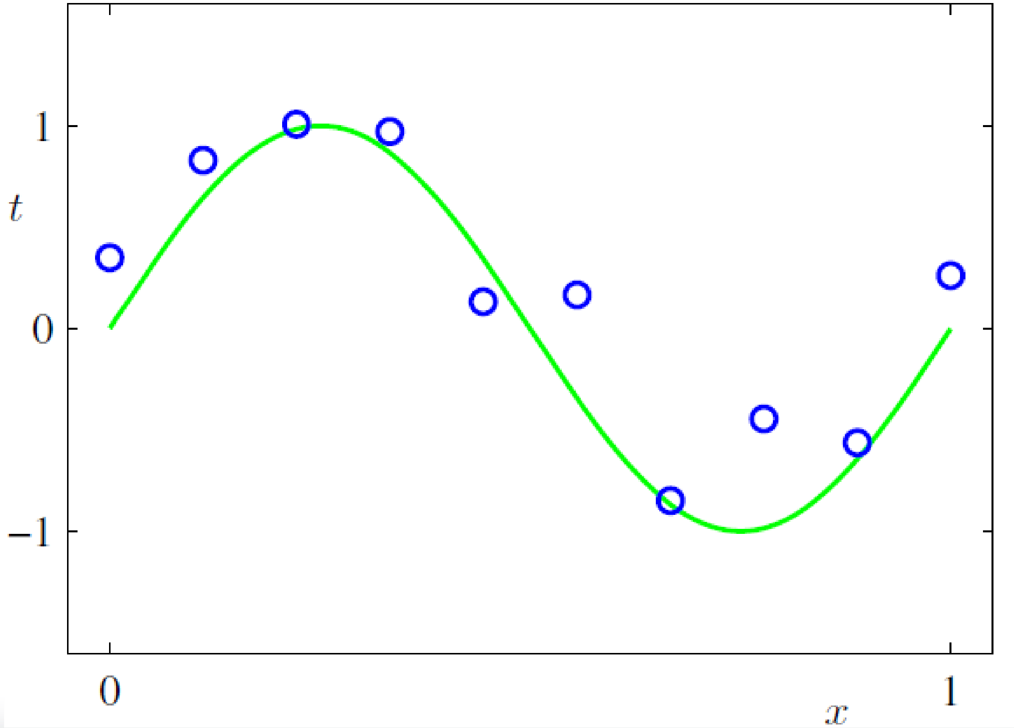
2022年春季

第7章约束问题的最优化方法(Methods of Constrained Optimization)

- **无约束问题**
  - 稳定点
  - 方法回顾
  - 最速下降
  - 牛顿法
  - 拟牛顿法
- **非线性最小二乘**
- **有约束问题**
  - **KT(KKT)**点
  - 可行方向法
  - 近似规划法
  - 制约函数法
  - 二次规划
- **总结**

拟合

- 10个训练数据点：圆圈代表 输入变量x的观察值，t代表对应的目标值，绿色线表示$sin(2\pi x)$，用来生成数据
- 问题：如何预测任意输入变量x所对应的输出值t?注意,绿色线未知

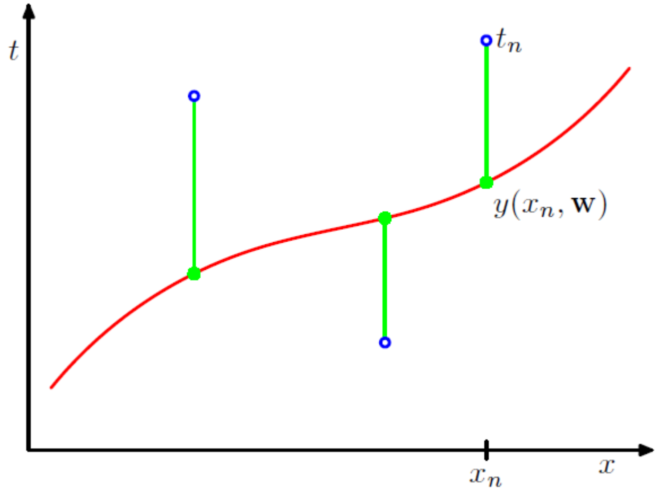


拟合

◆简单采用多项式来进行逼近

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

◆未知量与目标，函数之间为线性关系，称为线性模型，如何来确定系数w_i呢?



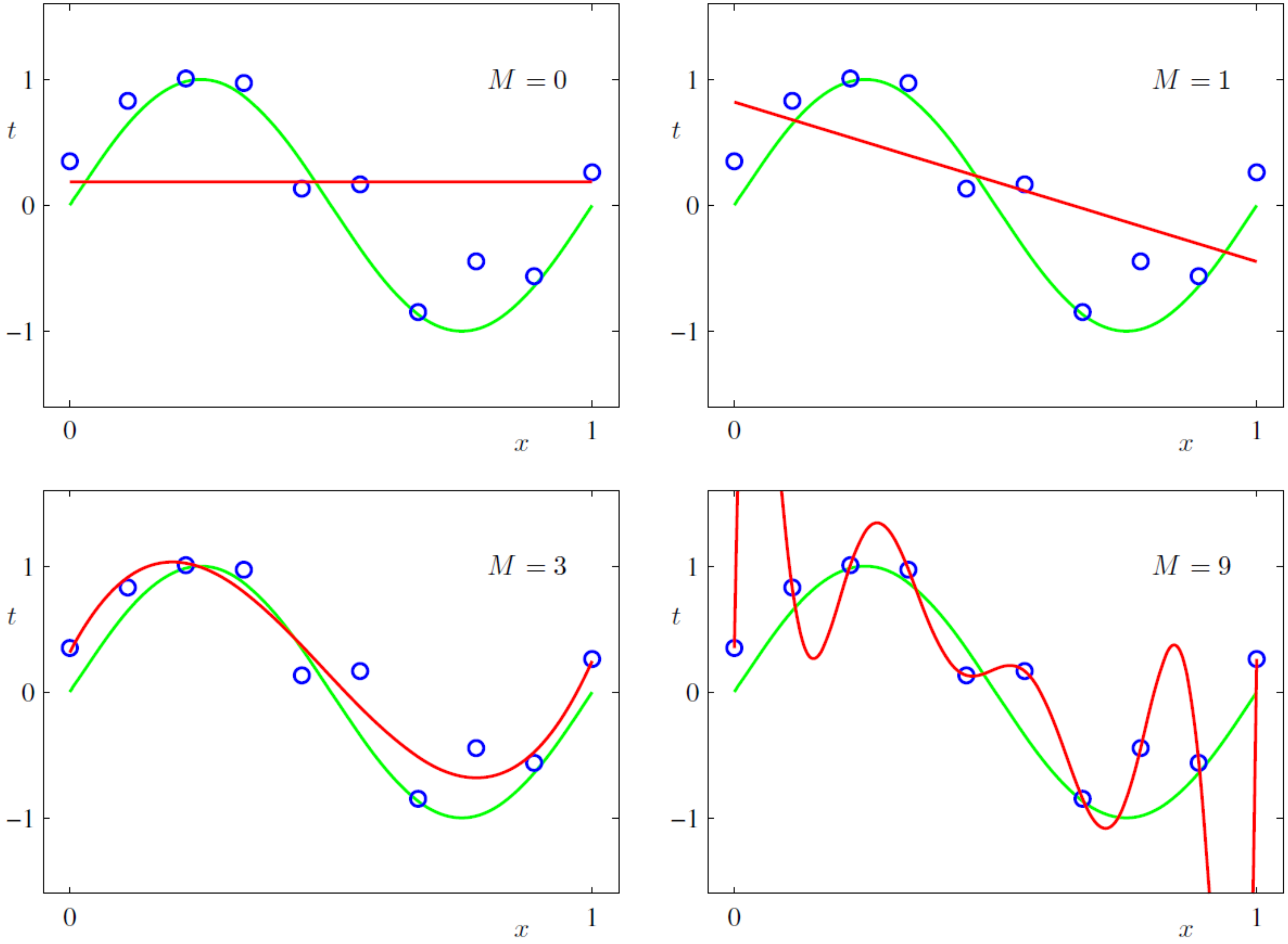
◆通过训练数据，如何做?

◆形式化为一个最优化问题

$$\min E(\boldsymbol{w}), E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2$$

拟合

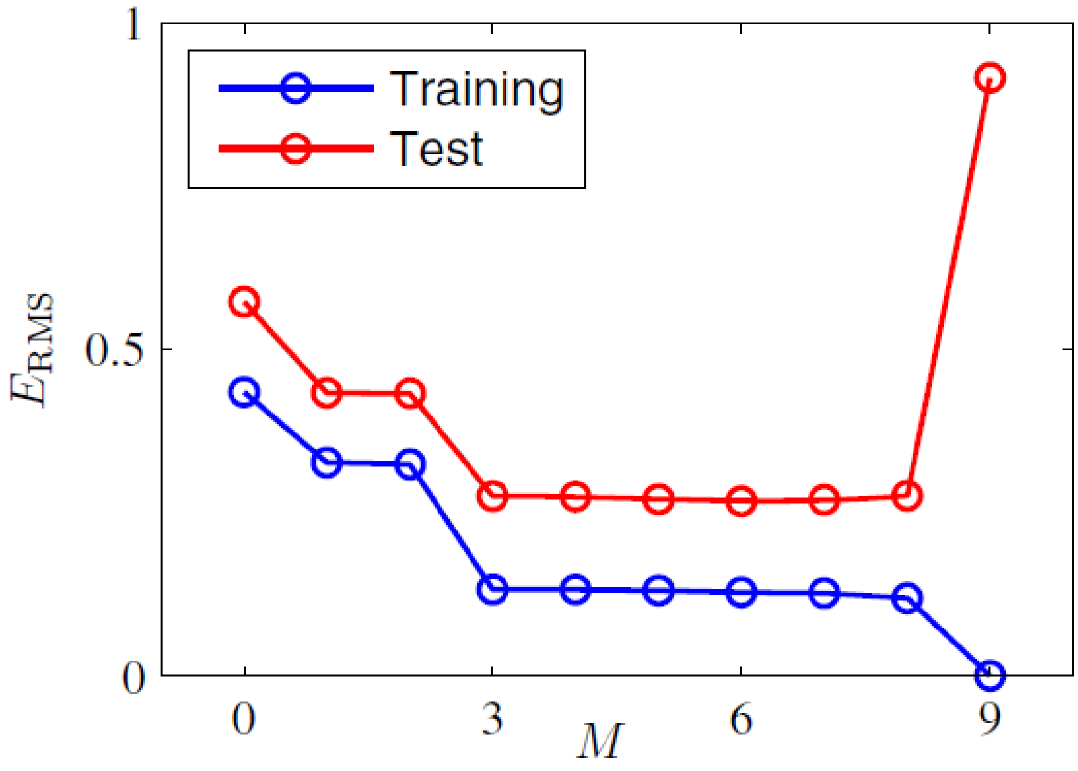
◆如何求解？不同的*M*有什么影响？



拟合

◆采用$Root-Mean-Squre(RMS)$误差评价：

$$E_{RMS}=\sqrt{2E(w^*)/N}$$



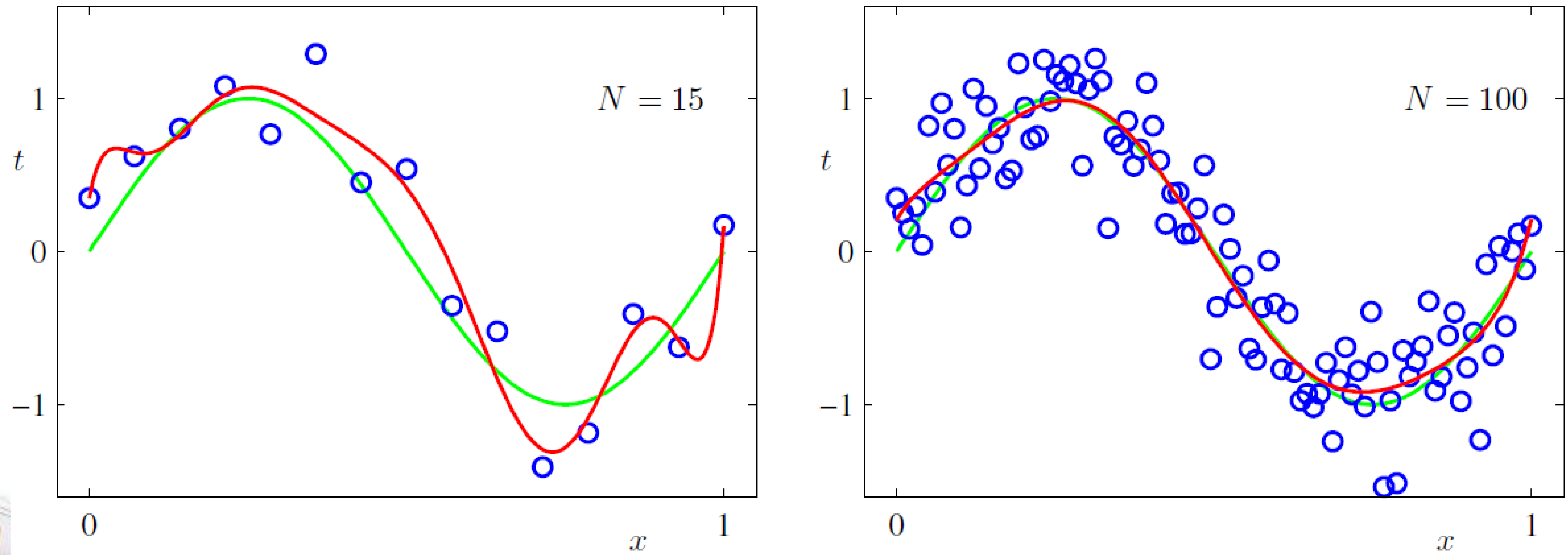
拟合

◆权系数的情况

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

拟合

◆提升训练数据量($M = 9$)



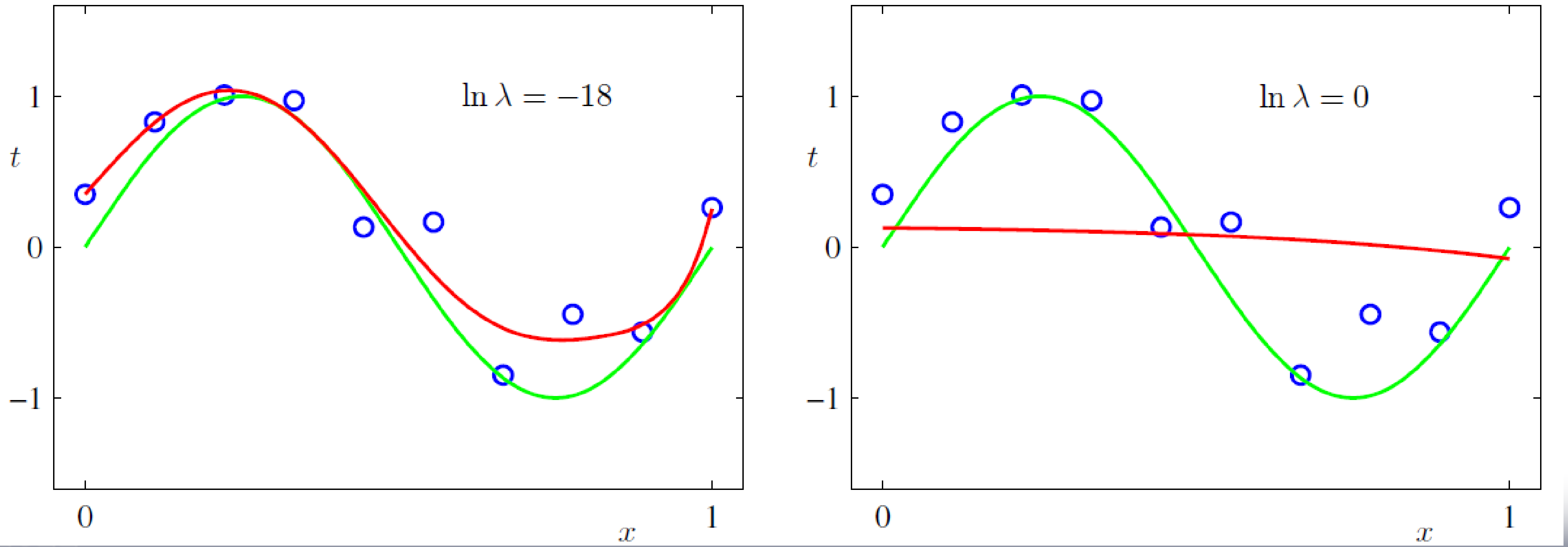
◆一般数据大小为参数的5-10倍左右，那是否需要限制模型参数的数量呢？

拟合

- 如何改进？正则化方法(Regularization)
- 对权系数进行限定，对近似模型进行约束

$$\widetilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}||w||_2^2$$

- 系数λ控制正则项与误差项之间的均衡程度



拟合

◆正则因子对权系数的影响

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

拟合

◆有什么用处？还能改进吗？具体应该如何求解？

- **Nonlinear Least-Squares Problems**
- **问题可表示为：**$\min_{x\in R^n} f(x) = \frac{1}{2} r(x)^T r(x) = \frac{1}{2}\sum_{i=1}^{m}\left(r_i(x)\right)^2, m \ge n$
- **其中**$r: R^n \to R^m$**是**x**的非线性函数，如果**$r(x)$**是线性函数，则问题成为线性最小二乘，例如线性回归**
- **本问题在数据拟合，参数估计，函数逼近等方面有广泛的应用。例如要求拟合数据**$(t_i, y_i), i = 1, 2, \cdots, m$**,拟合函数为**$\phi(t, x)$**是**$x$**的非线性函数，要求选取**$x$**使得拟合函数**$\phi(t, x)$**在残量平方和意义上尽可能好的拟合数据，其中残量**$r_i(x) = \phi(t_i, x) - y_i, i = 1, \cdots, m$**,通常**$m \gg n$**,从而可得到上述的非线性最小二乘问题**
- **设**$J(x)$**是**$r(x)$**的Jacobi矩阵，即**$J(x) = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \cdots & \frac{\partial r_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m}{\partial x_1} & \cdots & \frac{\partial r_m}{\partial x_n} \end{pmatrix}$**,则其梯度和Hessian矩阵可计算如下**

$$\nabla f(x) = g(x) = \sum_{i=1}^{m} r_i(x)\nabla r_i(x) = J(x)^T r(x),$$

$$H(f(x)) = G(x) = \sum_{i=1}^{m}\left(\nabla r_i(x)\nabla r_i(x)^T + r_i(x)\nabla^2 r_i(x)\right) = J(x)^T J(x) + S(x),$$

其中$S(x) = \sum_{i=1}^{m} r_i(x)\nabla^2 r_i(x)$

- 此时考虑目标函数的二次近似模型为$m_k(x)$

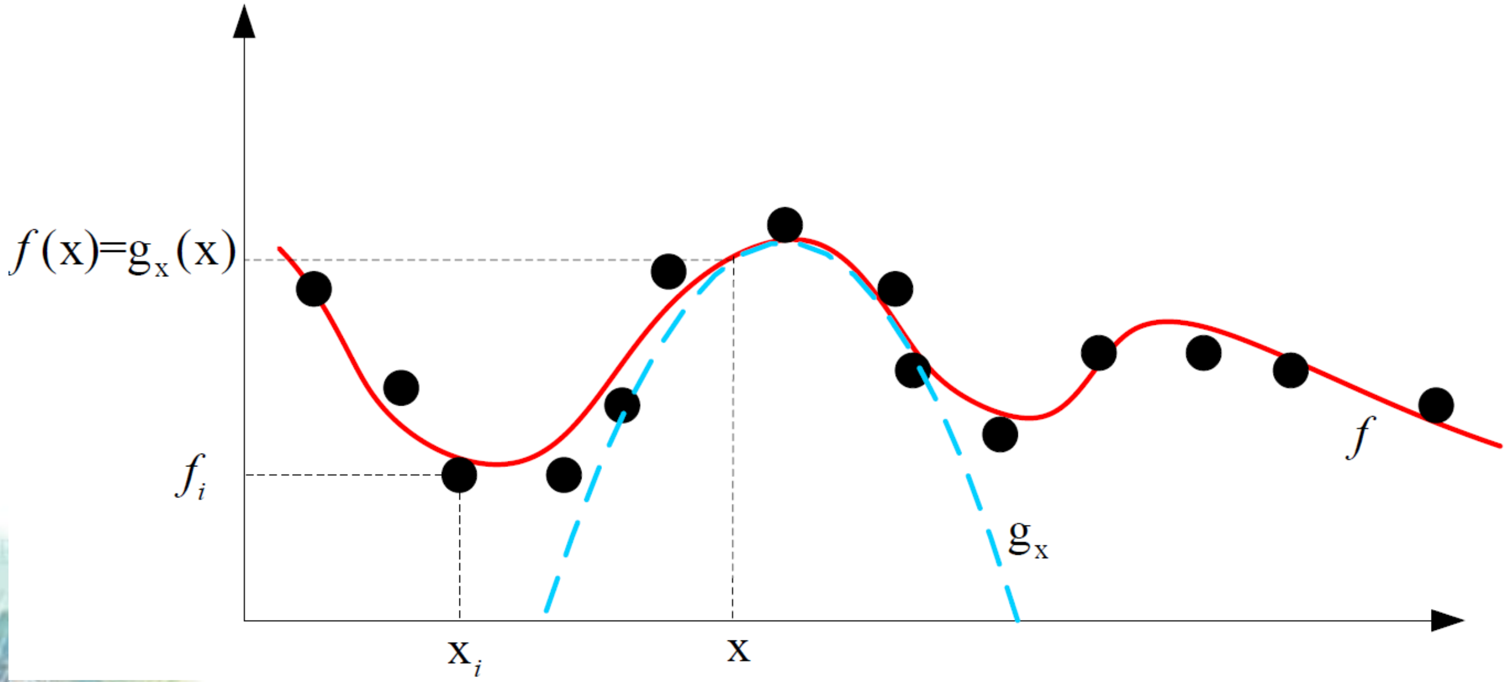
$$m_k(x)=f(x_k)+g(x_k)^T(x-x_k)+\frac{1}{2}(x-x_k)^T G(x_k)(x-x_k)=\frac{1}{2}r(x_k)^T r(x_k)+\left(J(x)^T r(x)\right)^{\mathrm{T}}(x-x_k)+\frac{1}{2}(x-x_k)^{\mathrm{T}}(J(x)^T J(x)+S(x))(x-x_k),$$

因此，如果采用牛顿法，则其迭代公式应该是？

$$x_{k+1}=x_k-\left(J(x)^T J(x)+S(x)\right)^{-1}J(x_k)r(x_k)$$

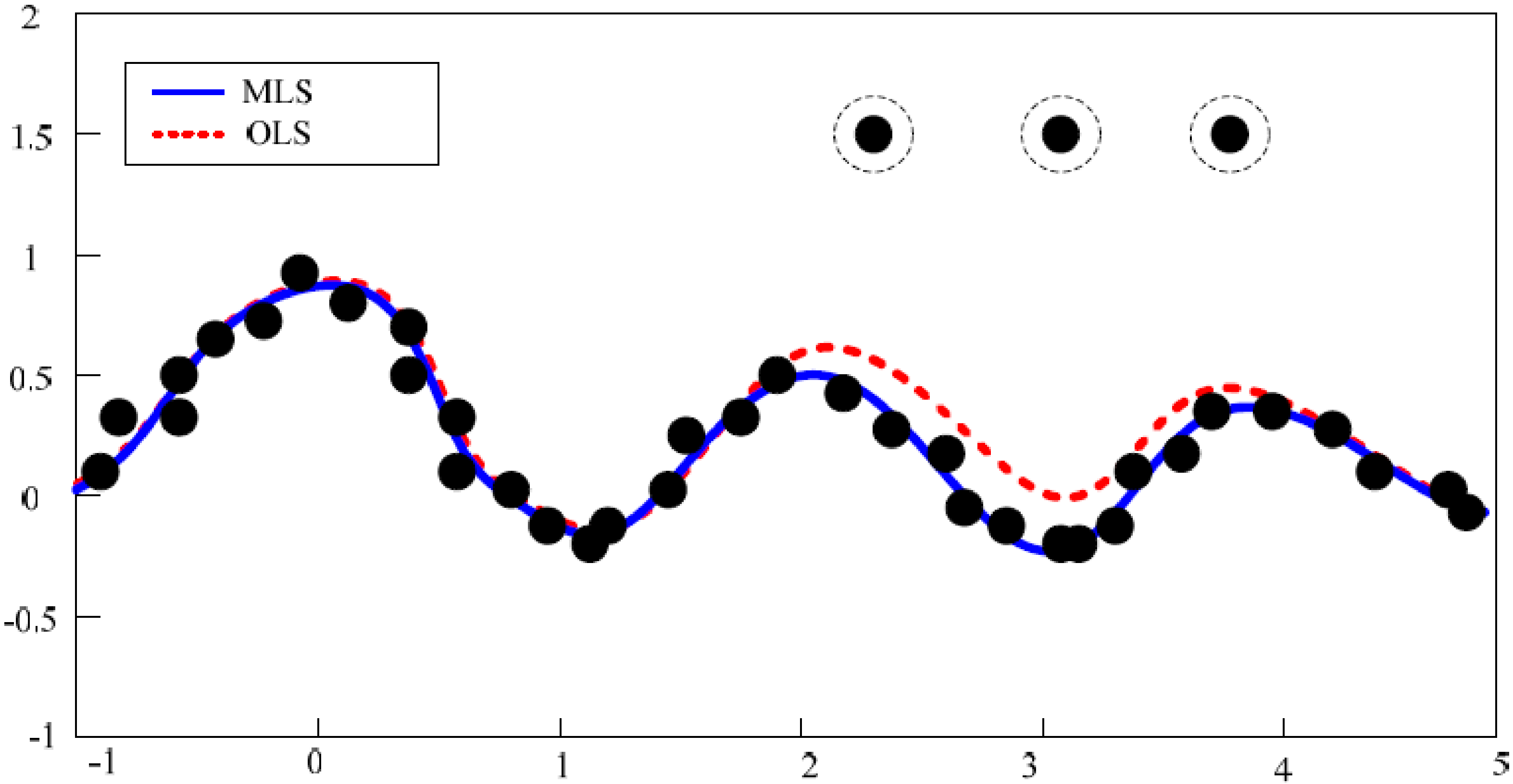
- 牛顿法在标准假设下具有局部二阶收敛，但上述问题一般*Hessian*矩阵中的$S(x)$难以计算或者花费工作量很大，但计算梯度时，Hessian矩阵$J(x)^T J(x)+S(x)$ 中的一阶信息项$J(x)^T J(x)$已知。
- 因此，为了简化计算，获得有效的算法，我们常忽略S(x)或者用一阶导数近似S(x)，但由$S(x)=\sum_{i=1}^{m} r_i(x)\nabla^2 r_i(x)$ 可知，只有$r_i(x)$接近于零或者接近线性函数，从而$\nabla^2 r_i(x)$接近于零时，$S(x)$才可以忽略。对于这类问题，通常称为小残量问题(Small Residual Problem)，否则，便称为大残量问题(Large Residual Problem)

- **给定一个训练集$S=\{(x_1,y_1),\cdots,(x_k,y_k)\}$,需要估计一个近似函数来模拟数据点的变化趋势，定义模型为$M(X,W)=W^T\phi(X)=\sum_{i=1}^{d}w_i\phi_i(X)$,$\phi$为基函数矢量**
- **普通**的最小二乘问题$f_{ols}(W)=\sum_{i=1}^{k}||w^T\phi(x_i)-y_i||^2$



- **移动最小二乘$f_{mls}(w_x)=\sum_{i=1}^{k}\theta(x,x_i)||\phi(x_i)-y_i||^2$**

◆移动最小二乘



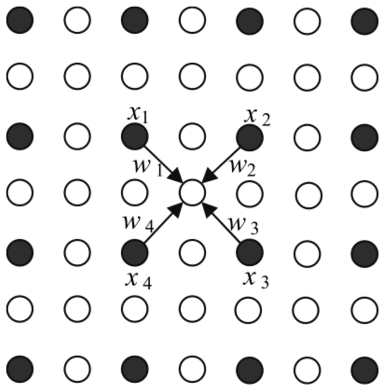
◆**扩展到图像插值，自然图像有着很强的局部结构性，这意味着自然图像中的像素点有着很强的局部依赖性。**

◆**根据这个特性，我们采用线性结合周围已知点的方式来估计当前未知点**

$$f_i(\mathbf{x}_i) = \langle \mathbf{a}_i, \phi(\mathbf{x}_i) \rangle + b_i$$

$$\Phi(\mathbf{x}_i)^T \leftarrow [\phi(\mathbf{x}_i)^T, 1] \qquad \mathbf{w}_i^T \leftarrow [\mathbf{a}_i^T, b_i]$$

$$f_i(\mathbf{x}_i) = \mathbf{w}_i^T \Phi(\mathbf{x}_i)$$



- 图像插值是一个病态问题
- 先验假设：自然图像可以建模成局部平滑变化的高斯过程，也就是说相邻的像素点有着相同或者相似的插值模型。
- 核岭回归，目标函数：

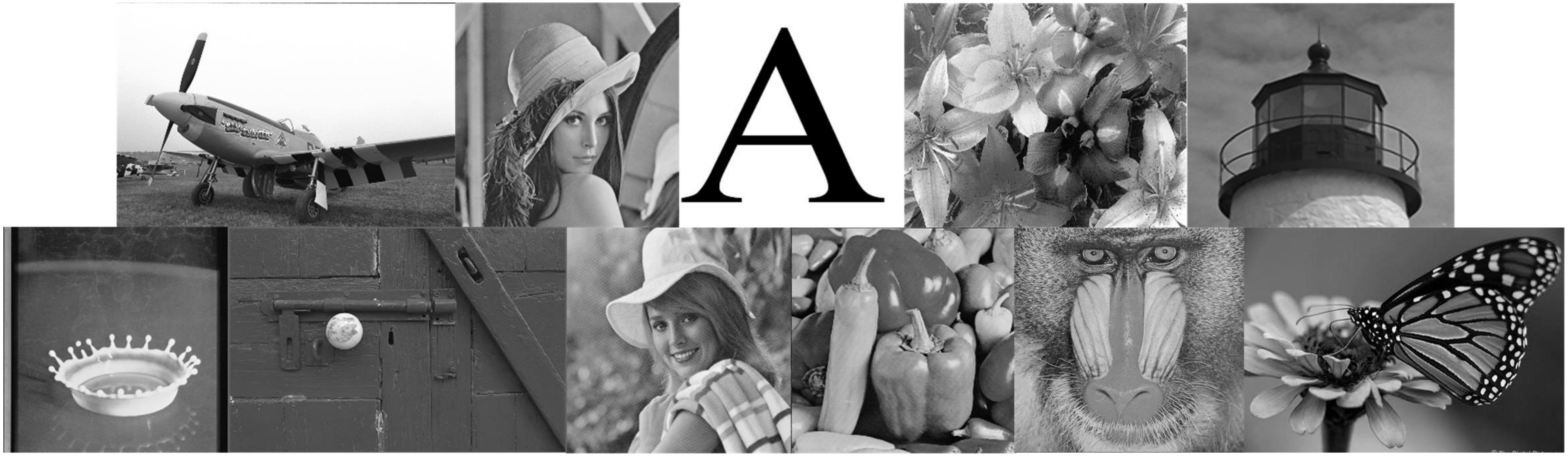
$$J(\mathbf{w}_i) = \sum_{j=1}^{l} \theta(\mathbf{x}_i, \mathbf{x}_j) \left\| y_i - \mathbf{w}_i^T \Phi(\mathbf{x}_j) \right\|^2 + \lambda \left\| \mathbf{w}_i \right\|^2$$

- 高分辨率图像中的临近点应该具有相似的像素值
- 将这样的先验知识通过一个额外地局部平滑保持约束引入到目标函数当中。给定局部邻域的已知点和未知点，我们定义如下的目标函数

$$J(\mathbf{w}_i) = \sum_{j=1}^{l} \theta(\mathbf{x}_i, \mathbf{x}_j) \| y_j - \mathbf{w}_i^T \Phi(\mathbf{x}_j) \|^2 + \lambda \| \mathbf{w}_i \|^2$$

$$+ \eta \sum_{m=l+1}^{l+u} \sum_{n=l+1}^{l+u} \theta(\mathbf{x}_m, \mathbf{x}_n) \| \mathbf{w}_m^T \Phi(\mathbf{x}_m) - \mathbf{w}_n^T \Phi(\mathbf{x}_n) \|^2$$

◆测试图像

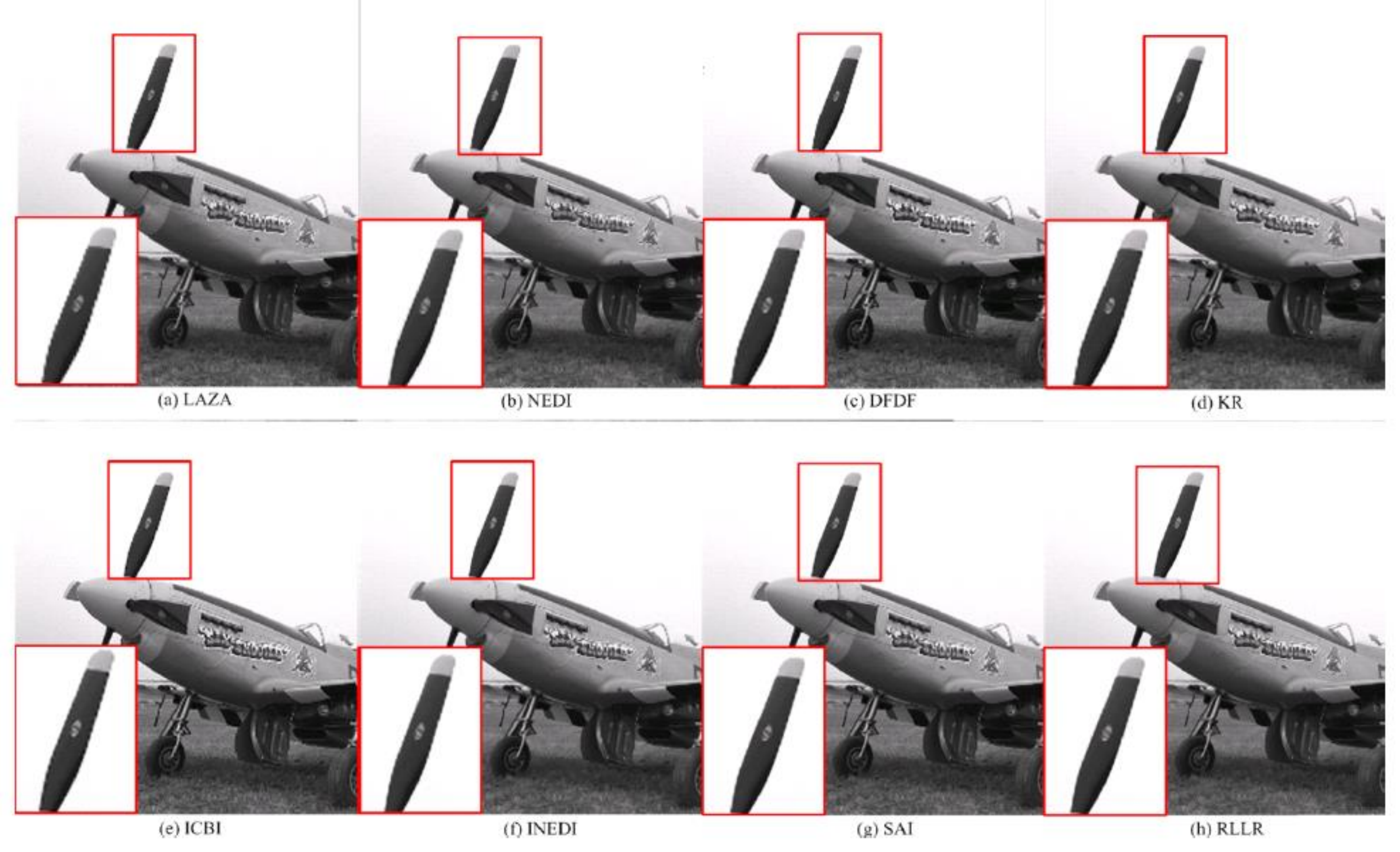


◆ $$PSNR = 10log\left(\frac{255}{MSE}\right)^2 = 20log\left(\frac{255}{MSE}\right)$$

Method	LAZA		NEDI		DFDF		KR		SAI		RLLR	
	PSNR	EPSNR	PSNR	EPSNR	PSNR	EPSNR	PSNR	EPSNR	PSNR	EPSNR	PSNR	EPSNR
Airplane	30.17	19.48	28.69	15.42	30.53	19.44	29.11	16.01	30.72	19.25	**30.97**	**19.86**
Lena	33.36	28.31	33.57	27.75	33.96	28.10	33.96	27.97	**34.72**	**28.97**	34.41	28.65
Flowers	25.65	20.37	25.62	19.94	25.74	20.40	25.79	20.23	25.96	**20.78**	**26.20**	20.75
Girl	31.84	29.50	31.84	28.30	31.81	29.41	31.92	29.10	31.77	29.60	**32.29**	**29.74**
Door	32.28	26.23	32.14	25.73	32.27	25.93	32.20	25.86	32.46	26.07	**32.60**	**26.39**
Peppers	31.51	22.42	29.30	17.83	31.87	22.53	31.02	20.53	31.84	22.08	**32.15**	**22.77**
Splash	33.54	19.74	31.38	15.49	33.79	19.79	33.38	19.07	33.54	19.14	**33.99**	19.87
Baboon	20.09	18.06	19.91	17.75	19.87	18.01	19.99	17.79	19.92	18.04	**20.67**	**18.77**
Tower	38.45	28.70	39.91	28.36	39.69	27.64	40.28	28.57	41.49	28.82	**41.97**	**29.43**
Butterfly	28.70	21.10	28.96	21.02	29.67	20.76	29.54	21.25	29.93	20.46	**30.32**	**21.56**
Letter	26.07	13.53	26.47	13.59	27.21	14.22	27.02	13.13	27.23	13.98	27.43	14.75
Average	30.15	22.49	29.79	21.02	30.58	22.38	30.38	21.78	30.88	22.47	**31.18**	**22.96**

第7章约束问题的最优化方法(Methods of Constrained Optimization)-非线性最小二乘

◆客观图像质量对比



- **高斯牛顿法(Gauss-Newton)**
  - 上述问题如果忽略$S(x)$,则迭代公式成为$x_{k+1}=x_k-\left(J(x)^TJ(x)\right)^{-1}J(x_k)r(x_k)$，此时即为Gauss-Newton法，其第$k$步如下
  - 1.解$J(x)^TJ(x)s=-J(x_k)r(x_k)$得到$s_k$;
  - 2.令$x_{k+1}=x_k+s_k$
- **实际上，此时相当于考虑$r(x)$在x_k附近的仿射模型$\bar{M}_k(x)=r(x_k)+J(x_k)(x-x_k)$,从而求线性最小二乘问题$Min\frac{1}{2}\|\bar{M}_k(x)\|^2$的解，从上述迭代解的过程可知，只需要残量的一阶导数信息，并且$J(x)^TJ(x)$至少为半正定**
- **但Gauss-Newton方法的收敛速度与忽略项$S(x)$在Hessian矩阵$G(x)$中的重要性有关，若$S(x^*)=0$,则二阶收敛，若相对于$J(x)^TJ(x)$是小的，则Q-线性收敛，否则可能不收敛**
- **缺点：残量越大，算法收敛越慢并有可能不收敛，$J(x_k)$不满秩则没有定义，不一定总体收敛**
- 但无论如何，高斯牛顿法是解非线性最小二乘的最基本的方法，如果2.令$x_{k+1}=x_k+s_k$中采用线性搜索来确定迭代公式，则称为阻尼牛顿法(Damped Gauss-Newton)
- 另外，求解高斯牛顿方程时，一般采用矩阵分解，如QR分解来求解$J(x)^TJ(x)s=-J(x_k)r(x_k)$
- 注意：上述方法要求$J(x^*)$是满秩的，但经常发生$J(x^*)$是奇异的，使得算法常常收敛到一个非驻点。一旦$J(x^*)$是奇异的，则在距离解点的某处，s_k,g_k便在数值上直交，这时线搜索便得不到进一步的下降，只能得到极小点的一个差的估计，为了克服这些困难，考虑采用信赖域或信任域方法

第7章约束问题的最优化方法(Methods of Constrained Optimization)-非线性最小二乘

- 通常$r(x)$是非线性函数，而Gauss-Newton方法用线性化模型近似$r(x)$，这种线性化并不是对所有$(x-x_k)$都成立，因此考虑约束线性最小二乘问题，即考虑信赖域模型：$min\ \|r(x_k)+J(x_k)(x-x_k)\|_2, s.t. \|x-x_k\|_2 \le h_k$
- 而这个模型的解可由解方程组$\left(J(x_k)^T J(x_k)+\mu_k I\right)s=-J(x_k)^T r(x_k)$来表征,从而$x_{k+1}=x_k-\left(J(x_k)^T J(x_k)+\mu_k I\right)^{-1} J(x_k)^T r(x_k)$,如果$\|(J(x_k)^T J(x_k))^{-1} J(x_k) r(x_k)\| \le h_k$,则$\mu_k=0$;否则$\mu_k>0$,由$\left(J(x_k)^T J(x_k)+\mu_k I\right)$正定，故产生的方向为下降方向，该方法称为Levenberg-Marqurdt方法
- 这是一类方法，有很多的不同实现算法，例如分别采用h_k, μ_k来控制迭代
- 但对于大残量问题，上述两类方法都可能收敛速度慢，因为都没有利用Hessian矩阵的信息，实际上却是难以计算，通过构造$S(x)$的割线近似，从而形成拟牛顿法
- 假设B_k是S(x_k)的割线近似，则牛顿法的迭代公式为:$x_{k+1}=x_k-\left(J(x)^T J(x)+S(x)\right)^{-1} J(x_k) r(x_k)$成为$\left(J(x)^T J(x)+B_k\right)d_k=-J(x_k)^T r(x_k)$,要求$B_k$满足某种拟牛顿条件。由于$S(x_{k+1})=\sum_{i=1}^m r_i(x_{k+1})\nabla^2 r_i(x_{k+1})$,从而用$\sum_{i=1}^m r_i(x_{k+1})(H_i)_{k+1}$来近似$S(x_{k+1})$，这里$(H_i)_{k+1}$是$\nabla^2 r_i(x_{k+1})$的拟牛顿近似，故有
- $(H_i)_{k+1}(x_{k+1}-x_k)=\nabla r_i(x_{k+1})-r_i(x_k)$,于是$B_{k+1}(x_{k+1}-x_k)=\sum_{i=1}^m r_i(x_{k+1})(H_i)_{k+1}(x_{k+1}-x_k)=\sum_{i=1}^m r_i(x_{k+1})(\nabla r_i(x_{k+1})-r_i(x_k))=\left(J(x_{k+1})-J(x_k)\right)^T r(x_{k+1}) \stackrel{\text{def}}{=} y_k$
- 这就是B_k满足的拟牛顿条件
- Biggs给出如下的拟牛顿条件下的秩一校正公式

$$B_{k+1}=B_k+\frac{(y_k-B_k s_k)(y_k-B_k s_k)^T}{(y_k-B_k s_k)^T s_k}$$

第7章约束问题的最优化方法(Methods of Constrained Optimization)-最优性条件

◆无约束问题的最优性极值

- 一阶必要条件(⇒):$f(x)\in R^1, x\in R^n$,在点x^*处可微，若该点为$\min f(x)$的局部最优解,则$\nabla f(x^*)=0$
- 二阶必要条件(⇒):$f(x)\in R^1, x\in R^n$,在点x^*处二次可微，若该点为$\min f(x)$的局部最优解,则$\nabla f(x^*)=0$，且$\nabla^2 f(x^*)$半正定
- 充分条件(⇐):$f(x)\in R^1, x\in R^n$,在点x^*处二次可微，且满足$\nabla f(x^*)=0$，且$\nabla^2 f(x^*)$正定，则该点为$\min f(x)$的严格局部最优解
- 充分条件(⇐):$f(x)\in R^1, x\in R^n$,在点x^*处的一个δ邻域$N(x^*,\delta)$内二次可微，满足$\nabla f(x^*)=0$，且对$\forall x\in N(x^*,\delta)$,$\nabla^2 f(x)$半正定，则该点为$\min f(x)$的局部最优解

◆无约束问题的最优性极值

- 对于正定二次函数$f(x)=\frac{1}{2}x^TAx+b^Tx+C$（$A$为$n$阶对称正定阵）有唯一极小点:$x^*=-A^{-1}b$
- 优势
- 正定二次函数的极小化简单
- 实际问题中如果目标函数在极小点处的Hessian矩阵正定，则局域内目标函数可近似为一个正定二次函数

◆正常求解

- 求出平稳点，计算平稳点的Hessian矩阵，判定其正定性
  - ✓正定，说明为严格局部最优解
  - ✓半正定，在判断其局部邻域内任意点的半正定性，若半正定，则为局部最优解

- **有约束问题的最优性极值**

  (问题P): $Min\ f(x), s.t.\ g(x) \le 0, h(x) = 0$

- **可行方向、下降方向、可行下降方向**

- **起作用约束**

- 一阶最优性条件，问题P：$Min\ f(x), s.t.\ g(x) \le 0, h(x) = 0$,要求起作用约束集$I$的导数线性无关即$\nabla h_i(x^*)$， $\nabla g_i(x^*)$,$i \in I$,线性无关，由于涉及导数，要求$x^*, f, h_i, g_i, i \in I$可微，$h_i, g_i, i \notin I$连续，则分别针对只有等式约束，和不等式约束的情况下，有如下定理
- 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 $\nabla f(x^*) + {\lambda^*}^T \nabla h(x^*) = 0$。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
- 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

◆(接上页)设x^*是问题P的一个局部极小点，则必存在$\lambda_i^*, \mu_i^* (i = 1, 2, \cdots, m + l)$, 使得

$$\nabla f(x^*) + \sum_{i=1}^{l} \lambda_i^* \nabla h_i(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) = 0 \quad (6-6);$$

$$\mu_i^* \geq 0, \mu_i^* g_i(x^*) = 0 \quad (6\text{-}7)$$

成立

◆注意

- 等式约束前的拉格朗日乘子没有正负约束，因此前面的±号没有实际意义，但不等式约束前的±号对应不等式约束的方向
- 上述条件针对的问题约束是$g(x) \leq 0$,因此，如果约束变为$g(x) \geq 0$,则此时有$-g(x) \leq 0$,对应$\nabla g_i(x^*)$前的+号变为减号，其它情况不变
- 若为凸规划，则成为充要条件：只需要证明$\nabla f^T(x^*)(x - x^*) \geq 0$

- 问题1：$min\ f(x) = x_1 + x_2$,
- $s.t.\ \begin{matrix} h_1(x) = (x_1-1)^2 + x_2^2 - 1 = 0 \\ h_2(x) = (x_1-2)^2 + x_2^2 - 4 = 0 \end{matrix}$，最优解为 $(0,0)$
- 请验证是否满足有约束优化问题的拉格朗日函数的一阶必要条件？为什么？
- $\{\nabla g_i(x^*) | i \in I\}$线性无关
- 一阶可行方向形成的子空间必须满足:$V(x^*) = \{x^* - x | \nabla h_i(x^*)^T (x^* - x) = 0, i = 1, \cdots, m\}$
- 即：$\nabla f(x^*)^T (x^* - x) = 0$,与无约束的零梯度条件类似

思考:$V(x^*)$与真实可行方向集的关系

- **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：

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

- 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$d^T \nabla_x^2 L(x^*, u, v) d \ge 0$

第7章约束问题的最优化方法(Methods of Constrained Optimization)-最优性条件-二阶必要条件例子

- $\min f(x)=\frac{1}{2}\left(x_1^2+x_2^2+x_3^2\right), s.t.\, h_1(x)=x_1+x_2+x_3=3$
- **验证是否满足二阶必要条件?**
- $\min f(x)=-\frac{1}{2}\left(x_1^2+x_2^2+x_3^2\right), s.t.\, h_1(x)=x_1+x_2+x_3=3$
- **验证是否满足二阶必要条件?**

◆二阶充分条件

◆问题$min\ f(x), s.t.\ g(x) \leq 0, h(x) = 0$,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$d \in S_h \cap \{d|\nabla f^T(x^*)d = 0, \nabla g_i^T(x^*)d \leq 0, i \in I\}$,且$d \neq 0$, 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：$min\ x_1 + x_2, s.t.\ 2 - x_1^2 - x_2^2 \geq 0$

◆其最优解为$x^* = (-1, -1)^T$, 乘子为$\lambda_1^* = \frac{1}{2}$,计算其拉格朗日函数的二阶导数$\nabla_x^2 L(x^*, \lambda^*) = \begin{pmatrix} 2\lambda_1^* & 0 \\ 0 & 2\lambda_1^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$,正定，因此KKT点就是其严格局部最优解。

◆ **例子**

$$\min f(x) = x_1^2 + x_2^2 - x_1x_2 - 2x_1 - 3x_2$$

$$s.t. \begin{cases} x_1 + x_2 \le 2 \\ x_1 + 5x_2 \le 5 \\ x_1, x_2 \ge 0 \end{cases}$$

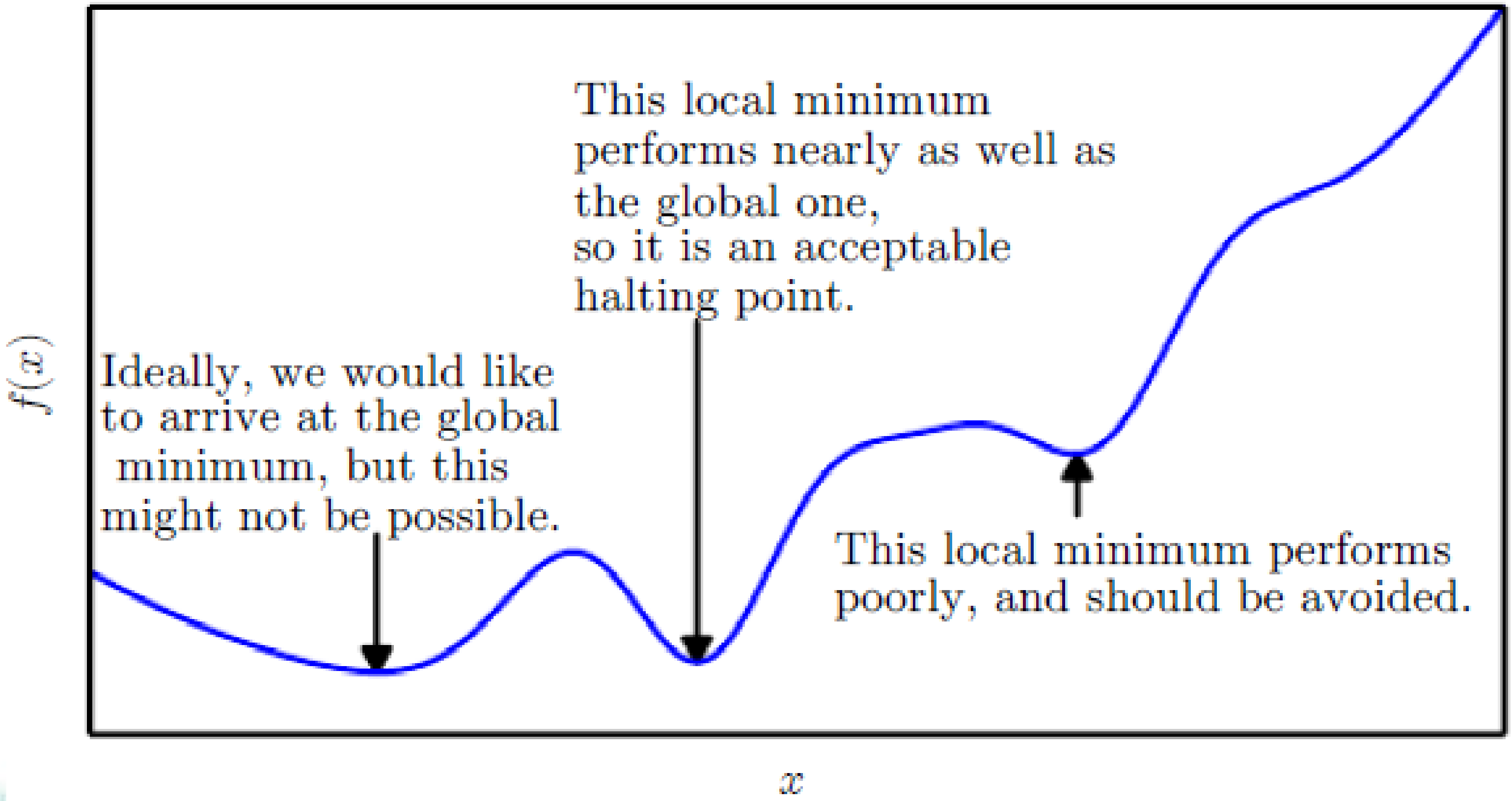
◆ **试验证$\left(\frac{35}{31}, \frac{24}{31}\right)^T$是否是最优点，为什么？**

◆ **例子：$\min\ -0.1(x_1-4)^2 + x_2^2, s.t.\ x_1^2 + x_2^2 - 1 \ge 0$,验证$\lambda = 0.3, x^* = (1,0)^T$是否是最优点？为什么？**

➢ 解：二阶充分条件中的临界方向d满足：$(0, w_2)^T$的形式，其中$w_2 \ne 0$,因此验证：$w^T\nabla_{xx}^2\mathcal{L}(x^*, \lambda^*)w > 0$即可!

◆ **请用拉格朗日对偶理论，求解问题P: $\min\ -c^Tx, s.t. \begin{cases} Ax = b, c \in R^n, A_{m\times n}, b \in R^m \\ x \ge 0 \end{cases}$的对偶问题。**

◆优化问题不一定非要全局最优点



◆是否有约束，是否可微，是否二阶可微…

第7章约束问题的最优化方法(Methods of Constrained Optimization)-如何求解

- ◆ **一般原则**
  - ➢ 无约束要比有约束简单
  - ➢ 可微要比不可微简单
- ◆ **可微，无约束**
  - ➢ 一阶、二阶导数(**Hessian**)矩阵
- ◆ **可微，有约束**
  - ➢ 拉格朗日函数的一阶、二阶导数
- ◆ **不可微，无约束**
  - ➢ 次梯度...
- ◆ **不可微，有约束**
  - ➢ 投影梯度...

第7章约束问题的最优化方法(Methods of Constrained Optimization)-如何求解

- **无约束:导数信息，二阶导数信息(曲率)，稳定点/临界点:**
  - 一阶、二阶条件
  - 梯度下降法
  - 牛顿法
  - 练习：用梯度下降法求解$\min_x f(x) = \frac{1}{2}||Ax - b||_2^2$
- **约束**
  - 拉格朗日函数，KKT条件
  - 考虑约束，修改梯度下降法
  - 逼近法，设计一个无约束最优化问题，其解转化为原问题的解

✓例如$min\ f(x), x \in R^2, ||x||_2^2 = 1 \Rightarrow min_\theta\ f((cos\theta, sin\theta)^T)$

◆如何求解 $min_x f(x), x^T x \leq 1$?

$$L(x,\lambda) = f(x) + \lambda(x^T x - 1)$$

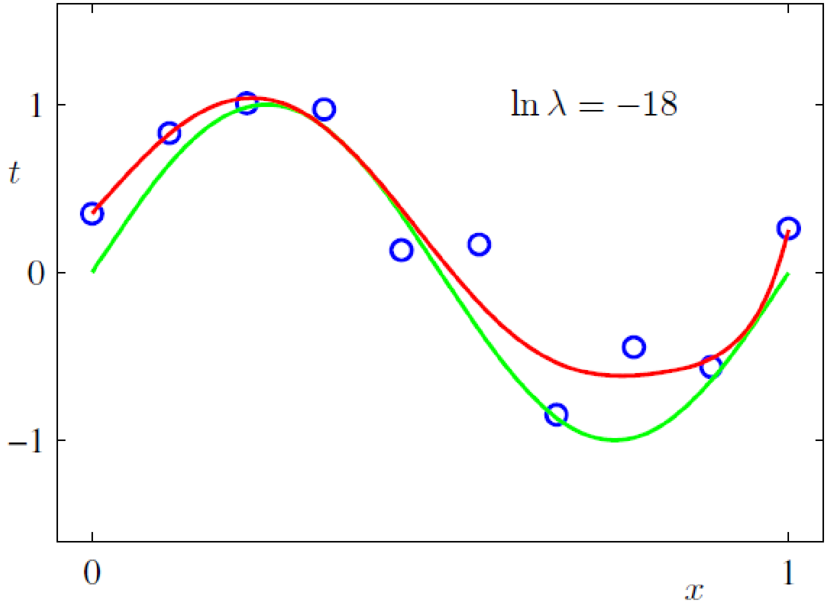
$\nabla L_x(x,\lambda) = A^T A x - A^T b + 2\lambda x = 0 \Rightarrow x = \left(A^T A + 2\lambda I\right)^{-1} A^T b,$

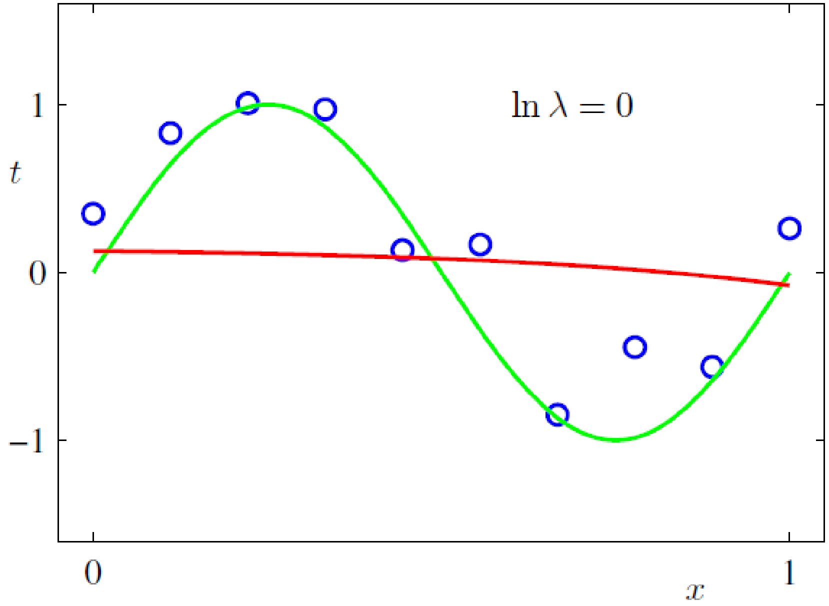
◆λ的幅度必须使得结果满足约束，若对其执行梯度增加可以找到这个值

◆$\nabla L_\lambda(x,\lambda) = x^T x - 1$，显然当$||x|| > 1$, 偏导数为正，所以提升$\lambda$,则增加拉格朗日函数值，其偏导数遵循上升趋势

◆约束作为正则化的手段，经常在机器学习、图像处理中广泛使用

- ◆拟合：$\widetilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}||w||_2^2$





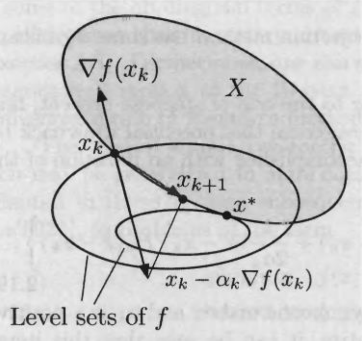
- ◆问题：正则化的目的是什么？
- ◆大量样本如何使用梯度下降算法？Stochastic gradient descent(SGD)

- **大多数优化方法都与精确的梯度和Hessian矩阵的值有很大关系**
  - 实际使用中，通常采用近似估计量，例如几乎所有的 **Deep learning** 算法使用基于样本的估计值
  - 甚至目标函数很难处理，导致其梯度也很难处理，采用一些逼近的办法来处理
- **稳定点，KKT点，然后采用充分条件来判断，可以借助一阶和二阶的条件**
- **难求？**
  - 与无约束中类似，也需要采用迭代的方式求解
  - 迭代下降：$\{x_k\} \in X, \phi(x_{k+1}) < \phi(x_k)$
  - 逼近思想：$\{x_k\}, x_{k+1} \in argmin_{x \in X_k} F_k(x)$

将一个约束优化问题转化为一系列无约束优化问题

第7章约束问题的最优化方法(Methods of Constrained Optimization)-如何求解

- **迭代下降：** $\{x_k\} \in X, \phi(x_{k+1}) < \phi(x_k)$
- **逼近思想：** $\{x_k\}, x_{k+1} \in argmin_{x \in X_k} F_k(x)$
- **目标函数和约束函数的可微性质有什么作用？**
  - 连续性
    - ✓无约束：梯度类方法，最速下降，共轭梯度，牛顿法
    - ✓有约束：可行方向法，如条件梯度，梯度投影
    - ✓$x_{k+1} = P_X(x_k - \alpha_k D_k \nabla f(x_k)), D_k$为修正矩阵，
    - ✓块坐标下降法

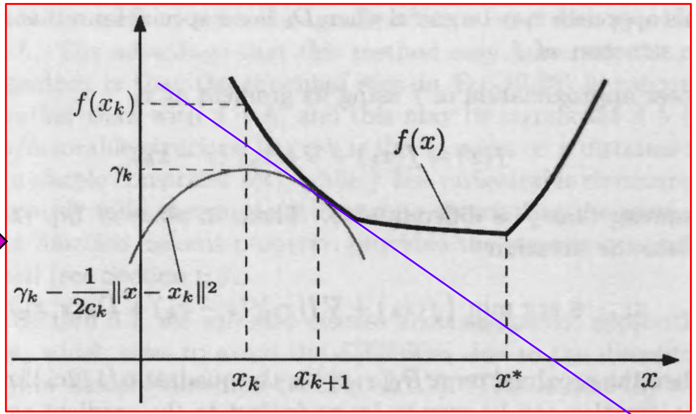


  - 不可微转化为可微
    - ✓$min\, max\{f_1(x), \cdots, f_m(x)\}, s.t.\, x \in R^n$,函数均可微，可转化为$min\, z\,; s.t.\, f_j(x) \le z, j = 1, 2, \cdots, m$,直接用罚函数或增广拉格朗日求解
    - ✓不可微：次梯度法
    - ✓交替下降法：$x_{k+1} \in argmin_{x \in R^n} \left\{ f(x) + \frac{1}{2c_k} ||x - x_k||^2 \right\}$

◆ **上面的交替下降法，有时也称为邻近算法，常用于代价函数是凸函数当中，体现了代价改进和逼近的思想,注意与增广拉格朗日方法的对偶性：**

◆ $\min f(x) + \lambda_k(Ax-b) + \frac{c_k}{2}||Ax-b||^2$

$x_{k+1} \in argmin_{x\in R^n}\left\{f(x) + \frac{1}{2c_k}||x-x_k||^2\right\}$ ⟺



斜率=$\frac{x_k - x_{k+1}}{c_k}$

◆ **与此类似的一种方法称为：Tikhonov正则化方法**

◆ $x_{k+1} \in argmin_{x\in R^n}\left\{f(x) + \frac{1}{2c_k}||x||^2\right\}$

◆ **与交替下降法的区别就是x_k不直接进入最小化当中来确定x_{k+1},因此方法收敛依赖于$c_k \to \infty$，但前面的交替下降法就没有这个要求，甚至可以令$c_k = cons$. Tikhonov正则化方法与$\min f(x) + \frac{c_k}{2}||Ax-b||^2$形成对偶关系**

◆ **邻近算法(Proximal algorithm)**

- 当采用合适的正则项的时候，则有ADMM方法(Alternating direction method of multipliers)，类似增广拉格朗日方法，又正好适合某些具有特殊结构的问题（可分离并且有很多成份函数求和的问题，例如正则化回归，分类和极大似然问题中）

$$\min f_1(x)+f_2(Ax), s.t.\ x\in R^n, A: m\times n\text{矩阵，} f_i\text{都为凸函数}$$

- 这可等价转化为：

$$\min f_1(x)+f_2(z), s.t.\ x\in R^n, z\in R^m, Ax=z$$

- 引入增广拉格朗日函数：
- $\mathcal{L}_c(x,z,\lambda)=f_1(x)+f_2(z)+\lambda'(Ax-z)+\frac{c}{2}||Ax-z||^2$，$c$为正的参数

ADMM方法：

- $x_{k+1}\in argmin_{x\in R^n}\mathcal{L}_c(x,z_k,\lambda_k)$
- $z_{k+1}\in argmin_{z\in R^m}\mathcal{L}_c(x_{k+1},z,\lambda_k)$
- $\lambda_{k+1}=\lambda_k+c(Ax_{k+1}-z_{k+1})$
- 这种分离变量方式可以联合最小化x,z，（否则多变量的耦合导致其复杂性增加！）而增广拉格朗日方法则不行，但收敛快！

◆ **注意：ALM(augmented Lagrangian method)c_k增加会加快收敛，但ADMM方法中则没有这样的性质，但收敛会慢一些。但是，分离性导致可以并行和分布式处理！**

◆可行方向法

- 问题P：$min\ f(x), g(x) \le 0$,设其可行域为X,给定可行点x_k,为求其极小点，则应在点x_k处的可行下降方向中选取方向d_k,然后采用线搜索求步长，产生新的迭代点
- $\begin{cases} x_{k+1} = x_k + \lambda_k d_k \\ f(x_{k+1}) < f(x_k) \end{cases}$,然后判断新点是否满足精度要求，满足，则停止；不满足，则继续迭代或者迭代次数超过预定次数也停止

◆约束为线性函数的情况

- ➢$\min f(x); s.t. Ax \le b; Ex = e; f$可微，$A_{m\times n}, E_{l\times n}, x \in R^n, b \in R^m, e \in R^l$ (7-3)

◆**定理：设$\overline{x}$是上述问题的一个可行解，且在点$\overline{x}$处有$A_1\overline{x} = b_1, A_2\overline{x} < b_2$,其中$A = [A_1\ A_2]^T, b = [b_1\ b_2]^T$则**

- ➢(1)向量$d(d \in R^n, d \neq 0)$是点$\overline{x}$处的可行方向的充要条件为$A_1 d \le 0, Ed = 0$;
- ➢(2)若此时d又满足:$\nabla f(\overline{x})^T d < 0$,则$d$是一个可行下降方向

◆**这条定理表明，在可行点处找一个下降方向，等价于求解一个线性规划问题**

◆**定理：设$\overline{x}$是上述问题的一个可行解，且在点$\overline{x}$处有$A_1\overline{x} = b_1, A_2\overline{x} < b_2$,其中$A = [A_1\ A_2]^T, b = [b_1\ b_2]^T$则**

- (1)向量$d(d \in R^n, d \neq 0)$是点$\overline{x}$处的可行方向的充要条件为$A_1 d \leq 0, Ed = 0$;
- (2)若此时d又满足:$\nabla f(\overline{x})^T d < 0$,则$d$是一个可行下降方向

◆**证明:(1)=>由定义可行方向知存在$\delta > 0, \forall\lambda \in (0,\delta), (\overline{x} + \lambda d) \in X$,从而$A(\overline{x} + \lambda d) \leq b, E(\overline{x} + \lambda d) = e \Rightarrow b_1 + \lambda A_1 d \leq b_1, b_2 + \lambda A_2 d \leq b_2, Ed = 0$**

◆**$\Leftarrow A_1(\overline{x} + \lambda d) = b_1 + \lambda A_1 d \leq b_1, E(\overline{x} + \lambda d) = e$,且$A_2\overline{x} < b_2$，因而必存在一个$\delta > 0, \forall\lambda \in (0,\delta), A_2(\overline{x} + \lambda d) < b_2$,从而得证**

◆**(2)显然.**

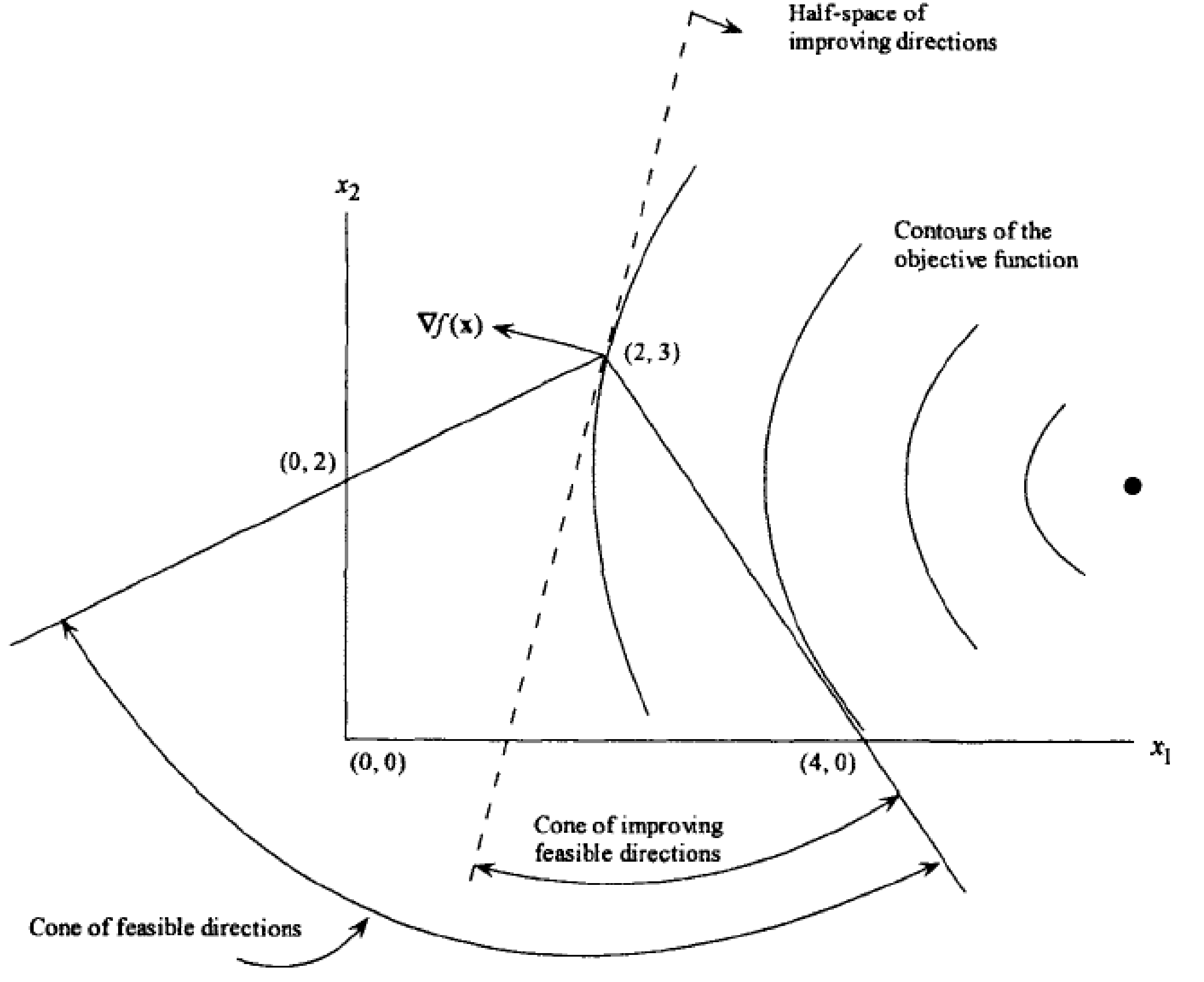
◆可行方向法的几何解释

◆$\min\ (x_1-6)^2+(x_2-2)^2, s.t.\begin{cases}-x_1+2x_2\le 4\\3x_1+2x_2\le 12\\-x_1\le 0\\-x_2\le 0\end{cases}$

◆令$x=(2,3)^T$,$A_1=\begin{pmatrix}-1 & 2\\3 & 2\end{pmatrix}$,由前面的定理，$d$在$x$处为可行方向$\Leftrightarrow A_1d\le 0$

◆即$\begin{cases}-d_1+2d_2\le 0\\3d_1+2d_2\le 0\end{cases}$

◆可行方向法的几何解释


Half-space of
improving directions
x_2
Contours of the
objective function
$\nabla f(\mathbf{x})$
(2, 3)
(0, 2)
(0, 0)
(4, 0)
x_1
Cone of improving
feasible directions
Cone of feasible directions

- $(7-4)\, Min\, z = \nabla f(\bar{x})^T d$, s.t. $\begin{cases} A_1 d \le 0 \\ Ed = 0 \\ -1 \le d_j \le 1, j = 1, \cdots, n \end{cases}$
- 最后一个约束主要是确保方向是有界方向。从这可知$d = 0$是可行解，因此目标函数的解必小于或等于0，若存在小于0的解，表明存在可行下降方向；若等于0，则表明$\bar{x}$即是KT点
- 定理：设$\bar{x}$是线性约束最优化问题的一个可行解，且在点$\bar{x}$处有$A_1\bar{x} = b_1, A_2\bar{x} < b_2$,其中$A = [A_1\ A_2]^T, b = [b_1\ b_2]^T$则$\bar{x}$是KT点的充要条件就是$(7-4)$式的最优目标函数值为0.

◆上述的两条定理表明， 求解$(7-4)$式的结果或是可行下降方向，或是KT(库恩-塔克)点。

◆$(7-5)\, Min\, f(x); s.t.\, g(x) \le 0$,其中$x \in R^n, f(x), g(x)$均为可微函数，如何求解其可行下降方向呢？

◆**定理：**设$\bar{x}$是问题$(7-5)$的一个可行解，$I = \{i | g_i(\bar{x}) = 0\}$，若$f(x), g_i(x)(i \in I)$在点$\bar{x}$处可微，$g_i(x)(i \notin I)$在点$\bar{x}$处连续，如果$\nabla f(\bar{x})^T d < 0, \nabla g_i(\bar{x})^T d < 0, i \in I$,则方向$d$是可行下降方向。

◆**证明：** $\nabla f(\bar{x})^T d < 0$表明方向$d$是下降方向，$\nabla g_i(\bar{x})^T d < 0$表明方向是可行方向(因为$g_i(\bar{x} + \lambda d) \le 0, i \in I$,泰勒展开即可)

- 因此，(7−5)在可行解$\bar{x}$处的可行方向d必然满足$\begin{cases}\nabla f(\bar{x})^T d < 0, \\ \nabla g_i(\bar{x})^T d < 0, i \in I\end{cases}$,如果右边的限制用数$\eta$替换，则等价于下述方程组求向量$d$,和$\eta$ $\begin{cases}\nabla f(\bar{x})^T d \le \eta, \\ \nabla g_i(\bar{x})^T d \le \eta, i \in I, \\ \eta < 0\end{cases}$
  显然，这个方程组存在很多可能的(d, η),实际求解时，希望能求出使目标函数值下降最多的方向d,因此，可进一步转化为对η求极小值的一个线性规划问题：
- (7−6)$Min\ \eta; s.t. \begin{cases}\nabla f(\bar{x})^T d \le \eta, \\ \nabla g_i(\bar{x})^T d \le \eta, i \in I \\ -1 \le d_j \le 1, j = 1, 2, \cdots, n\end{cases}$,求解后的$(d^*, \eta^*)$,$\eta^*$必然$\le 0$,若$\eta^* < 0$,则$d^*$为可行下降方向；若$\eta^* = 0$,则在一定条件下$\bar{x}$为KT点

◆可行方向法的计算步骤

◆迭代: $x_{k+1}=x_k+\lambda_k d_k$, 获得方向$d_k$,为使新的迭代点可行，并且新的函数值$f(x_{k+1})$下降，求解:$\min_{0\le\lambda\le\bar{\lambda}} f(x_k+\lambda d_k)$, $\bar{\lambda}=\max\{\lambda \mid x_k+\lambda d_k\in X\}$

◆ 约束为线性:$\min f(x); s.t.\ Ax\le b; Ex=e$; f可微，$A_{m\times n}, E_{l\times n}, x\in R^n, b\in R^m, e\in R^l$, 通过(7 − 4)$Min\ z=\nabla f(\bar{x})^T d$, s.t. $\begin{cases} A_1 d\le 0 \\ Ed=0 \\ -1\le d_j\le 1, j=1,\cdots,n \end{cases}$ 求得方向

◆约束为非线性:(7 − 5)$Min\ f(x); s.t.\ g(x)\le 0$，通过求解线性规划(7 − 6)$Min\ \eta; s.t. \begin{cases} \nabla f(\bar{x})^T d\le\eta, \\ \nabla g_i(\bar{x})^T d\le\eta, i\in I \\ -1\le d_j\le 1, j=1,2,\cdots,n \end{cases}$ 得到x_k及可行下降方向d_k,

- 约束为线性:$min\ f(x); s.t. Ax \le b; Ex = e$; f可微，$A_{m\times n}, E_{l\times n}, x \in R^n, b \in R^m, e \in R^l$
- 先求解$min\ f(x_k + \lambda d_k), s.t. \begin{cases} A(x_k+\lambda d_k) \le b \\ E(x_k+\lambda d_k) = e \end{cases}$(7-7),考虑到$d_k$为可行下降方向 $\Rightarrow Ed_k = 0$，x_k是可行点$\Rightarrow Ex_k = e$,因此(7-7)中只需考虑第一个约束，并将$A = [A_1 A_2]^T$,其中A_1为起作用约束，此时第一个约束为$A_1(x_k+\lambda d_k) \le b_1$, $A_2(x_k+\lambda d_k) \le b_2$, d_k为可行下降方向$\Rightarrow A_1 d_k \le 0$,又设$A_1 x_k = b_1, A_2 x_k < b_2, \lambda \ge 0 \Rightarrow A_1(x_k+\lambda d_k) \le b_1$,此时只剩第2个约束$A_2(x_k+\lambda d_k) \le b_2$，因此，

$$min\ f(x_k+\lambda d_k), s.t. \begin{cases} A(x_k+\lambda d_k) \le b \\ E(x_k+\lambda d_k) = e \end{cases} \Leftrightarrow min\ f(x_k+\lambda d_k), s.t. A_2(x_k+\lambda d_k) \le b_2, \lambda \ge 0 \quad (7-8)$$

- 下面推导$\bar{\lambda}$. $A_2(x_k+\lambda d_k) \le b_2 \to \lambda A_2 d_k \le b_2 - A_2 x_k$,分别记$\hat{d} = A_2 d_k$,$\hat{b} = b_2 - A_2 x_k$,则约束化为$\lambda \hat{d} \le \hat{b}, \lambda \ge 0$,且$\hat{b} < 0$,由此可得$\lambda$的上界:$\bar{\lambda} = \begin{cases} min\left\{\frac{\hat{b}_i}{\hat{d}_i} | \hat{d}_i > 0\right\}, 若\hat{d}不是\le 0 \\ \infty, 若\hat{d} \le 0 \end{cases}$(7-9)，因此(7-7)化简为$min\ f(x_k+\lambda d_k), s.t. 0 \le \lambda \le \bar{\lambda}$,

- 因此，可行方向法求解等式约束的非线性规划问题的步骤如下
- 1.给定初始可行点$x_0 \in X$,允许误差$\epsilon_1 > 0, \epsilon_2 > 0, k = 0$;
- 2.在点x_k处，把A, b分解为$A = [A_1\ A_2]$^T,$b = [b_1\ b_2]^T$,使得A_1为起作用约束即$A_1 x_k = b_1, A_2 x_k < b_2$;
- 3.判断x_k是否是问题(7-3) $min\ f(x); s.t. Ax \le b; Ex = e$的可行域的内点
  - 3.1 若x_k是可行域的内点(表明此时问题(7-3)没有起作用约束，即$E = 0, A_1 = 0$),而且$\|\nabla f(x_k)\| < \epsilon_1$,则停止迭代，得到近似极小点$x_k$;
  - 3.2若x_k是可行域的内点,且$\|\nabla f(x_k)\| \ge \epsilon_1$,则取搜索方向$d_k = -\nabla f(x_k)$,Goto 6,即用目标函数得负梯度方向作为搜索方向再求步长，此时类似无约束问题
  - 3.3若x_k是可行域的内点,即在可行域的边界上，则要寻找可行下降方向,Goto4

◆ 4.求解线性规划问题(7-4) $Min\ z = \nabla f(\overline{x})^T d$, s.t.

$$\begin{cases} A_1 d \le 0 \\ Ed = 0 \\ -1 \le d_j \le 1, j = 1, \cdots, n \end{cases}$$,设求得的解为(d_k, z_k);

◆ 5.判断精度，若$|z_k| = \left|\nabla f(\overline{x_k})^T d_k\right| < \epsilon_2$,则停止迭代，输出$x_k$;否则，以$d_k$为搜索方向，Goto 6;

◆ 6. 作一维搜索，先根据(7-9) $\hat{d} = A_2 d_k$,$\hat{b} = b_2 - A_2 x_k$, $:\bar{\lambda} =$

$$\begin{cases} min\left\{\frac{\hat{b}_i}{\hat{d}_i} \middle| \hat{d}_i > 0\right\}, 若\hat{d}不是 \le 0 \\ \infty, 若\hat{d} \le 0 \end{cases}$$ (7-9)计算得到λ上限$\bar{\lambda}$,然后作如下的一维搜索$min\ f(x_k + \lambda d_k), s.t. 0 \le \lambda \le \bar{\lambda}$,求得最优解$\lambda_k$,令$x_{k+1} = x_k + \lambda_k d_k$;

◆ 7.置k:=k+1,Goto 2.

第7章约束问题的最优化方法(Methods of Constrained Optimization)-可行方向法- 线性约束-计算例子

- 用可行方向法求解$min\ f(x)=(x_1-6)^2+(x_2-2)^2;s.t.\begin{cases}-x_1+2x_2\le 4\\3x_1+2x_2\le 12\\-x_1\le 0\\-x_2\le 0\end{cases}$,取$x_0=(2,3)^T,\epsilon_1=0.001,\epsilon_2=0.001$
- 解:将x_0代入约束条件，得$g_1(x_0)=4,g_2(x_0)=12,g_3(x_0)=-2,g_4(x_0)=-3$,可见第1,2约束为起作用约束，故$A_1=\begin{pmatrix}-1&2\\3&2\end{pmatrix},b_1=\binom{4}{12}$,令$d=\binom{d_1}{d_2}$,求解线性规划问题$min\ z=\nabla f(x_0)^Td;s.t.\begin{cases}A_1d\le 0\\Ed=0\\-1\le d_i\le 1,i=1,2\end{cases}$,这里$E=0,\nabla f(x)=(2x_1-12,2x_2-4)^T$,故$\nabla f(x_0)^T=(-8\ \ 2)$,线性规划问题成为：$min\ z=-8d_1+2d_2;s.t.\begin{cases}-d_1+2d_2\le 0\\3d_1+2d_2\le 0\\-1\le d_i\le 1,i=1,2\end{cases}$,
- 将其化为如下的标准形式,$x_1=d_1+1,x_2=d_2+2$,
- $max\ z=8x_1-2x_2-6;$
- $s.t.\begin{cases}-x_1+2x_2+x_3=1\\3x_1+2x_2+x_4=5\\x_1+x_5=2\\x_2+x_6=2\\x_i\ge 0,i=1\cdots,6\end{cases}$,

◆从而用单纯形法解得$d_1 = x_1 - 1 = \frac{2}{3}, d_2 = x_2 - 1 = -1$,即得

$$x_1 = x_0 + \lambda d_0 = \left(2 + \frac{2}{3}\lambda \;\; 3 - \lambda\right)^T,$$

◆作一维搜索，先由:$\hat{d} = A_2 d_k, \hat{b} = b_2 - A_2 x_k, \bar{\lambda} =$

$$\begin{cases} min\left\{\frac{\hat{b}_i}{\hat{d}_i} \middle| \hat{d}_i > 0\right\}, 若\hat{d}不是 \le 0 \\ \infty, 若\hat{d} \le 0 \end{cases}$$

◆求得$\bar{\lambda} = min\left\{-\frac{3}{-1}\right\} = 3$

◆解规划问题$min\, f(x_1), s.t. 0 \le \lambda \le 3$,得$\lambda_1 = \frac{33}{13}, f(x_1) = \frac{100}{13}$,此时得到新的迭代点$x_1 = \left(\frac{48}{13}, \frac{6}{13}\right)^T, f(x_1) = \frac{100}{13}, \nabla f(x_1)^T = \left(-\frac{60}{13}, -\frac{40}{13}\right)^T, ||\nabla f(x_1)|| > \epsilon_1$,继续寻找可行下降方向$g_1(x_1) < 4, g_2(x_1) = 12, g_3(x_1) < 0, g_4(x_1) < 0$

◆ 因此$A_1=(3\ 2), b_1=12$,再次求解求解线性规划问题$min\ z=\nabla f(x_1)^T d; s.t.\begin{cases} A_1 d\le 0 \\ Ed=0 \\ -1\le d_i\le 1, i=1,2\end{cases}$,同样引入变量$x_1=d_1+1, x_2=d_2+1$,形成标准的单纯形，用单纯形法解得得目标函数值为0,$d_1=\frac{2}{3}, d_2=-1$，从而得最优点为$x^*=\left(\frac{48}{13},\frac{6}{13}\right)^T, f(x^*)=\frac{100}{13}$

- 约束为非线性函数时(7 − 5)$Min\ f(x); s.t.\ g(x) \le 0$，通过求解线性规划(7 − 6)$Min\ \eta; s.t. \begin{cases} \nabla f(\bar{x})^T d \le \eta, \\ \nabla g_i(\bar{x})^T d \le \eta, i \in I \\ -1 \le d_j \le 1, j = 1, 2, \cdots, n \end{cases}$ 得到x_k及可行下降方向d_k,现在沿该方向进行一维搜索求得λ_k,即求解$min\ f(x_k + \lambda d_k), s.t. 0 \le \lambda \le \bar{\lambda}$,其中$\bar{\lambda} = sup\{\lambda | g_i(x_k + \lambda d_k) \ge 0, i = 1, 2, \cdots, m\}$,约束为非线性约束时，非线性规划(7-5)的可行方向法的计算步骤如下

- 1.给定初始可行点$x_0 \in X$,允许误差$\epsilon_1 > 0, \epsilon_2 > 0, k = 0$;
- 2.确定点x_k处的起作用约束集$I(x_k)$：$I(x_k) = \{i | g_i(x_k) = 0, 1 \le i \le m\}$;
  - 2.1若约束集$I(x_k) = \emptyset$为空集，且$\|\nabla f(x_k)\| < \epsilon_1$,则停止迭代，得到近似极小点$x_k$;
  - 2.2若约束集$I(x_k) = \emptyset$为空集，且$\|\nabla f(x_k)\| \ge \epsilon_1$,则取搜索方向$d_k = -\nabla f(x_k)$,Goto 5,这是因为类似无约束优化问题
  - 2.3若约束集$I(x_k) \ne \emptyset$不为空集，Goto 3;
- 3. 求解线性规划(7 − 6)$Min\ \eta; s.t. \begin{cases} \nabla f(\bar{x})^T d \le \eta, \\ \nabla g_i(\bar{x})^T d \le \eta, i \in I \\ -1 \le d_j \le 1, j = 1, 2, \cdots, n \end{cases}$，设求得的解为$(d_k, \eta_k)$;

- **4.判断精度，若$|\eta| < \epsilon_2$,则停止迭代，因为$\eta_k \approx 0$,说明找不到可行下降方向，可以认为x_k是一个KT点;否则，以d_k为搜索方向，Goto 5;**
- **5. 作一维搜索，先根据$\bar{\lambda} = sup\{\lambda | g_i(x_k + \lambda d_k) \le 0$,得到$\lambda$上限$\bar{\lambda}$,然后作如下的一维搜索$min\, f(x_k + \lambda d_k), s.t. 0 \le \lambda \le \bar{\lambda}$,求得最优解$\lambda_k$,令$x_{k+1} = x_k + \lambda_k d_k$;**
- **6.**置k:=k+1,**Goto 2.**

第7章约束问题的最优化方法(Methods of Constrained Optimization)-可行方向法-非线性约束

- 例子，用可行方向法求解$min\ f(x)=x_1^2+x_2^2-4x_1-4x_2, s.t. x_1+2x_2\le 4$，取$x_0=(0\ 0)^T, \epsilon_1=0.001, \epsilon_2=0.001$
- 解：计算目标函数和约束函数的导数$\nabla f(x)=(2x_1-4, 2x_2-4)^T$，$\nabla g_1(x)=(1,2)^T$,故$\nabla f(x_0)=(-4\ \ -4)^T, g_1(x_0)=-4<0$,起作用约束为空集，$\|\nabla f(x_0)\|>\epsilon_1$,说明$x_0$为可行域内点，按最速下降法取$d_0=(4\ 4)^T$,得$x_1=(4\lambda\ 4\lambda)^T$,此时将$x_1$代入约束条件，$g_1(x_1)=12\lambda-4\le 0$,可得$\lambda\le\frac{1}{3}=\bar{\lambda}$,求解$min_{0\le\lambda\le\frac{1}{3}} f(x_0+\lambda d_0)$, 求稳定点得到$\lambda_0'=\frac{1}{2}>\bar{\lambda}$，说明$x_1=x_0+\lambda_0' d_0$不在可行域内，原函数$f(x_0+\lambda d_0)$在$[0,\frac{1}{3}]$单调递减，因而取最优值为$\lambda_0=\frac{1}{3}=\bar{\lambda}$,得$x_1=\begin{bmatrix}\frac{4}{3} & \frac{4}{3}\end{bmatrix}^T, f(x_1)=-\frac{64}{9}, \nabla f(x_1)=\left(-\frac{4}{3}\ \ -\frac{4}{3}\right)^T$, $g_1(x_1)=0$,说明起作用约束集不为空
- 构造下列线性规划$Min\ \eta; s.t. \begin{cases} -\frac{4}{3}d_1-\frac{4}{3}d_2\le\eta, \\ d_1+2d_2\le\eta, \\ -1\le d_j\le 1, j=1,2 \end{cases}$ 与前述类似，化为标准型，然后单纯形法求解：$x_1=d_1+1, x_2=d_2+1$

◆可得最优解为$\eta=-\frac{4}{10}$,从而搜索方向为$d_1=(1\ \ -0.7)^T$, $x_2=x_1+\lambda d_1=\left(\frac{4}{3}+\lambda\ \ \frac{4}{3}-0.7\lambda\right)^T$,代入得到$f(x_2)=1.49\lambda^2-0.4\lambda-\frac{64}{9}\Rightarrow\lambda=\frac{0.4}{2.98}\approx0.134$,再进一步计算得到$x_2=(1.467\ 1.240)^T$，从而$g_1(x_2)=-0.053<0$,说明$x_2$在可行域内，继续迭代得到最优解$x^*=(1.6\ 1.2)^T, f(x^*)=-7.2$

- 不管是线性约束还是非线性约束，都转化为基本的线性规划问题来求解，然后借助一维搜索和迭代方法(无约束最优化问题)
- 近似规划法：将非线性模型线性化，然后通过解线性规划来求原问题的近似最优解
- 制约函数法：通过构造制约函数，将约束问题转化为一系列无约束问题，进而用无约束最优化方法求解，亦称为序列无约束极小化方法(SUMT:Sequential Unconstrained Minimization Technique)。按照取罚函数的方法不同，分为罚函数法(外点法)和闸函数法(内点法)。例如原问题为$min\ f(x), s.t.\ x \in X$,可以构造惩罚函数$\alpha(x) = \begin{cases} 0, x \in X \\ +\infty, x \notin X \end{cases}$;而闸函数法的基本思想是把惩罚加在约束集的边界，在内部时，受极小的惩罚或不受惩罚，在可行域边界时，惩罚项→∞,从而使迭代点始终在可行域内部
- 乘子法：罚函数的主要困难在于其辅助问题的最优解要达到对原问题解的较好的近似，常常在罚因子的极限情况才能实现。但如果用拉格朗日函数来取代$f(x)$构造辅助问题，根据K-T条件，在最优解处，存在最优乘子使得$\nabla_x L = 0$,从而弥补了$\nabla f(x^*) \neq 0$造成的困难

第7章约束问题的最优化方法(Methods of Constrained Optimization)-可行方向法-非线性约束

◆假定所有问题都是凸的

- ➢二次问题求解：封闭形式的解
- ➢等式约束二次问题：**KKT**条件得封闭形式的解
- ➢等式约束光滑问题：牛顿法将其化解为一系列等式约束二次问题
- ➢不等式和等式约束光滑问题：使用内点法化解为一系列等式约束光滑问题

第7章约束问题的最优化方法(Methods of Constrained Optimization)-制约函数法

- 对于常见的约束优化问题COP:$min f(x), s.t. c_i(x)=0, i=1,\cdots,m_e, c_j(x)\geq 0, j=m_e+1,\cdots,m$，构造具有惩罚性质的函数:
$$P(x)=\bar{P}\big(f(x),c(x)\big)$$
- 要求对原问题的可行点$x\in\mathcal{X}$均有$P(x)=f(x)$;而当约束条件破坏很大时有$P(x)$远大于$f(x)$.
- 例如：定义违反度函数$c^{(-)}(x)=\left(c_1^{(-)}(x),\cdots,c_m^{(-)}(x)\right)^T$如下：
$$c_i^{(-)}=c_i(x), i=1,\cdots,m_e,\text{表示等式约束}$$
$$c_i^{(-)}=min\{0,c_i(x)\}, i=m_e+1,\cdots,m$$

注意：$c_j(x)\leq 0$，则$c_i^{(-)}=max\{0,c_i(x)\}, i=m_e+1,\cdots,m$

- 定义集合$C=\{c|c\in R^m, c_i=0, i\in E; c_i\geq 0, i\in I\}$,显然，$x$是可行点$\Leftrightarrow c(x)\in C$.对任何$x\in R^n$都有：
- $||c^{(-)}(x)||_2=dist(c(x),C)=min\{||c(x)-y||_2|y\in C\}$,从而罚函数一般可表示为目标函数与一项与约束有关的“罚项”之和：
- $P(x)=f(x)+h\left(c^{(-)}(x)\right)$,其中$h\left(c^{(-)}(x)\right)$满足：$h(0)=0, lim_{||c||\to+\infty}h(c)=+\infty$

第7章约束问题的最优化方法(Methods of Constrained Optimization)-可行方向法-制约函数法

- $P(x) = f(x) + \sigma||c^{(-)}(x)||^{\alpha}, \alpha > 0$是罚因子，两种特殊的形式如下：
- $P_1(x) = f(x) + \sigma||c^{(-)}(x)||_1$
- $P_\infty(x) = f(x) + \sigma||c^{(-)}(x)||_\infty$
- 分别称为L_1, L_∞罚函数
- 若约束均为等式约束，此时定义其罚函数为：
- $P_2(x) = f(x) + \sigma||c^{(-)}(x)||_2^2$
- 注意，上述罚函数中，点都在可行域外，在边界上的时候取值为最优，罚函数为0，因此称为外点罚函数。
- 在可行域中，$P(x, \alpha)$的全局极小点与约束最优化问题COP的最优解相同。注意罚因子大了，小了都不好！
  - 小了，不可行解处的函数下降抵消了罚函数对违反约束的惩罚
  - 因此，一般逐步取$\sigma_{k+1} = \rho \cdot \sigma_k, \rho > 1$,在LASSO中称为连续化
  - 增长太快，导致求解不容易，增长太慢则迭代次数增加
  - 子问题求解精度必须满足要求：足够精确！

- **定理：设x_{k+1}是$P_2(x,\sigma_k)$的全局极小点，σ_k单调递增趋于正无穷，则$\{x_k\}$的每个极限点都是原问题的全局极小点。**
  - 注意：全局极小点很难求解，因此只能用逼近解来代替，但要求其在一定精度范围内！
- **定理：设$f(x)$和$c_i(x), i \in E$都是连续可微的，正数序列$\epsilon_k \to 0, \sigma_k \to +\infty$,如果子问题$min_x\{P_2(x,\sigma_k)\} = x_{k+1}$,满足$||\nabla_x P_2(x_{k+1},\sigma_k)|| \le \epsilon_k$,则对$\{x_k\}$的任何极限点$x^*$,都有$\{\nabla c_i(x^*), i \in E\}$线性无关，则$x^*$是等式约束最优化问题的KKT点，且**

$$lim_{k\to\infty}\left(-\sigma_k c_i(x_{k+1})\right) = \lambda_i^*, \forall i \in E.$$

- **其中λ_i^*是约束$c_i(x^*) = 0$对应的拉格朗日乘子法。**
  - 证明：$||\nabla P_2(x,\sigma_k)|| = ||\nabla f(x) + \sum_{i\in E}\sigma_k c_i(x)\nabla c_i(x)|| \le \epsilon_k \Rightarrow || = || \sum_{i\in E} c_i(x)\nabla c_i(x)|| \le \frac{1}{\sigma_k}(\epsilon_k + ||\nabla f(x_{k+1})||)$,两边取极限，令$k \to \infty$,可知$\sum_{i\in E} c_i(x)\nabla c_i(x) = 0$，又因为$\{\nabla c_i(x^*), i \in E\}$线性无关，因此$c_i(x^*) = 0$即$x^*$是可行点。
  - 后面证明其满足KKT条件中的梯度条件，令$\lambda_i^k = -\sigma_k c_i(x_{k+1})$

- 因此有：$\nabla c(x_{k+1})\lambda^k = \nabla f(x_{k+1}) - \nabla_x P_2(x_{k+1}, \sigma_k)$, $\nabla c(x^*)$是列满秩矩阵，而$x_k \to x^*, k \to \infty$, $\nabla c(x_{k+1})$应该是列满秩矩阵，因此用前述的广义逆表示λ_k:
- $\lambda_k = \left(\nabla c(x_{k+1})^T \nabla c(x_{k+1})\right)^{-1} \nabla c(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla_x P_2(x_{k+1}, \sigma_k))$
- 令$k \to \infty$,且$\nabla_x P_x(x_{k+1}, \sigma_k) \to 0$，从而有

$$\lambda^* = lim_{k\to\infty}\lambda_k = \left(\nabla c(x^*)^T \nabla c(x^*)\right)^{-1} \nabla c(x^*)^T \nabla f(x^*)$$

- 在$\nabla_x P_2(x, \sigma_k) = \nabla f(x) + \sigma_k c_i(x) \nabla c_i(x)$中令$k \to \infty$则可得
- $\nabla f(x^*) - \nabla c(x^*)\lambda^* = 0$
- 即证x^*满足KKT的梯度条件，因此λ^*就是其对应的拉格朗日乘子。
- 分析：
  - 迭代给出的解x_k的聚点总是$||c(x)||^2$的稳定点,即便没找到解，也是使得约束违反度较小的一个解
  - 定理不要求每一个子问题精确求解，但要求精度越来越高
- 如在可行域边界上取值为无穷，则称为内点罚函数，内点罚函数仅适合不等式约束问题，两个常见的内点罚函数是倒数和对数罚函数：
- $P(x) = f(x) + \sigma^{-1} \sum_{i=1}^m \frac{1}{c_i(x)}$， $P(x) = f(x) - \sigma^{-1} \sum_{i=1}^m log c_i(x)$

注意：内点罚函数要求初始点是可行解

- 设x^*是COP问题的KT点，则有$\nabla P(x^*) = \nabla f(x^*)$,因此$x^*$一般不是Courant罚函数的稳定点。为克服这一问题，引入参数$\theta_i(i = 1, \cdots, m)$,令不等式对应的参数$\theta_i \geq 0$,令$\theta = (\theta_1, \cdots, \theta_m)^T$.则Courant罚函数可修改为：
- $P(x) = f(x) + \sum_{i=1}^{m} \frac{\sigma_i}{2}\left(\left[(c(x) - \theta)_i^{(-)}\right]^2 - \theta_i^2\right) = f(x) + \sum_{i=1}^{m_e}\left[-\lambda_i c_i(x) + \frac{1}{2}\sigma_i\left(c_i(x)\right)^2\right] + \sum_{i=m_e+1}^{m}\begin{cases} -\lambda_i c_i(x) + \frac{1}{2}\sigma_i\left(c_i(x)\right)^2, & c_i(x) < \lambda_i/\sigma_i \\ -\frac{1}{2}\frac{\lambda_i^2}{\sigma_i}, & \text{否则} \end{cases}$
- 这里，$\lambda_i = \sigma_i\theta_i, i = 1, \cdots, m$,上式可以看作是Lagrange函数加上一项具有罚性质的项，因而称为增广拉格朗日函数.
- 假设x^*是COP的KT点，$\lambda_i^*(i = 1, \cdots, m)$是对应的Lagrange乘子，则对采用其构造的增广拉格朗日函数满足：$\nabla P(x^*) = 0$。
- 实际上，在约束规划求解之前，并不知道乘子，因而基于增广拉格朗日函数的罚函数方法需要逐步修正乘子λ_i
- 对于等式约束问题，可定义：$\lambda(x) = \left(A(x)\right)^+ g(x), A(x) = \nabla c(x)^T, g(x) = \nabla f(x)$, A^+表示其广义逆矩阵。因而乘子是问题的最小二范数解:$\min_{\lambda} ||\nabla f(x) = \sum_{i=1}^{m} \lambda_i \nabla c_i(x)||_2^2$
- 利用上面的$\lambda(x) = \left(A(x)\right)^+$，可得Fletcher光滑精确罚函数法：

$P(x) = f(x) - \lambda(x)^T c(x) + \frac{1}{2}\sum_{i=1}^{m} \sigma_i\left(c_i(x)\right)^2, \sigma_i > 0$是罚因子!

- **以等式约束的二次罚函数构造增广拉格朗日函数**
  - $c_i(x_{k+1}) \approx -\frac{\lambda_i^*}{\sigma_k}, \forall i \in E$
  - 为保证可行性，罚因子必须趋于正无穷，此时子问题条件数过大而难以求解，是否可以通过罚函数修正，使得对有限的罚因子，得到的近似最优解也是可行的呢？
- $L_\sigma(x,\lambda) = f(x) + \sum_{i\in E} \lambda_i c_i(x) + \frac{1}{2}\sigma \sum_{i\in E} c_i^2(x)$
  - 在拉格朗日函数的基础上增加二次惩罚项
  - 在第k步迭代，给定罚因子σ_k和乘子λ_k,增广拉格朗日函数$L_{\sigma_k}(x,\lambda^k)$的最小值点$x_{k+1}$满足
  - $\nabla_x L_{\sigma_k}\left(x_{k+1},\lambda^k\right) = \nabla f(x_{k+1}) + \sum_{i\in E}\left(\lambda_i^k + \sigma_k c_i(x_{k+1})\right)\nabla c_i(x_{k+1}) = 0$
  - 而拉格朗日函数最优点满足：
  - $\nabla f(x^*) + \sum_{i\in E} \lambda_i^* \nabla c_i(x^*) = 0$
  - 为使得增广拉格朗日函数的迭代解最终与拉格朗日函数一致，必定对于充分大的k,有$\lambda_i^k + \sigma_k c_i(x_{k+1}) \approx \lambda_i^*, \forall i \in E$

◆ **以等式约束的二次罚函数构造增广拉格朗日函数**

- $\nabla_x L_{\sigma_k}\left(x_{k+1},\lambda^k\right)=\nabla f(x_{k+1})+\sum_{i\in E}\left(\lambda_i^k+\sigma_k c_i(x_{k+1})\right)\nabla c_i(x_{k+1})=0$
- 而拉格朗日函数最优点满足：
- $\nabla f(x^*)+\sum_{i\in E}\lambda_i^*\nabla c_i(x^*)=0$
- 为使得增广拉格朗日函数的迭代解最终与拉格朗日函数一致，必定对于充分大的k,有$\lambda_i^k+\sigma_k c_i(x_{k+1})\approx\lambda_i^*,\forall i\in E$
- 即：$c_i(x_{k+1})\approx\frac{1}{\sigma_k}(\lambda_i^*-\lambda_i^k)$,违反度与$\frac{1}{\sigma_k}$成正比，接近最优解时，约束违反度远小于$\frac{1}{\sigma_k}$，注意，减少违反度，可以通过更新乘子来达到
- 即：$\lambda_i^{k+1}=\lambda_i^k+\sigma_k c_i(x_{k+1}),\forall i\in E$

◆ **可以扩展到一般的不等式约束**

- 将不等式添加变量松弛到等式约束，然后构造增广拉格朗日函数

◆ **凸规划更没问题！**

第7章

- 本章介绍了几种基本的约束优化求解方法
- 基本思想：转化为无约束优化问题求解，采用近似、逼近、迭代的思想构造子问题求解
- 一般可以利用问题的结构，其中最典型的结构就是可分离结构，可写成各项之和
  - 思考坐标下降法$f: R^{n+m} \rightarrow R$,将f视为$f(x, y), x \in R^n, y \in R^m$,从而迭代时：$x^{(k+1)} \leftarrow min_x f\left(x, y^{(k)}\right); y^{(k+1)} \leftarrow min_y f(x^{(k+1)}, y)$
  - 但一般无法证明交替法能达到全局或局部最优解，但实际运行对很多问题都效果良好（例如，广义PCA）
- 最后，智能优化算法一般都可以在实际中采用，例如各种生物启发式算法!

End！