智能优化方法及其应用

第一讲 经典优化算法1

课程纲要

- 最优化理论数学基础复习
- 线搜索方法
- 梯度法和牛顿法
- 共轭梯度法
- 拟牛顿法
- 最小二乘法
- 最优性条件(*)
- 线性规划
- 二次规划

课程纲要

- 最优化理论数学基础复习
- 线搜索方法
- 梯度法和牛顿法
- 共轭梯度法
- 拟牛顿法
- 最小二乘法
- 最优性条件(*)
- 线性规划
- 二次规划

最优化理论数学基础

- **最优化问题**就是求一个多元函数在某个给定集合上的极值问题，几乎所有类型的最优化问题都可以描述如下：

$$\min f(\boldsymbol{x}),$$
$$\text{s.t. } \boldsymbol{x} \in \Omega$$

其中：Ω为某个给定的集合，称为**可行集或可行域**；

$f(x)$为定义在集合Ω上连续可微的多元实值函数，称为**目标函数**；$\boldsymbol{x} = (x_1, x_2 \cdots, x_n)^T$为**决策变量**；s.t. 为**subject to (受限于)**的缩写.

对于极大化问题，在目标函数前添加负号可以转化为极小化问题. 因此，只需考虑目标函数极小化问题.

最优化理论数学基础

- **最优化问题:**

$$\min f(\boldsymbol{x}),$$
$$\text{s.t. } \boldsymbol{x} \in \Omega$$

可行域Ω一般常用等式和不等式来描述

$$\Omega = \{\boldsymbol{x} \in R^n | c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}; c_i(\boldsymbol{x}) \geq 0, i \in \mathcal{I}\}$$

其中$c_i(\boldsymbol{x})\ (i \in \mathcal{E} \cup \mathcal{I})$为定义在$\boldsymbol{R}^n$上连续可微的多元实值函数, 称为**约束函数**.

等式约束: $c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}$; $\mathcal{E}$为等式约束指标集;

不等式约束: $c_i(\boldsymbol{x}) \geq 0, i \in \mathcal{I}$; $\mathcal{I}$为不等式约束指标集

最优化理论数学基础

- 最优化问题:

$$\min f(\boldsymbol{x}),$$
$$\text{s.t. } \boldsymbol{x} \in \Omega$$
$$\Omega = \{\boldsymbol{x} \in R^n | c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}; c_i(\boldsymbol{x}) \geq 0, i \in \mathcal{I}\}$$

无约束优化问题: $\mathcal{E} \cup \mathcal{I} = \emptyset$, 否则为约束优化问题

等式约束优化问题: $\mathcal{E} \neq \emptyset$ 且 $\mathcal{I} = \emptyset$

不等式约束优化问题: $\mathcal{I} \neq \emptyset$ 且 $\mathcal{E} = \emptyset$

线性规划问题: 目标函数和约束函数都是线性函数

二次规划问题: 目标函数二次, 而约束函数是线性

向量和矩阵范数

- 在n维实向量空间$\boldsymbol{R}^n$中定义向量的范数
- 向量$\boldsymbol{x} \in \boldsymbol{R}^n$的范数$\|\cdot\|$是一个非负数，须满足以下条件：
  - ① $\|\boldsymbol{x}\| \geq 0, \|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$
  - ② $\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|, \forall \alpha \in \boldsymbol{R}$
  - ③ $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- p-范数：$\|\boldsymbol{x}\|_p = (\sum_{i=1}^{n}|x_i|^p)^{\frac{1}{p}},\ p \geq 1$
  - 1-范数：$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n}|x_i|$
  - 2-范数：$\|\boldsymbol{x}\|_2 = (\sum_{i=1}^{n}|x_i|^2)^{\frac{1}{2}}$
  - ∞-范数：$\|\boldsymbol{x}\|_\infty = \max\limits_{1\leq i\leq n}|x_i|$

向量和矩阵范数

- 按照向量范数的定义引入矩阵$\boldsymbol{A} \in \boldsymbol{R}^{m\times n}$的**矩阵范数**
- 矩阵范数$\|\cdot\|$还需满足乘法性质:

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|, \qquad \forall \boldsymbol{A} \in \boldsymbol{R}^{m\times n}, \boldsymbol{B} \in \boldsymbol{R}^{n\times q}$$

- 称矩阵范数$\|\cdot\|_{\mu}$与向量范数$\|\cdot\|$是**相容**的, 当

$$\|\boldsymbol{Ax}\| \leq \|\boldsymbol{A}\|_{\mu}\|\boldsymbol{x}\|, \qquad \forall \boldsymbol{A} \in \boldsymbol{R}^{m\times n}, \boldsymbol{x} \in \boldsymbol{R}^{n}$$

- 若存在$\boldsymbol{x} \neq \boldsymbol{0}$使得

$$\|\boldsymbol{A}\|_{\mu} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{Ax}\|$$

称矩阵范数$\|\cdot\|_{\mu}$是由向量范数$\|\cdot\|$诱导出来的**算子范数**, 也称为从属于向量范数$\|\cdot\|$的矩阵范数. 此时, 向量范数和算子范数通常采用相同的符号$\|\cdot\|$.

向量和矩阵范数

- 显然，从属于向量范数$\|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_2, \|\boldsymbol{x}\|_\infty$的矩阵范数为：

$$\|\boldsymbol{A}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$$

$$\|\boldsymbol{A}\|_2 = max\{\sqrt{\lambda} | \lambda \in \lambda(\boldsymbol{A}^T\boldsymbol{A})\}$$

$$\|\boldsymbol{A}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$$

分别称为**列和范数**，**谱范数**，**行和范数**.

注：$\lambda(\boldsymbol{A}^T\boldsymbol{A})$表示$\boldsymbol{A}^T\boldsymbol{A}$的特征值

向量和矩阵范数

- F-范数:

$$\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \sqrt{tr(\boldsymbol{A}^T\boldsymbol{A})}$$

例如:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow \boldsymbol{A}^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}$$

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} a_{11}^2 + a_{21}^2 & a_{11}a_{12} + a_{21}a_{22} \\ a_{11}a_{12} + a_{21}a_{22} & a_{12}^2 + a_{22}^2 \end{bmatrix}$$

$$tr(\boldsymbol{A}^T\boldsymbol{A}) = a_{11}^2 + a_{12}^2 + a_{21}^2 + a_{22}^2$$

谱范数和F-范数常用来讨论各种迭代算法的收敛性.

向量序列与矩阵序列的收敛性

- 若向量序列$\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty} \subset \boldsymbol{R}^n$, 则:

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} \Leftrightarrow \lim_{k\to\infty} x_i^{(k)} = x_i$$

- 类似, 若矩阵序列$\{\boldsymbol{A}^{(k)}\}_{k=1}^{\infty} \subset \boldsymbol{R}^{m\times n}$, 则:

$$\lim_{k\to\infty} \boldsymbol{A}^{(k)} = \boldsymbol{A} \Leftrightarrow \lim_{k\to\infty} a_{ij}^{(k)} = a_{ij}$$

定理 1.1 (1) 设 $\|\cdot\|$ 和 $\|\cdot\|'$ 是定义在 $\mathbf{R}^n$ 上的两个向量范数, 则存在两个正数 c_1 和 c_2, 对所有 $\boldsymbol{x} \in \mathbf{R}^n$ 均成立

$$c_1\|\boldsymbol{x}\| \leqslant \|\boldsymbol{x}\|' \leqslant c_2\|\boldsymbol{x}\|.$$

(2) 设 $\|\cdot\|$ 和 $\|\cdot\|'$ 是定义在 $\mathbf{R}^{m\times n}$ 上的两个矩阵范数, 则存在两个正数 m_1 和 m_2, 对所有 $\boldsymbol{A} \in \mathbf{R}^{m\times n}$ 均成立

$$m_1\|\boldsymbol{A}\| \leqslant \|\boldsymbol{A}\|' \leqslant m_2\|\boldsymbol{A}\|.$$

向量序列与矩阵序列的收敛性

定理 1.2 (1) 设 $\{\boldsymbol{x}^{(k)}\}$ 为 n 维向量序列, $\|\cdot\|$ 为定义在 $\mathbf{R}^n$ 上的向量范数, 则

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} \Longleftrightarrow \lim_{k\to\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = 0.$$

(2) 设 $\{\boldsymbol{A}^{(k)}\}$ 为 $m \times n$ 矩阵序列, $\|\cdot\|$ 为定义在 $\mathbf{R}^{m\times n}$ 上的矩阵范数, 则

$$\lim_{k\to\infty} \boldsymbol{A}^{(k)} = \boldsymbol{A} \Longleftrightarrow \lim_{k\to\infty} \|\boldsymbol{A}^{(k)} - \boldsymbol{A}\| = 0.$$

多元函数分析

- 主要介绍三个概念
  - ✓ n元函数的一阶导数
  - ✓ n元函数的二阶导数
  - ✓ n元函数的泰勒展开式

多元函数分析

- **定义1.1** $\boldsymbol{x}=(x_1,x_2\cdots,x_n)^T\in \boldsymbol{R}^n$. 称向量

$$\nabla f(\boldsymbol{x})=\left(\frac{\partial f(\boldsymbol{x})}{\partial x_1},\frac{\partial f(\boldsymbol{x})}{\partial x_2},\cdots,\frac{\partial f(\boldsymbol{x})}{\partial x_n}\right)^T$$

为$f(\boldsymbol{x})$在$\boldsymbol{x}$处的**一阶导数**或**梯度**. 称下列矩阵为$f(\boldsymbol{x})$在$\boldsymbol{x}$处的**二阶导数**或**Hesse阵**

$$\nabla^2 f(\boldsymbol{x})=\begin{pmatrix}\frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1\partial x_n}\\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2\partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n\partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2}\end{pmatrix}$$

多元函数分析

- 若梯度$\nabla f(\boldsymbol{x})$的每个分量函数在$\boldsymbol{x}$处都连续，则称f在$\boldsymbol{x}$处**一阶连续可微**.
- 若Hesse阵$\nabla^2 f(\boldsymbol{x})$的各个分量函数在$\boldsymbol{x}$处都连续，则称$f$在$\boldsymbol{x}$处**二阶连续可微**.
- 若f在开集$\mathcal{D}$的每一点都一阶连续可微，则称**f在$\mathcal{D}$上一阶连续可微**.
- 若f在开集$\mathcal{D}$的每一点都二阶连续可微，则称**f在$\mathcal{D}$上二阶连续可微**.

多元函数分析

- 显然，若f在$\boldsymbol{x}$处二阶连续可微，则

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i}, \qquad i,j = 1,2,\cdots,n$$

即Hesse阵$\nabla^2 f(\boldsymbol{x})$是对称阵.

$$\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{pmatrix}$$

多元函数分析

- **泰勒展开：** 设函数$f: \boldsymbol{R}^n \to \boldsymbol{R}$连续可微，则

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \int_0^1 \nabla f(\boldsymbol{x}+t\boldsymbol{h})^T \boldsymbol{h} dt$$
$$= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}+\theta\boldsymbol{h})^T \boldsymbol{h} \ \left(\theta \in (0,1)\right)$$
$$= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + o(\|\boldsymbol{h}\|)$$

$o(\|h\|)$是泰勒展开的余项，是一个一价无穷小

泰勒公式是一个用函数在某点的信息描述其附近取值的公式。如果函数足够平滑的话，在已知函数在某一点的各阶导数值的情况之下，泰勒公式可以用这些导数值做系数构建一个多项式来近似函数在这一点的邻域中的值。

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k$$

多元函数分析

- **泰勒展开：** 设函数$f: \boldsymbol{R}^n \to \boldsymbol{R}$连续可微，则

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + o(\|\boldsymbol{h}\|)$$

若函数f是二次连续可微的，则有

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + \int_0^1 (1-t)\boldsymbol{h}^T \nabla^2 f(\boldsymbol{x}+t\boldsymbol{h})\boldsymbol{h}dt$$

拉格朗日余项

$$= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})\boldsymbol{h} \left(\theta \in (0,1)\right)$$

$$= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{h} + o(\|\boldsymbol{h}\|^2)$$

多元函数分析

- **泰勒展开：** 设函数f: $\boldsymbol{R}^n \to \boldsymbol{R}$ 连续可微, 则

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + o(\|\boldsymbol{h}\|)$$

若函数f是二次连续可微的, 则有

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{h} + o(\|\boldsymbol{h}\|^2)$$

以及

$$\begin{aligned}\nabla f(\boldsymbol{x}+\boldsymbol{h}) &= \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{x}+t\boldsymbol{h})^T \boldsymbol{h} dt \\ &= \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})^T \boldsymbol{h} \ \left(\theta \in (0,1)\right) \\ &= \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})^T \boldsymbol{h} + o(\|\boldsymbol{h}\|)\end{aligned}$$

多元函数分析

- 例1.1: 设二次函数

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

式中: $\boldsymbol{b} \in \boldsymbol{R}^n$ 且 $\boldsymbol{A} \in \boldsymbol{R}^{n\times n}$ 是**对称阵**. 求 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处的梯度和Hesse阵.

- **解:** 显然 $\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}, \nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}$

例: $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} \\ \boldsymbol{a_2} & \boldsymbol{a_3} \end{bmatrix}$

对称阵条件实际中通常满足，如: 距离矩阵等

多元函数分析

- 向量值函数

$$\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \cdots, F_m(\boldsymbol{x}))^T : \boldsymbol{R}^n \to \boldsymbol{R}^m$$

- 若每个分量函数F_i都是(连续)可微的, 则称$\boldsymbol{F}$是(连续)可微的.
- 向量值函数$\boldsymbol{F}$在$\boldsymbol{x}$处的导数$\boldsymbol{F}' \in \boldsymbol{R}^{m\times n}$称为它在$\boldsymbol{x}$处的**Jacobi矩阵**,记为$\boldsymbol{F}'(\boldsymbol{x})$或$\boldsymbol{J}_{\boldsymbol{F}}(\boldsymbol{x})$, 其转置称为梯度矩阵$\nabla \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{J}_{\boldsymbol{F}}(\boldsymbol{x})^T$:

$$\boldsymbol{F}'(\boldsymbol{x}) := \boldsymbol{J}_F(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial F_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial F_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial F_1(\boldsymbol{x})}{\partial x_n} \\ \frac{\partial F_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial F_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial F_2(\boldsymbol{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m(\boldsymbol{x})}{\partial x_1} & \frac{\partial F_m(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial F_m(\boldsymbol{x})}{\partial x_n} \end{pmatrix}$$

多元函数分析

- 同理: 若向量值函数$\boldsymbol{F}$: $\boldsymbol{R}^n \to \boldsymbol{R}^m$是连续可微的, 则对于任意的$\boldsymbol{x}, \boldsymbol{h} \in \boldsymbol{R}^n$, 有:

$$\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{h} + o(\|\boldsymbol{h}\|)$$

定义 1.2 设向量值函数 $\boldsymbol{F}: \mathbf{R}^n \to \mathbf{R}^m$, $\boldsymbol{x} \in \mathbf{R}^n$, 称 $\boldsymbol{F}$ 在 $\boldsymbol{x}$ 处是 Lipschitz 连续的, 是指存在常数 $L > 0$, 使得对任意的 $\boldsymbol{y} \in \mathbf{R}^n$, 满足

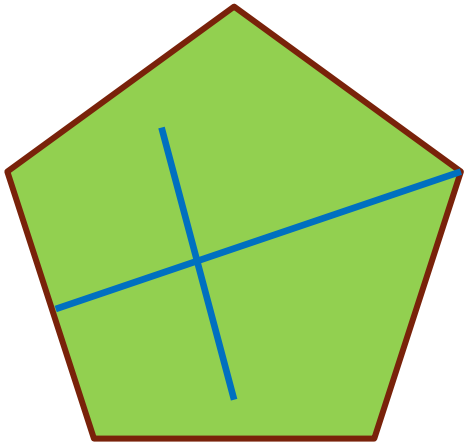
$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|, \tag{1.8}$$

式中: L 为 Lipschitz 常数.

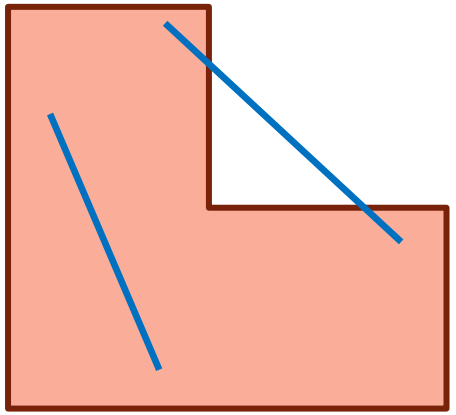
若式 (1.8) 对任意的 $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbf{R}^n$ 都成立, 则称 $\boldsymbol{F}$ 在 $\mathbf{R}^n$ 内是 Lipschitz 连续的.

凸集与凸函数

- 介绍凸集、锥和凸函数的有关概念
- **定义1.3** 设集合$\mathcal{D} \subset \boldsymbol{R}^n$. 称集合$\mathcal{D}$为**凸集**, 是指对任意的$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$及任意的实数$\lambda \in [0,1]$, 都有$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{D}$.



凸集



非凸

凸集与凸函数

- **性质1.1** 设$\mathcal{D}, \mathcal{D}_1, \mathcal{D}_2$是凸集, α是一实数, 那么

  (1) $\alpha\mathcal{D} := \{\boldsymbol{y}|\boldsymbol{y} = \alpha\boldsymbol{x}, \boldsymbol{x} \in \mathcal{D}\}$是凸集

  (2) 交集$\mathcal{D}_1 \cap \mathcal{D}_2$是凸集

  (3) 和集$\mathcal{D}_1 + \mathcal{D}_2 := \{\boldsymbol{z}|\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} \in \mathcal{D}_1, \boldsymbol{y} \in \mathcal{D}_2\}$是凸集

- **例1.3** n维欧几里得空间中的m个点的凸组合是一个凸集, 即集合

$$\left\{x = \sum_{i=1}^{m} \alpha_i x_i \,|x_i \in R^n, \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i = 1\right\}$$

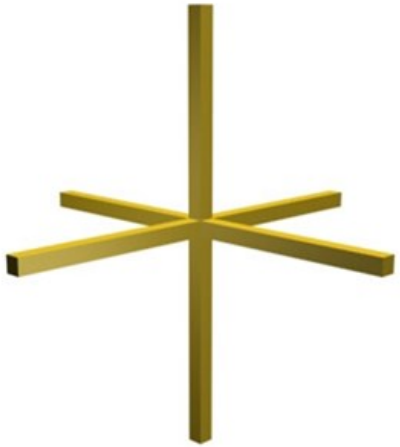
  是凸集

凸集与凸函数

- **定义1.4** 集合$\mathcal{D} \subset \boldsymbol{R}^n$的**凸包(convex hull)**是指所有包含$\mathcal{D}$的凸集的交集，记为

$$conv(\mathcal{D}) := \bigcap_{\mathcal{G} \supseteq \mathcal{D}} \mathcal{G}$$

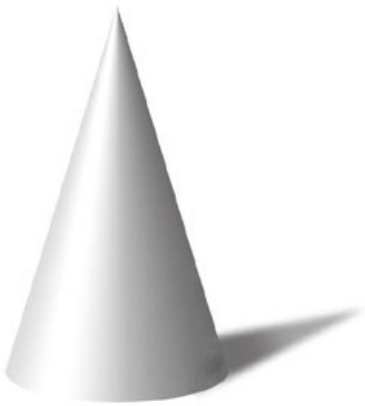
式中：$\mathcal{G}$为凸集



凸集与凸函数

- **定义1.5** 设非空集合$\mathcal{G} \subset \boldsymbol{R}^n$. 若对任意的$\boldsymbol{x} \in \mathcal{G}$和任意的实数$\lambda > 0$, 有$\lambda \boldsymbol{x} \in \mathcal{G}$, 则称$\mathcal{G}$为一个**锥(cone)**. 若$\mathcal{G}$同时也是凸集, 则称$\mathcal{G}$为一个**凸锥(convex cone)**. 此外, 对于锥$\mathcal{G}$, 若$\boldsymbol{0} \in \mathcal{G}$, 则称$\mathcal{G}$为一个**尖锥(pointed cone)**. 相应地, 包含$\boldsymbol{0}$的凸锥成为**尖凸锥**.

- **例1.6** 多面体$\{\boldsymbol{x} \in \boldsymbol{R}^n | \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{0}\}$是一个尖凸锥, 通常称为多面锥



凸集与凸函数

- 定义凸集上的凸函数
- **定义1.6** 设函数$f: \mathcal{D} \subset \boldsymbol{R}^n \to \boldsymbol{R}$, 其中$\mathcal{D}$为凸集.

(1) 称f是$\mathcal{D}$上的**凸函数**, 是指对任意的$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$及任意的实数$\lambda \in [0,1]$, 都有

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

(2) 称f是$\mathcal{D}$上的**严格凸函数**, 是指对任意的$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$及任意的实数$\lambda \in (0,1)$, 都有

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

(3) 称f是$\mathcal{D}$上的**一致凸函数**, 是指存在常数$\gamma > 0$, 使对任意的$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$及任意的实数$\lambda \in [0,1]$, 都有

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) + \frac{1}{2}\lambda(1-\lambda)\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

凸集与凸函数

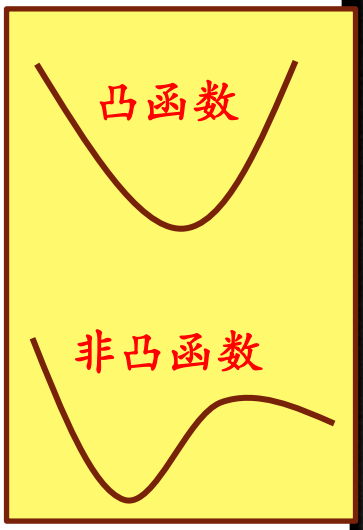
- 定义凸集上的**凸函数**
- **性质1.2** 设f, f_1, f_2都是凸集$\mathcal{D}$上的凸函数, $c_1, c_2 \in \boldsymbol{R}_+$, $\alpha \in \boldsymbol{R}$, 则有

  (1) $c_1 f_1(\boldsymbol{x}) + c_2 f_2(\boldsymbol{x})$也是$\mathcal{D}$上的凸函数;

  (2) 水平集$\mathcal{L}(f, \alpha) = \{\boldsymbol{x} | \boldsymbol{x} \in \mathcal{D}, f(\boldsymbol{x}) \leq \alpha\}$是凸集.

  - 凸集和凸函数在传统优化理论中起着举足轻重的作用, 但是用凸函数的定义来判断一个函数是否具有凸性相当困难.
  - 如果函数是一阶或二阶连续可微的, 那么利用函数的梯度或Hesse矩阵来判别或验证函数的凸性要相对容易.

函数的凸性



- **定理 1.4** 设 f 在凸集 $\mathcal{D} \subset \mathbf{R}^n$ 上一阶连续可微, 则

(1) f 在 $\mathcal{D}$ 上为凸函数的充要条件是

$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}), \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D};$$

(2) f 在 $\mathcal{D}$ 上为严格凸函数的充要条件是, 当 $\boldsymbol{x} \neq \boldsymbol{y}$ 时, 成立

$$f(\boldsymbol{x}) > f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}), \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D};$$

(3) f 在 $\mathcal{D}$ 上为一致凸函数的充要条件是, 存在常数 $c > 0$, 使对任意的 $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, 成立

$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) + c\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

函数的凸性

- 在一元函数中，若$f(x)$在区间(a,b)上二阶可微且$\boldsymbol{f''(x)\geq 0(>0)}$，则$f(x)$在$(a,b)$内**凸(严格凸)**.
- 对于二阶连续可微的多元函数$f:\mathcal{D}\subset \boldsymbol{R}^n\to \boldsymbol{R}$，也可以由其二阶导数(Hesse矩阵)给出凸性的一个近乎完整的表述.

定义 1.7 设 n 元实函数 f 在凸集 $\mathcal{D}$ 上是二阶连续可微的. 若对一切 $\boldsymbol{h}\in\mathbf{R}^n$, 有 $\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{h}\geqslant 0$, 则称 $\nabla^2 f$ 在点 $\boldsymbol{x}$ 处是半正定的. 若对一切 $\mathbf{0}\neq\boldsymbol{h}\in\mathbf{R}^n$, 有 $\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{h}>0$, 则称 $\nabla^2 f$ 在点 $\boldsymbol{x}$ 处是正定的. 进一步, 若存在常数 $c>0$, 使得对任意的 $\boldsymbol{h}\in\mathbf{R}^n$, $\boldsymbol{x}\in\mathcal{D}$, 有 $\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{h}\geqslant c\|\boldsymbol{h}\|^2$, 则称 $\nabla^2 f$ 在 $\mathcal{D}$ 上是一致正定的.

函数的凸性

- 在一元函数中，若$f(x)$在区间(a,b)上二阶可微且$\boldsymbol{f''(x)\geq 0(>0)}$，则$f(x)$在$(a,b)$内**凸(严格凸)**.
- 推广到多元函数
- **定理1.5** 设n元实函数f在凸集$\mathcal{D}\subset \boldsymbol{R}^n$上二阶连续可微，则：

  (1) f在$\mathcal{D}$上为**凸函数**的**充要条件**是$\nabla^2 f(x)$对一切$x\in\mathcal{D}$为**半正定**;

  (2) f在$\mathcal{D}$上为**严格凸函数**的**充分条件**是$\nabla^2 f(x)$对一切$x\in\mathcal{D}$为**正定**;

  (3) f在$\mathcal{D}$上为**一致凸函数**的**充要条件**是$\nabla^2 f(x)$对一切$x\in\mathcal{D}$为**一致正定**.

无约束问题的最优性条件

- 讨论无约束优化问题

$$\min_{x \in R^n} f(x)$$

的**最优性条件**

- ✓ 一阶必要条件
- ✓ 二阶必要条件
- ✓ 二阶充分条件
- ✓ 凸函数全局极小点的充要条件

无约束问题的最优性条件

- **定义 1.8** 若对于任意的 $\boldsymbol{x} \in \mathbf{R}^n$, 都有

$$f(\boldsymbol{x}^*) \leqslant f(\boldsymbol{x}),$$

则称 $\boldsymbol{x}^*$ 为 f 的一个全局极小点. 若上述不等式严格成立且 $\boldsymbol{x} \neq \boldsymbol{x}^*$, 则称 $\boldsymbol{x}^*$ 为 f 的一个严格全局极小点.

- **定义 1.9** 若对于任意的 $\boldsymbol{x} \in N(\boldsymbol{x}^*, \delta) = \{\boldsymbol{x} \in \mathbf{R}^n | \|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta\}$, 都有

$$f(\boldsymbol{x}^*) \leqslant f(\boldsymbol{x}),$$

则称 $\boldsymbol{x}^*$ 为 f 的一个局部极小点, 其中 $\delta > 0$ 为某个常数. 若上述不等式严格成立且 $\boldsymbol{x} \neq \boldsymbol{x}^*$, 则称 $\boldsymbol{x}^*$ 为 f 的一个严格局部极小点.

无约束问题的最优性条件

- 一般讨论的求极小点的方法都是指局部极小点.
- 为了表述的方便, 引入下列记号:

$$\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}), \boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$$

$$\boldsymbol{G}(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}), \boldsymbol{G}_k = \nabla^2 f(\boldsymbol{x}_k)$$

无约束问题的最优性条件

- **定理1.6** (一阶必要条件) 设$f(\boldsymbol{x})$在开集$\mathcal{D}$上一阶连续可微. 若$\boldsymbol{x}^* \in \mathcal{D}$是问题$\min\limits_{\boldsymbol{x}\in\boldsymbol{R}^n} f(\boldsymbol{x})$的一个局部极小点, 则必有$\boldsymbol{g}(\boldsymbol{x}^*) = 0$.

证明 取 $\boldsymbol{x} = \boldsymbol{x}^* - \alpha \boldsymbol{g}(\boldsymbol{x}^*) \in \mathcal{D}$, 其中 $\alpha > 0$ 为某个常数, 则有

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}^*) + \boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) + o(\|\boldsymbol{x} - \boldsymbol{x}^*\|) \\ &= f(\boldsymbol{x}^*) - \alpha \boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}^*) + o(\alpha) \\ &= f(\boldsymbol{x}^*) - \alpha \|\boldsymbol{g}(\boldsymbol{x}^*)\|^2 + o(\alpha). \end{aligned}$$

注意到 $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*)$ 及 $\alpha > 0$, 于是有

$$0 \leqslant \|\boldsymbol{g}(\boldsymbol{x}^*)\|^2 \leqslant \frac{o(\alpha)}{\alpha}.$$

上式两边令 $\alpha \to 0$, 得 $\|\boldsymbol{g}(\boldsymbol{x}^*)\| = 0$, 即 $\boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{0}$. 证毕.

无约束问题的最优性条件

- **定理1.7** (二阶必要条件) 设$f(\boldsymbol{x})$在开集$\mathcal{D}$上二阶连续可微. 若$\boldsymbol{x}^* \in \mathcal{D}$是问题$\min\limits_{\boldsymbol{x}\in\boldsymbol{R}^n} f(\boldsymbol{x})$的一个局部极小点, 则必有$\boldsymbol{g}(\boldsymbol{x}^*)=0$且$\boldsymbol{G}(\boldsymbol{x}^*)$是**半正定矩阵**.

证明 设 $\boldsymbol{x}^*$ 是一局部极小点, 那么由定理 1.6 可知 $\boldsymbol{g}(\boldsymbol{x}^*)=\boldsymbol{0}$. 下面只需证明 $\boldsymbol{G}(\boldsymbol{x}^*)$ 的半正定性. 任取 $\boldsymbol{x}=\boldsymbol{x}^*+\alpha\boldsymbol{d}\in\mathcal{D}$, 其中 $\alpha>0$ 且 $\boldsymbol{d}\in\mathbf{R}^n$. 由泰勒展开式, 得

$$f(\boldsymbol{x}+\boldsymbol{h})=f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^T\boldsymbol{h}+\frac{1}{2}\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{h}+o(\|\boldsymbol{h}\|^2)$$

$$0\leqslant f(\boldsymbol{x})-f(\boldsymbol{x}^*)=\frac{1}{2}\alpha^2\boldsymbol{d}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x}^*)\boldsymbol{d}+o(\alpha^2),$$

即

$$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x}^*)\boldsymbol{d}+\frac{o(2\alpha^2)}{\alpha^2}\geqslant 0.$$

对上式令 $\alpha\to 0$, 即得 $\boldsymbol{d}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x}^*)\boldsymbol{d}\geqslant 0$, 从而定理成立. 证毕.

无约束问题的最优性条件

- **定理1.8** (二阶充分条件) 设$f(\boldsymbol{x})$在开集$\mathcal{D}$上二阶连续可微. 若$\boldsymbol{x}^* \in \mathcal{D}$满足条件$\boldsymbol{g}(\boldsymbol{x}^*) = 0$且$\boldsymbol{G}(\boldsymbol{x}^*)$是**正定**矩阵, 则$\boldsymbol{x}^*$是问题$\min\limits_{\boldsymbol{x}\in\boldsymbol{R}^n} f(\boldsymbol{x})$的一个局部极小点.

- 目标函数的稳定点(驻点, 一阶导数为0)不一定是极小点.
- 但是, **目标函数是凸函数**的无约束优化问题, 其**稳定点、局部极小点和全局极小点等价**.

无约束问题的最优性条件

- **定理1.9** 设$f(\boldsymbol{x})$在R^n上是凸函数并且一阶连续可微. 则$\boldsymbol{x}^* \in R^n$是问题$\min\limits_{\boldsymbol{x}\in\boldsymbol{R}^n} f(\boldsymbol{x})$的全局极小点的充要条件是$\boldsymbol{g}(\boldsymbol{x}^*) = 0$.

证明 必要性是显然的, 所以只需证明充分性. 设 $\boldsymbol{g}(\boldsymbol{x}^*) = \mathbf{0}$. 由凸函数的判别定理 1.4 (1), 得

$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*) + \boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = f(\boldsymbol{x}^*), \quad \forall \boldsymbol{x} \in \mathbf{R}^n,$$

f 在 $\mathcal{D}$ 上为凸函数的充要条件是

$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}), \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D};$$

无约束优化问题的数值优化方法

- 在数值优化中，一般采用迭代法求解无约束优化问题的极小点.
- 基本思路如下:
  - ✓ 给定一初始点x_0;
  - ✓ 按照某迭代规则产生一迭代序列$\{x_k\}$，使该序列是**有限**的，则最后一个点就是极小点;
  - ✓ 若序列$\{x_k\}$是**无限**的，它有极限点且该点即为极小点.
- 迭代细节:
  - ➢ x_k为第k次迭代点，d_k为第k次**搜索方向**，α_k为第k次**步长因子**，则可得第$k+1$次迭代点$x_{k+1} = x_k + \alpha_k d_k$.

无约束优化问题的数值优化方法

- 无约束优化问题的一般算法框架

算法 1.1（无约束优化问题的一般算法框架）

步骤 0，给定初始化参数及初始迭代点 $\boldsymbol{x}_0$. 令 $k := 0$.

步骤 1，若 $\boldsymbol{x}_k$ 满足某种终止准则，停止迭代，以 $\boldsymbol{x}_k$ 作为近似极小点.

步骤 2，通过求解 $\boldsymbol{x}_k$ 处的某个子问题确定下降方向 $\boldsymbol{d}_k$.

步骤 3，通过某种搜索方式确定步长因子 α_k，使得 $f(\boldsymbol{x}_k+\alpha_k\boldsymbol{d}_k) < f(\boldsymbol{x}_k)$

步骤 4，令 $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$，$k := k+1$，转步骤 1.

- 称上述算法中的$\boldsymbol{s}_k = \alpha_k\boldsymbol{d}_k$为第$k$次迭代的**位移**.
- 设计不同的位移(不同搜索方向和步长因子)会产生不同的迭代算法. 为了保证收敛性，一般要求搜索方向为所谓的下降方向.

无约束优化问题的数值优化方法

- **定义 1.10** 若存在$\bar{\alpha} > 0$, 使得对任意的$\alpha \in (0, \bar{\alpha})$和$\boldsymbol{d}_k \neq \boldsymbol{0}$, 有$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$, 则称$\boldsymbol{d}_k$为$f(\boldsymbol{x})$在$\boldsymbol{x}_k$处的一个**下降方向**.

- **引理 1.1** 设函数$f: \mathcal{D} \subset \boldsymbol{R}^n \to \boldsymbol{R}$在开集$\mathcal{D}$上一阶连续可微. 则$\boldsymbol{d}_k$为$f(\boldsymbol{x})$在$\boldsymbol{x}_k$处的一个**下降方向**的充要条件是.

$$\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k < 0$$

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \alpha \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{d}_k + o(\alpha).$$

无约束优化问题的数值优化方法

- **定义 1.11** 若某算法只有当初始点x_0**充分接近**极小点x^*时, 由算法产生的点列$\{x_k\}$才收敛于x^*, 则称该算法具有**局部收敛性**. 若对于**任意的初始点**x_0, 由算法产生的点列$\{x_k\}$都收敛于x^*, 则称该算法具有**全局收敛性**.

- 算法的局部收敛速度是衡量一个算法好坏的重要指标, 通常有两种衡量收敛速度的尺度: **Q-收敛**和**R-收敛**.

无约束优化问题的数值优化方法

- Q-收敛

定义 1.12 设算法产生的点列 $\{\boldsymbol{x}_k\}$ 收敛于极小点 $\boldsymbol{x}^*$, 且

$$\limsup_{k\to\infty}\frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^p}\leqslant c.$$

(1) 若 $p=1$ 且 $0<c<1$, 则称该算法具有 Q-线性收敛速度 (或 Q-线性收敛的).

(2) 若 $p=1$ 且 $c=0$, 则称该算法具有 Q-超线性收敛速度 (或 Q-超线性收敛的).

(3) 若 $p=2$ 且 $0<c<\infty$, 则称该算法具有 Q-平方收敛速度 (或 Q-平方收敛的).

(4) 一般地, 若 $p>2$ 且 $0<c<\infty$, 则称该算法具有 Q-p 阶收敛速度 (或 Q-p 阶收敛的).

无约束优化问题的数值优化方法

- R–收敛

定义 1.13 设算法产生的点列 $\{\boldsymbol{x}_k\}$ 收敛于极小点 $\boldsymbol{x}^*$.

(1) 若存在常数 $c>0$ 和 $q\in(0,1)$ 使得

$$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\leqslant cq^k,$$

则称序列 $\{\boldsymbol{x}_k\}$ R-线性收敛到 $\boldsymbol{x}^*$.

(2) 若存在常数 $c>0$ 和收敛于零的正数列 $\{q_k\}$ 使得

$$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\leqslant c\prod_{i=0}^{k}q_i,$$

则称序列 $\{\boldsymbol{x}_k\}$ R-超线性收敛到 $\boldsymbol{x}^*$.

无约束优化问题的数值优化方法

- **迭代算法常用的终止条件：**

(1) 位移的绝对误差或相对误差充分小，即

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|<\varepsilon \quad \text{或} \quad \frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|}{\|\boldsymbol{x}_k\|}<\varepsilon,$$

式中：ε 为充分小的正数；

(2) 目标函数的绝对误差或相对误差充分小，即

$$|f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}_k)|<\varepsilon \quad \text{或} \quad \frac{|f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}_k)|}{|f(\boldsymbol{x}_k)|}<\varepsilon,$$

式中：ε 为充分小的正数；

(3) 目标函数的梯度的范数充分小，即

$$\|\nabla f(\boldsymbol{x}_k)\|<\varepsilon,$$

式中：ε 为充分小的正数.

线搜索方法

- 线搜索方法是求解无约束优化问题$\min\limits_{x\in R^n} f(x)$的一个最基本的方法. 此处介绍**一维线搜索**方法.

算法 1.1（无约束优化问题的一般算法框架）

步骤 0，给定初始化参数及初始迭代点 x_0. 令 $k := 0$.

步骤 1，若 x_k 满足某种终止准则，停止迭代，以 x_k 作为近似极小点.

步骤 2，通过求解 x_k 处的某个子问题确定下降方向 d_k.

步骤 3，通过某种搜索方式确定步长因子 α_k，使得 $f(x_k+\alpha_k d_k) < f(x_k)$

步骤 4，令 $x_{k+1} := x_k + \alpha_k d_k$，$k := k+1$，转步骤 1.

线搜索方法

- 线搜索方法是求解无约束优化问题$\min\limits_{x\in R^n} f(x)$的一个最基本的方法. 此处介绍**一维线搜索**方法.

步骤 3, 通过某种搜索方式确定步长因子 α_k, 使得 $f(x_k+\alpha_k d_k)<f(x_k)$

- 这实际上是将原优化问题转变为(n个变量)目标函数$f(x)$在一个规定方向上移动所形成的单变量优化问题, 也即所谓的“**线搜索**”或“**一维搜索**”. 令

$$\phi(\alpha)=f(x_k+\alpha d_k)$$

如此, **搜索步骤3**等价于求步长因子α_k使得$\phi(\alpha_k)<\phi(0)$.

线搜索方法

$$\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

求步长因子α_k使得$\phi(\alpha_k) < \phi(0)$.

- **精确线搜索:** 求α_k使目标函数f沿方向$\boldsymbol{d}_k$达到极小, 即

$$\phi(\alpha_k) = \min_{\alpha>0} \phi(\alpha)$$

若$f(x)$连续可微, 那么步长因子α_k具有如下性质:

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k = 0 \text{ (也即} \boldsymbol{g}_{k+1}^T \boldsymbol{d}_k = 0)$$

- **非精确线搜索:** 求α_k使目标函数f得到可接受的下降量, 即

$$\Delta f_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) > 0$$

是可接受的.

精确线搜索方法

- **精确线搜索基本思想**: 首先**确定**包含问题最优解的搜索**区间**, 然后采用某种插值或分割技术**缩小**这个**区间**, 进行搜索求解.
- **搜索区间定义**:

定义 2.1 设 ϕ 是定义在实数集上一元实函数, $\alpha^* \in [0, +\infty)$, 并且

$$\phi(\alpha^*) = \min_{\alpha \geqslant 0} \phi(\alpha). \tag{2.5}$$

若存在区间 $[a, b] \subset [0, +\infty)$, 使 $\alpha^* \in (a, b)$, 则称 $[a, b]$ 是极小化问题 (2.5) 的搜索区间. 进一步, 若 α^* 使得 $\phi(\alpha)$ 在 $[a, \alpha^*]$ 上严格递减, 在 $[\alpha^*, b]$ 上严格递增, 则称 $[a, b]$ 是 $\phi(\alpha)$ 的单峰区间, $\phi(\alpha)$ 是 $[a, b]$ 上的单峰函数.

精确线搜索方法

- **精确线搜索**：首先**确定**包含问题最优解的搜索**区间**，然后采用某种插值或分割技术**缩小**这个**区间**，进行搜索求解.

① 如何确定搜索区间并保证具有近似单峰性质？

- **进退法基本思想**：从一点出发，按一定步长，试图确定函数值呈现“高-低-高”的三点，从而得到一个近似的单峰区间.

精确线搜索方法

① 如何确定搜索区间并保证具有近似单峰性质?

- **进退法:**

算法 2.1 (确定近似单峰区间的进退法)

步骤 0, 选取 $\alpha_0 \geqslant 0$, $h_0 > 0$. 计算 $\phi_0 := \phi(\alpha_0)$. 令 $k := 0$.

步骤 1, 令 $\alpha_{k+1} := \alpha_k + h_k$, 计算 $\phi_{k+1} := \phi(\alpha_{k+1})$. 若 $\phi_{k+1} < \phi_k$, 转步骤 2; 否则, 转步骤 3.

步骤 2, 加大步长. 令 $h_{k+1} := 2h_k$, $\alpha := \alpha_k$, $\alpha_k := \alpha_{k+1}$, $\phi_k := \phi_{k+1}$, $k := k+1$, 转步骤 1.

步骤 3, 反向搜索或输出. 若 $k = 0$, 令 $h_1 := -h_0$, $\alpha := \alpha_1$, $\alpha_1 := \alpha_0$, $\phi_1 := \phi_0$, $k := 1$, 转步骤 1; 否则, 停止迭代, 令

$$a = \min\{\alpha, \alpha_{k+1}\}, \quad b = \max\{\alpha, \alpha_{k+1}\}.$$

输出 $[a, b]$.

α实际上相当于是α_{k-1}

精确线搜索方法

- **精确线搜索**：首先**确定**包含问题最优解的搜索**区间**，然后采用某种插值或分割技术**缩小**这个**区间**，进行搜索求解.

② 如何采用某种插值或分割技术缩小这个区间?

- **方法一般分为两类**：1) 使用导数的搜索，如插值法、牛顿法及**抛物线法**等；2) 不使用导数的搜索，如**黄金分割法**、分数法等.

- **黄金分割法基本思想**：通过试探点函数值的比较，使包含极值点的搜索区间不断缩小. **该方法仅需要计算函数值**.

- **抛物线法基本思想**：通过在搜索区间中不断地使用二次多项式去近似目标函数，并逐步用插值多项式的极小点去逼近线搜索问题的极小点. **该方法也不需要计算导数值**.

精确线搜索方法—黄金分割法

算法 2.2 (黄金分割法)

步骤 0, 确定初始搜索区间 $[a_0, b_0]$ 和容许误差 $0 \leqslant \varepsilon \ll 1$. 令

$t = (\sqrt{5}-1)/2$, 计算初始试探点

$p_0 = a_0 + (1-t)(b_0 - a_0)$, $q_0 = a_0 + t(b_0 - a_0)$

及相应的函数值 $\phi(p_0)$, $\phi(q_0)$. 令 $i := 0$.

步骤 1, 若 $\phi(p_i) \leqslant \phi(q_i)$, 转步骤 2; 否则, 转步骤 3.

步骤 2, 计算左试探点. 若 $|q_i - a_i| \leqslant \varepsilon$, 停算, 输出 p_i 否则, 令

$$a_{i+1} := a_i, \quad b_{i+1} := q_i, \quad \phi(q_{i+1}) := \phi(p_i),$$

$$q_{i+1} := p_i, \quad p_{i+1} := a_{i+1} + (1-t)(b_{i+1} - a_{i+1}).$$

计算 $\phi(p_{i+1})$, $i := i+1$, 转步骤 1.

步骤 3, 计算右试探点. 若 $|b_i - p_i| \leqslant \varepsilon$, 停算, 输出 q_i 否则, 令

$$a_{i+1} := p_i, \quad b_{i+1} := b_i, \quad \phi(p_{i+1}) := \phi(q_i),$$

$$p_{i+1} := q_i, \quad q_{i+1} := a_{i+1} + t(b_{i+1} - a_{i+1}).$$

计算 $\phi(q_{i+1})$, $i := i+1$, 转步骤 1.

Why? t为何取该值?

保持高低高结构

精确线搜索方法—黄金分割法

- 第i次迭代时，搜索区间为$[a_i, b_i]$. 取两个试探点为$p_i, q_i \in [a_i, b_i]$且$p_i < q_i$，并满足以下两个条件：
  - ✓ $[a_i, q_i]$与$[p_i, b_i]$的**长度相同**，即$b_i - p_i = q_i - a_i$
  - ✓ 区间长度的**缩短率相同**，即$b_{i+1} - a_{i+1} = t(b_i - a_i)$

黄金分割法每次迭代搜索区间搜索率是0.618，只是线性收敛，计算效率不高，但是**每次迭代只需计算一次函数值.**

$p_i = a_i + (1-t)(b_i - a_i), \quad q_i = a_i + t(b_i - a_i).$

考虑情形 $\phi(p_i) \leqslant \phi(q_i)$, 新的搜索区间为

$$[a_{i+1}, b_{i+1}] = [a_i, q_i].$$

$$q_{i+1} = a_{i+1} + t(b_{i+1} - a_{i+1}) = a_i + t(q_i - a_i)$$

$$= a_i + t^2(b_i - a_i).$$

若令 $t^2 = 1 - t, \quad t > 0,$

则 $q_{i+1} = a_i + (1-t)(b_i - a_i) = p_i.$

新的试探点 q_{i+1} 就不需要重新计算.

得区间长度缩短率为 $t = \dfrac{\sqrt{5}-1}{2} \approx 0.618.$

精确线搜索方法—抛物线法

- **抛物线法基本思想:** 通过在搜索区间中不断地使用二次多项式去近似目标函数, 并逐步用插值多项式的极小点去逼近线搜索问题的极小点.

算法 2.3 (抛物线法)

步骤 0, 由算法 2.1 确定三点 $s_0 < s_1 < s_2$, 对应的函数值 ϕ_0, ϕ_1, ϕ_2 满足 $\phi_1 < \phi_0, \quad \phi_1 < \phi_2$. 设定容许误差 $0 \leqslant \varepsilon \ll 1$.

步骤 1, 若 $|s_2 - s_0| < \varepsilon$, 停算, 输出 $s^* \approx s_1$.

步骤 2, 计算插值点. 根据式 (2.9) 计算 $\bar{s}$ 和 $\bar{\phi} := \phi(\bar{s})$. 若 $\phi_1 \leqslant \bar{\phi}$, 转步骤 4; 否则, 转步骤 3.

步骤 3, 若 $s_1 > \bar{s}$, 则 $s_2 := s_1$, $s_1 := \bar{s}$, $\phi_2 := \phi_1$, $\phi_1 := \bar{\phi}$, 转步骤 1; 否则, $s_0 := s_1$, $s_1 := \bar{s}$, $\phi_0 := \phi_1$, $\phi_1 := \bar{\phi}$, 转步骤 1.

步骤 4, 若 $s_1 < \bar{s}$, 则 $s_2 := \bar{s}$, $\phi_2 := \bar{\phi}$, 转步骤 1; 否则, $s_0 := \bar{s}$, $\phi_0 := \bar{\phi}$, 转步骤 1.

$$\bar{s} = s_0 + \frac{(3\phi_0 - 4\phi_1 + \phi_2)h}{2(\phi_0 - 2\phi_1 + \phi_2)}$$

精确线搜索方法—抛物线法

- 抛物线法**原理解释**: 1)如何得到$\bar{s}$？ 2)搜索的最优性保证?

已知三点 $s_0,\ s_1 = s_0 + h,\ s_2 = s_0 + 2h\ \ (h > 0)$ 处的函数值 ϕ_0, ϕ_1, ϕ_2, 且满足

$\phi_1 < \phi_0,\quad \phi_1 < \phi_2$. 保证了函数 ϕ 在区间 $[s_0, s_2]$ 上是单峰函数. 则三点间 二次拉格朗日插值多项式为

$$q(s) = \frac{(s-s_1)(s-s_2)}{2h^2}\phi_0 - \frac{(s-s_0)(s-s_2)}{h^2}\phi_1 + \frac{(s-s_0)(s-s_1)}{2h^2}\phi_2.$$

拟合三点间的二次曲线

$q(s)$ 的一阶导数为

$$q'(s) = \frac{2s-s_1-s_2}{2h^2}\phi_0 - \frac{2s-s_0-s_2}{h^2}\phi_1 + \frac{2s-s_0-s_1}{2h^2}\phi_2.$$

问：为何不直接解$\phi'(s) = 0$?

令 $q'(s) = 0$, 解得

$$\begin{aligned}\bar{s} &= \frac{(s_1+s_2)\phi_0 - 2(s_0+s_2)\phi_1 + (s_0+s_1)\phi_2}{2(\phi_0 - 2\phi_1 + \phi_2)} \\ &= \frac{(2s_0+3h)\phi_0 - 2(2s_0+2h)\phi_1 + (2s_0+h)\phi_2}{2(\phi_0 - 2\phi_1 + \phi_2)} \\ &= s_0 + \frac{(3\phi_0 - 4\phi_1 + \phi_2)h}{2(\phi_0 - 2\phi_1 + \phi_2)} := s_0 + \bar{h}, \qquad (2.9)\end{aligned}$$

求拟合曲线的极小值

凸二次连续可微, 保证最优性

因 $q(s)$ 的二阶导数为

$$q''(s) = \frac{\phi_0}{h^2} - \frac{2\phi_1}{h^2} + \frac{\phi_2}{h^2} = \frac{\phi_0 - 2\phi_1 + \phi_2}{h^2} > 0,$$

故 $q(s)$ 为凸二次函数.

$s_{\min}$ 是 $q(s)$ 的全局极小点.

非精确线搜索方法

- **精确线搜索缺点**: 往往需要计算很多的函数值和梯度值, 特别当迭代点远离最优点时, 不是十分有效合理.
- **非精确线搜索**: 既能保证目标函数具有可接受的下降量, 又能使最终形成的迭代序列收敛.
  - ✓ Wolfe准则
  - ✓ Armijo准则

非精确线搜索方法—Wolfe准则

- **Wolfe准则**是指给定$\rho \in (0,0.5), \sigma \in (\rho, 1)$, 求使得下面两个不等式同时成立:

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \leq f(\boldsymbol{x}_k) + \rho \alpha_k \boldsymbol{g}_k^T \boldsymbol{d}_k$$

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k \geq \sigma \boldsymbol{g}_k^T \boldsymbol{d}_k$$

其中: $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_k)$

有时用更强的条件代替: (**强Wolfe准则**)

$$|\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k| \leq -\sigma \boldsymbol{g}_k^T \boldsymbol{d}_k$$

$\sigma > 0$充分小时, 可使其变为近似精确线搜索.

由该Wolfe准则得到的新的迭代点$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$在$\boldsymbol{x}_k$的某一领域内并且使目标函数值有一定的下降量.

由于$\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$, 有定理可以证明Wolfe准则的有限终止性.

非精确线搜索方法——Armijo准则

- **Armijo准则**是指给定$\beta \in (0,1), \sigma \in (0,0.5)$. 令步长因子$\alpha_k = \beta^{m_k}$, 其中$m_k$为满足下列不等式的最小非负整数:

$$f(\boldsymbol{x}_k + \beta^m \boldsymbol{d}_k) \leq f(\boldsymbol{x}_k) + \sigma \beta^m \boldsymbol{g}_k^T \boldsymbol{d}_k$$

可以证明, 若$f(\boldsymbol{x})$是连续可微的且满足$\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$, 则Armijo准则是有限终止的, 即存在正数$\sigma$, 使得对于充分大的正整数$m$, 上式成立.

算法 2.4 (Armijo 准则)

给定参数 $\beta \in (0,1)$, $\sigma \in (0,0.5)$. 对 $m = 0, 1, \cdots$, 若不等式

$$f(\boldsymbol{x}_k + \beta^m \boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + \sigma \beta^m \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k$$

成立, 则置 $m_k := m$, $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \beta^{m_k} \boldsymbol{d}_k$, 第 k 步线搜索完成.

非精确线搜索方法——Armijo准则

- Armijo搜索的Matlab程序:

```
function mk=armijo(xk,dk)
beta=0.5;  sigma=0.2;
m=0;  maxm=20;
while (m<=maxm)
    if(fun(xk+beta^m*dk)<=fun(xk)+sigma*beta^m*gfun(xk)'*dk)
        mk=m;  break;
    end
    m=m+1;
end
alpha=beta^mk
newxk=xk+alpha*dk
fk=fun(xk)
newfk=fun(newxk)
```

fun: 目标函数

gfun: 目标函数的梯度函数

线搜索法总结

- **线搜索法:** 指用**线搜索策略求步长因子**的无约束优化问题下降类算法的简称, 一般算法框架如下.

算法 2.5 (线搜索法算法框架)

步骤 0, 初始化. 选取有关参数及初始迭代点 $\boldsymbol{x}_0 \in \mathbf{R}^n$. 设定容许误差 $0 \leqslant \varepsilon \ll 1$. 令 $k := 0$.

步骤 1, 检验终止判别准则. 计算 $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. 若 $\|\boldsymbol{g}_k\| \leqslant \varepsilon$, 停算, 输出 $\boldsymbol{x}^* \approx \boldsymbol{x}_k$.

步骤 2, 确定下降方向 $\boldsymbol{d}_k$, 使满足 $\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k < 0$.

步骤 3, 确定步长因子 α_k. 可在下列“精确”与“非精确”两种线搜索策略中选用其一:

(1) 用黄金分割法或抛物线法等精确线搜索策略求

$$\alpha_k = \arg\min_{\alpha>0} f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k);$$

(2) 用 Wolfe 准则或 Armijo 准则等非精确线搜索策略求 α_k

步骤 4, 更新迭代点. 令 $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, $k := k+1$, 转步骤 1.

梯度法

- **梯度法(最速下降法)**是求解无约束优化问题最简单最古老的方法之一. 前述下降类算法的一般框架中提到用不同的方式确定搜索方向或搜索步长, 就会得到不同算法. 梯度法是用负梯度方向

$$\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$$

作为搜索方向的.

- **两个问题:**
  - ✓ 为什么选择负梯度方向?
  - ✓ 为什么称为最速下降法?

梯度法

1）为什么选择负梯度方向？ 2）为什么称为最速下降法？

搜索步长

设 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}_k$ 附近连续可微，$\boldsymbol{d}_k$ 为搜索方向向量，$\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. 泰勒展开得

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \alpha \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + o(\alpha), \quad \alpha > 0.$$

那么目标函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}_k$ 处沿方向 $\boldsymbol{d}_k$ 下降的变化率为

$$\lim_{\alpha \to 0} \frac{f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) - f(\boldsymbol{x}_k)}{\alpha} = \lim_{\alpha \to 0} \frac{\alpha \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k + o(\alpha)}{\alpha}$$

$$= \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k = \|\boldsymbol{g}_k\| \|\boldsymbol{d}_k\| \cos \bar{\theta}_k,$$

$\bar{\theta}_k$ 为 $\boldsymbol{g}_k$ 与 $\boldsymbol{d}_k$ 的夹角.

小于0

也即搜索方向为负梯度方向，是目标函数在当前点的最速下降方向

要使变化率最小，也即变化率绝对值最大，只有 $cos\bar{\theta}_k = -1$，即 $\bar{\theta}_k = \pi$

梯度法

- 梯度法计算步骤:

用精确线搜索方法

$$\phi'(\alpha)=\frac{\mathrm{d}}{\mathrm{d}\alpha}f(\boldsymbol{x}_k+\alpha\boldsymbol{d}_k)|_{\alpha=\alpha_k}=\nabla f(\boldsymbol{x}_k+\alpha_k\boldsymbol{d}_k)^{\mathrm{T}}\boldsymbol{d}_k=0.$$

有 $$\boldsymbol{g}(\boldsymbol{x}_{k+1})^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}_k)=0,$$

即新点 $\boldsymbol{x}_{k+1}$ 处的梯度与旧点 $\boldsymbol{x}_k$ 处的梯度是正交的

算法 3.1 (梯度法)

步骤 0, 选取初始点 $\boldsymbol{x}_0\in\mathbf{R}^n$, 容许误差 $0\leqslant\varepsilon\ll 1$. 令 $k:=0$.

步骤 1, 计算 $\boldsymbol{g}_k=\nabla f(\boldsymbol{x}_k)$. 若 $\|\boldsymbol{g}_k\|\leqslant\varepsilon$, 停算, 输出 $\boldsymbol{x}_k$ 作为近似极小点.

步骤 2, 取方向 $\boldsymbol{d}_k=-\boldsymbol{g}_k$.

步骤 3, 由线搜索方法确定步长因子 α_k.

步骤 4, 令 $\boldsymbol{x}_{k+1}:=\boldsymbol{x}_k+\alpha_k\boldsymbol{d}_k$, $k:=k+1$, 转步骤 1.

可用精确/非精确线性搜索方法,在理论上均能保证全局收敛性.

梯度法

- 一些定理(了解)

定理 3.1 设目标函数 $f(\boldsymbol{x})$ 连续可微且其梯度函数 $\boldsymbol{g}(\boldsymbol{x})$ 是 Lipschitz 连续的, $\{\boldsymbol{x}_k\}$ 由梯度法产生, 其中步长因子 α_k 由精确线搜索, 或由 Wolfe 准则, 或由 Armijo 准则产生, 则有

$$\lim_{k\to\infty} \|\boldsymbol{g}(\boldsymbol{x}_k)\| = 0.$$

定理 3.2 设矩阵 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$ 对称正定, $\boldsymbol{b} \in \mathbf{R}^n$. 记 λ_1 和 λ_n 分别是 $\boldsymbol{A}$ 的最大和最小特征值, $\kappa = \lambda_1/\lambda_n$. 考虑如下极小化问题

$$\min f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}.$$

设 $\{\boldsymbol{x}_k\}$ 是用精确线搜索的梯度法求解上述问题所产生的迭代序列, 则对于所有的 k, 下面的不等式成立

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_A \leqslant \left(\frac{\kappa-1}{\kappa+1}\right)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_A,$$

式中: $\boldsymbol{x}^*$ 为问题的唯一解, $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}$.

κ接近1($\boldsymbol{A}$的最大和最小特征值接近时), 梯度法收敛很快.κ较大时($\boldsymbol{A}$近似于病态), 收敛很慢

牛顿法

- **牛顿法基本思想:** 用迭代点$\boldsymbol{x}_k$处的一阶导数(**梯度**)和二阶导数(Hesse**矩阵**)对目标函数进行二次函数近似, 然后把二次函数的极小点作为新的迭代点, 不断重复这一过程, 直至求得满足精度的近似极小点.

设 $f(\boldsymbol{x})$ 的 Hesse 阵 $\boldsymbol{G}(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ 连续, 取其在 $\boldsymbol{x}_k$ 处的泰勒展开式的前三项, 得

$$q_k(\boldsymbol{x}) = f_k + \boldsymbol{g}_k^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{G}_k(\boldsymbol{x} - \boldsymbol{x}_k),$$

求二次函数 $q_k(\boldsymbol{x})$ 的稳定点, 得

$$\nabla q_k(\boldsymbol{x}) = \boldsymbol{g}_k + \boldsymbol{G}_k(\boldsymbol{x} - \boldsymbol{x}_k) = \boldsymbol{0}.$$

若 $\boldsymbol{G}_k$ 非奇异, 那么解上面的线性方程组 得

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{G}_k^{-1}\boldsymbol{g}_k.$$

实际运算中为避免求逆, 可先解$\boldsymbol{G}_k\boldsymbol{d} = -\boldsymbol{g}_k$得$\boldsymbol{d}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, 然后令$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$

牛顿法

- **基本牛顿法优点:** 收敛速度快, 具有局部二阶收敛性

算法 3.2 (基本牛顿法)

步骤 0, 选取初始点 $\boldsymbol{x}_0 \in \mathbf{R}^n$, 容许误差 $0 \leqslant \varepsilon \ll 1$. 令 $k := 0$.

步骤 1, 计算 $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. 若 $\|\boldsymbol{g}_k\| \leqslant \varepsilon$, 停算, 输出 $\boldsymbol{x}^* \approx \boldsymbol{x}_k$.

步骤 2, 计算 $\boldsymbol{G}_k = \nabla^2 f(\boldsymbol{x}_k)$, 并求解线性方程组

$$\boldsymbol{G}_k \boldsymbol{d} = -\boldsymbol{g}_k,$$

得解 $\boldsymbol{d}_k$.

步骤 3, 令 $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \boldsymbol{d}_k$, $k := k + 1$, 转步骤 1.

牛顿法

- **基本牛顿法缺点:** 初始点需要**足够靠近极小点**, 否则算法有**可能不收敛**. 但通常精确极小点位置未知, 故而带来困难. 可引入线搜索方法以得到大范围收敛的算法→**阻尼牛顿法**

定理 3.3 设函数 $f(\boldsymbol{x})$ 有二阶连续偏导数, 在局部极小点 $\boldsymbol{x}^*$ 处, $\boldsymbol{G}(\boldsymbol{x}^*) = \nabla^2 f(\boldsymbol{x}^*)$ 是正定的且 $\boldsymbol{G}(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 的一个邻域内是 Lipschitz 连续的. 如果初始点 $\boldsymbol{x}_0$ 充分靠近 $\boldsymbol{x}^*$, 那么对一切 k, 牛顿迭代公式 (3.4) 是适定的, 并且当 $\{\boldsymbol{x}_k\}$ 为无穷点列时, 其极限为 $\boldsymbol{x}^*$ 且收敛阶至少是二阶的.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{G}_k^{-1}\boldsymbol{g}_k.$$

牛顿法

- 基于Armijo搜索的阻尼牛顿法:

算法 3.3 (基于 Armijo 搜索的阻尼牛顿法)

步骤 0, 选取参数 $\beta \in (0,1)$, $\sigma \in (0, 0.5)$, 初始点 $\boldsymbol{x}_0 \in \mathbf{R}^n$, 容许误差 $0 \leqslant \varepsilon \ll 1$. 令 $k := 0$.

步骤 1, 计算 $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. 若 $\|\boldsymbol{g}_k\| \leqslant \varepsilon$, 停算, 输出 $\boldsymbol{x}^* \approx \boldsymbol{x}_k$.

步骤 2, 计算 $\boldsymbol{G}_k = \nabla^2 f(\boldsymbol{x}_k)$, 并求解线性方程组

$$\boldsymbol{G}_k \boldsymbol{d} = -\boldsymbol{g}_k,$$

牛顿法求搜索方向

得解 $\boldsymbol{d}_k$.

步骤 3, 记 m_k 是满足下列不等式的最小非负整数 m:

$$f(\boldsymbol{x}_k + \beta^m \boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + \sigma \beta^m \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k.$$

Armijo法则搜索步长

步骤 4, 令 $\alpha_k := \beta^{m_k}$, $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, $k := k + 1$, 转步骤 1.

牛顿法

- 基于Armijo搜索的阻尼牛顿法(理论支撑，了解)

定理 3.4 设函数 $f(\boldsymbol{x})$ 二次连续可微且存在常数 $\gamma > 0$，使得

$$\boldsymbol{d}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x})\boldsymbol{d} \geqslant \gamma\|\boldsymbol{d}\|^2, \ \forall\, \boldsymbol{d} \in \mathbf{R}^n, \boldsymbol{x} \in \mathcal{L}(\boldsymbol{x}_0), \tag{3.8}$$

式中：$\mathcal{L}(\boldsymbol{x}_0) = \{\boldsymbol{x} | f(\boldsymbol{x}) \leqslant f(\boldsymbol{x}_0)\}$. 设 $\{\boldsymbol{x}_k\}$ 是由算法 3.3 产生的无穷点列，则该点列收敛于 f 在水平集 $\mathcal{L}(\boldsymbol{x}_0)$ 中的唯一全局极小点.

引理 3.1 设函数 $f : \mathbf{R}^n \to \mathbf{R}$ 二次连续可微，点列 $\{\boldsymbol{x}_k\}$ 由算法 3.3 产生. 设 $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^*$ 且 $\boldsymbol{g}(\boldsymbol{x}^*) = \mathbf{0}$, $\boldsymbol{G}(\boldsymbol{x}^*)$ 正定. 那么，若

$$\lim_{k\to\infty} \frac{\|\boldsymbol{G}(\boldsymbol{x}_k)\boldsymbol{d}_k + \boldsymbol{g}_k\|}{\|\boldsymbol{d}_k\|} = 0, \tag{3.10}$$

则 (1) 当 k 充分大时，步长因子 $\alpha_k \equiv 1$;

(2) 点列 $\{\boldsymbol{x}_k\}$ 超线性收敛于 $\boldsymbol{x}^*$.

定理 3.5 设定理 3.4 的条件成立，点列 $\{\boldsymbol{x}_k\}$ 由算法 3.3 产生，则 $\{\boldsymbol{x}_k\}$ 超线性收敛于 f 的全局极小点 $\boldsymbol{x}^*$. 此外，若 Hesse 阵 $\boldsymbol{G}(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处 Lipschitz 连续，则收敛阶至少是二阶的.

修正牛顿法

- **牛顿法的优点:** 不低于二阶的收敛速度
- **牛顿法的缺点:** 要求目标函数的Hesse矩阵$\boldsymbol{G}(\boldsymbol{x})$在每个迭代点$\boldsymbol{x}_k$处是**正定**的, 否则难以保证牛顿方向$\boldsymbol{d}_k = -\boldsymbol{G}_k^{-1}\boldsymbol{g}_k$是$f$在$\boldsymbol{x}_k$处的下降方向.

- **修正方法之一:** 和梯度法结合起来, 构造“**牛顿-梯度混合算法**”: 当Hesse阵正定, 用牛顿方向作为搜索方向; 否则, 若Hesse阵奇异, 或非奇异但牛顿方向不是下降方向, 采用负梯度方向作为搜索方向.

修正牛顿法--牛顿-梯度混合算法

算法 3.4 (牛顿 - 梯度混合算法)

步骤 0, 选取初始点 $\boldsymbol{x}_0 \in \mathbf{R}^n$, 容许误差 $0 \leqslant \varepsilon \ll 1$. 令 $k := 0$.

步骤 1, 计算 $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. 若 $\|\boldsymbol{g}_k\| \leqslant \varepsilon$, 停算, 输出 $\boldsymbol{x}_k$ 作为近似极小点.

步骤 2, 计算 $\boldsymbol{G}_k = \nabla^2 f(\boldsymbol{x}_k)$, 并求解线性方程组

$$\boldsymbol{G}_k \boldsymbol{d} = -\boldsymbol{g}_k.$$

若方程组有解 $\boldsymbol{d}_k$ 且满足 $\boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k < 0$, 转步骤 3;

否则, 令 $\boldsymbol{d}_k = -\boldsymbol{g}_k$, 转步骤 3.

步骤 3, 由线搜索方法确定步长因子 α_k.

步骤 4, 令 $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, $k := k+1$, 转步骤 1.

修正牛顿法

- **牛顿法的优点:** 不低于二阶的收敛速度
- **牛顿法的缺点:** 要求目标函数的Hesse矩阵$\boldsymbol{G}(\boldsymbol{x})$在每个迭代点$\boldsymbol{x}_k$处是**正定**的, 否则难以保证牛顿方向$\boldsymbol{d}_k = -\boldsymbol{G}_k^{-1}\boldsymbol{g}_k$是$f$在$\boldsymbol{x}_k$处的下降方向.

- **修正方法之二**: 引入阻尼因子$\mu_k \geq 0$, 即在每一迭代步适当选取参数μ_k使得矩阵$\boldsymbol{A}_k = \boldsymbol{G}(\boldsymbol{x}_k) + \mu_k \boldsymbol{I}$正定.

修正牛顿法

算法 3.5 (修正牛顿法)

步骤 0, 选取参数 $\beta \in (0,1)$, $\sigma \in (0,0.5)$, 初始点 $\boldsymbol{x}_0 \in \mathbf{R}^n$, 容许误差 $0 \leqslant \varepsilon \ll 1$, 参数 $\tau \in [0,1]$. 令 $k := 0$.

步骤 1, 计算 $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$, $\mu_k = \|\boldsymbol{g}_k\|^{1+\tau}$. 若 $\|\boldsymbol{g}_k\| \leqslant \varepsilon$, 停算, 输出 $\boldsymbol{x}_k$ 作为近似极小点.

步骤 2, 计算 Hesse 阵 $\boldsymbol{G}_k = \nabla^2 f(\boldsymbol{x}_k)$, 并求解线性方程组

$$(\boldsymbol{G}_k + \mu_k \boldsymbol{I})\boldsymbol{d} = -\boldsymbol{g}_k,$$

牛顿法求搜索方向

得解 $\boldsymbol{d}_k$.

步骤 3, 令 m_k 是满足下列不等式的最小非负整数 m:

$$f(\boldsymbol{x}_k + \beta^m \boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + \sigma \beta^m \boldsymbol{g}_k^{\mathrm{T}} \boldsymbol{d}_k.$$

Armijo法则搜索步长

步骤 4, 令 $\alpha_k := \beta^{m_k}$, $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, $k := k+1$, 转步骤 1.