最优化方法及其 Matlab 程序设计

马昌凤

2010 年 3 月 22 日

目 录

第一章 最优化理论基础

1.1 最优化问题的数学模型

通俗地说，所谓最优化问题，就是求一个多元函数在某个给定集合上的极值. 几乎所有类型的最优化问题都可以用下面的数学模型来描述:

$$\begin{aligned} &\min f(x), \\ &\text{s.t. } \ x \in K, \end{aligned} \tag{1.1}$$

这里，K 是某个给定的集合 (称为可行集或可行域)，$f(x)$ 是定义在集合 K 上的实值函数. 此外，在模型 (1.1) 中，x 通常称为决策变量, s.t. 是 subject to (受限于) 的缩写.

人们通常按照可行集的性质对最优化问题 (1.1) 进行一个大致的分类:

- 线性规划和非线性规划. — 可行集是有限维空间中的一个子集;
- 组合优化或网络规划. — 可行集中的元素是有限的;
- 动态规划. — 可行集是一个依赖时间的决策序列;
- 最优控制. — 可行集是无穷维空间中的一个连续子集.

本书主要考虑在工程设计中有着重要应用的非线性规划，其数学模型为

$$\min \quad f(x), \tag{1.2}$$

$$\text{s.t.} \quad h_i(x) = 0,\ i = 1, \cdots, l, \tag{1.3}$$

$$g_i(x) \geq 0,\ i = 1, \cdots, m, \tag{1.4}$$

其中，$f(x)$, $h_i(x)$ $(i = 1, \cdots, l)$ 及 $g_i(x)$ $(i = 1, \cdots, m)$ 都是定义在 $\mathbb{R}^n$ 上

连续可微的多元实值函数, 且至少有一个是非线性的. 记

$$E = \{i : h_i(x) = 0\}, \quad I = \{i : g_i(x) \geq 0\}. \tag{1.5}$$

若指标集 $E \cup I = \emptyset$, 称之为无约束优化问题，否则称为约束优化问题. 特别地, 把 $E \neq \emptyset$ 且 $I = \emptyset$ 的优化问题称为等式约束优化问题; 而把 $I \neq \emptyset$ 且 $E = \emptyset$ 的优化问题称为不等式约束优化问题. $f(x)$ 称为目标函数, $h_i(x)$, $g_j(x)$ $(i = 1, \cdots, l; j = 1, \cdots, m)$ 称为约束函数. 此外，通常把目标函数为二次函数而约束函数都是线性函数的优化问题称为二次规划；而目标函数和约束函数都是线性函数的优化问题称为线性规划.

1.2 向量和矩阵范数

在算法的收敛性分析中，需要用到向量和矩阵范数的概念及其有关理论.

设 $\mathbb{R}^n$ 表示实 n 维向量空间，$\mathbb{R}^{n\times n}$ 表示实 n 阶矩阵全体所组成的线性空间. 在这两个空间中，我们分别定义向量和矩阵的范数.

向量 $x \in \mathbb{R}^n$ 的范数 $\|x\|$ 是一个非负数，它必须满足以下条件：

（1）$\|x\| \geq 0$，$\|x\| = 0 \Longleftrightarrow x = 0$;

（2）$\|\lambda x\| = |\lambda|\|x\|, \lambda \in \mathbb{R}$;

（3）$\|x + y\| \leq \|x\| + \|y\|$.

向量 $x = (x_1, \cdots, x_n)^T$ 的 p-范数定义为

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{\frac{1}{p}}. \tag{1.6}$$

常用的向量范数有

1-范数：$\|x\|_1 = \sum_{i=1}^{n} |x_i|$;

2-范数：$\|x\|_2 = \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{\frac{1}{2}}$;

∞-范数：$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

矩阵 $A \in \mathbb{R}^{n \times n}$ 的范数是一个非负实数，它除了满足跟向量范数相似的三条性质之外，还必须具备乘法性质：

（4） $\|AB\| \le \|A\|\|B\|, \quad A, B \in \mathbb{R}^{n\times n}.$

如果一矩阵范数 $\|\cdot\|_\mu$ 相对于某向量范数 $\|\cdot\|$ 满足下面的不等式

（5） $\|Ax\| \le \|A\|_\mu \|x\|, \quad x \in \mathbb{R}^n,$

则称矩阵范数 $\|\cdot\|_\mu$ 和向量范数 $\|\cdot\|$ 是相容的. 进一步，若存在 $x \neq 0$ 使成立

$$\|A\|_\mu = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}, \tag{1.7}$$

则称矩阵范数 $\|\cdot\|_\mu$ 是由向量范数 $\|\cdot\|$ 诱导出来的算子范数，简称算子范数，有时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 $\|\cdot\|$.

不难验证，从属于向量范数 $\|x\|_\infty$, $\|x\|_1$, $\|x\|_2$ 的矩阵范数分别为

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

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

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

它们分别称作行和范数、列和范数和谱范数.

本书在讨论各种迭代算法的收敛性时，通常采用谱范数和按下述方式定义的 F-范数:

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \sqrt{\operatorname{tr}(A^T A)} \tag{1.8}$$

现在我们来讨论向量序列和矩阵序列的收敛性. 我们知道，若 $\{x^{(k)}\}_{k=1}^{\infty} \subset \mathbb{R}^n$，则

$$\lim_{k\to\infty} x^{(k)} = x \Longleftrightarrow \lim_{k\to\infty} x_i^{(k)} = x_i, \quad i = 1, \cdots, n.$$

类似地，若 $\{A^{(k)}\}_{k=1}^{\infty} \subset R^{n\times n}$，则

$$\lim_{k\to\infty} A^{(k)} = A \Longleftrightarrow \lim_{k\to\infty} a_{ij}^{(k)} = a_{ij}, \quad i,\ j = 1, \cdots, n.$$

为了利用范数来描述上述极限，必须建立向量范数的等价定理以及矩阵范数的等价定理.

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

$$c_1\|x\| \leq \|x\|' \leq c_2\|x\|.$$

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

$$m_1\|A\| \leq \|A\|' \leq m_2\|A\|.$$

下面，我们利用范数的概念来等价地定义向量序列和矩阵序列的收敛性.

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

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

(2) 设 $\{A^{(k)}\}$ 为 $n \times n$ 矩阵序列，$\|\cdot\|$ 为定义在 $\mathbb{R}^{n\times n}$ 上的向量范数，

则

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

1.3 函数的可微性与展开

本节主要介绍后文经常需要用到 n 元函数的一阶和二阶导数以及泰勒展开式.

定义 1.1 设有 n 元实函数 $f(x)$, 其中自变量 $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$. 称向量

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_n} \right)^T \tag{1.9}$$

为 $f(x)$ 在 x 处的一阶导数或梯度. 称矩阵

$$\nabla^2 f(x)=\begin{pmatrix} \dfrac{\partial^2 f(x)}{\partial x_1^2} & \dfrac{\partial^2 f(x)}{\partial x_1\partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_1\partial x_n} \\ \dfrac{\partial^2 f(x)}{\partial x_2\partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_2\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f(x)}{\partial x_n\partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_n\partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix} \tag{1.10}$$

为 $f(x)$ 在 x 处的二阶导数或 Hesse 矩阵. 若梯度 $\nabla f(x)$ 的每个分量函数在 x 都连续, 则称 f 在 x 一阶连续可微; 若 Hesse 阵 $\nabla^2 f(x)$ 的各个分量函数都连续, 则称 f 在 x 二阶连续可微.

若 f 在开集 D 的每一点都连续可微, 则称 f 在 D 上一阶连续可微; 若 f 在开集 D 的每一点都都二阶连续可微, 则称 f 在 D 上二阶连续可微.

由上述定义不难发现，若 f 在 x 二阶连续可微，则

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

即 Hesse 阵 $\nabla^2 f(x)$ 是对称阵.

例 1.1 设二次函数

$$f(x) = c^T x + \frac{1}{2} x^T H x,$$

其中 $c \in \mathbb{R}^n$, $H \in \mathrm{R}^{n\times n}$ 是对称阵. 那么，不难计算它在 x 的梯度及 Hesse 为

$$\nabla f(x) = c + Hx, \quad \nabla^2 f(x) = H.$$

例 1.2 (泰勒展开). 设函数 $f : \mathbb{R}^n \to \mathbb{R}$ 连续可微，那么

$$\begin{aligned} f(x+h) &= f(x) + \int_0^1 \nabla f(x+\tau h)^T h \mathrm{d}\tau \\ &= f(x) + \nabla f(x+\xi h)^T h, \quad \xi \in (0,1) \\ &= f(x) + \nabla f(x)^T h + o(\|h\|). \end{aligned}$$

进一步, 若函数 f 是二次连续可微的, 则有

$$
\begin{aligned}
f(x+h) &= f(x)+\nabla f(x)^T h+\int_0^1(1-\tau)h^T\nabla^2 f(x+\tau h)h\mathrm{d}\tau \\
&= f(x)+\nabla f(x)^T h+\frac{1}{2}h^T\nabla^2 f(x+\xi h)h, \quad \xi\in(0,1) \\
&= f(x)+\nabla f(x)^T h+\frac{1}{2}h^T\nabla^2 f(x)h+o(\|h\|^2)
\end{aligned}
$$

及

$$
\begin{aligned}
\nabla f(x+h) &= \nabla f(x)+\int_0^1\nabla^2 f(x+\tau h)^T h\mathrm{d}\tau \\
&= \nabla f(x)+\nabla^2 f(x+\xi h)^T h, \quad \xi\in(0,1) \\
&= \nabla f(x)+\nabla^2 f(x)^T h+o(\|h\|).
\end{aligned}
$$

下面简单介绍一下向量值函数的可微性及中值定理. 设有向量值函数 $F=(F_1,F_2,\cdots,F_m)^T:\mathbb{R}^n\to\mathbb{R}^m$. 若每个分量函数 F_i 都是 (连续) 可微的，则称 F 是 (连续) 可微的. 向量值函数 F 在 x 的导数 $F'\in\mathbb{R}^{m\times n}$

是指它在 x 的 Jacobi 矩阵, 记为 $F'(x)$ 或 $J_F(x)$, 即

$$F'(x) := J_F(x) := \begin{pmatrix} \dfrac{\partial F_1(x)}{\partial x_1} & \dfrac{\partial F_1(x)}{\partial x_2} & \cdots & \dfrac{\partial F_1(x)}{\partial x_n} \\ \dfrac{\partial F_2(x)}{\partial x_1} & \dfrac{\partial F_2(x)}{\partial x_2} & \cdots & \dfrac{\partial F_2(x)}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial F_m(x)}{\partial x_1} & \dfrac{\partial F_m(x)}{\partial x_2} & \cdots & \dfrac{\partial F_m(x)}{\partial x_n} \end{pmatrix}.$$

考虑到标量函数的梯度定义, 有时也把向量函数 F 的 Jacobi 矩阵的转置称为 F 在 x 的梯度，记为

$$\nabla F(x) = J_F(x)^T = \big(\nabla F_1(x), \nabla F_2(x), \cdots, \nabla F_m(x)\big).$$

不难发现，例 1.2 中关于多元实函数的一些结论可以推广到向量值函数的情形. 例如，若向量值函数 $F : \mathbb{R}^n \to \mathbb{R}^m$ 是连续可微的，则对于任意

的 $x,\ h \in \mathbb{R}^n$, 有

$$F(x+h) = F(x) + \int_0^1 \nabla F(x+\tau h)^T h \mathrm{d}\tau.$$

对于向量值函数 F, 也可以定义 Lipschitz 连续性的概念.

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

$$\|F(x) - F(y)\| \leq L\|x - y\|, \tag{1.11}$$

其中，L 称为 Lipschitz 常数. 若 (1.11) 式对任意的 $x,\ y \in \mathbb{R}^n$ 都成立, 则称 F 在 $\mathbb{R}^n$ 内是 Lipschitz 连续的.

在迭代法的收敛性分析中, 有时需要用到向量值函数的“中值定理”, 现引述如下.

定理 1.3 设向量值函数 $F : \mathbb{R}^n \to \mathbb{R}^m$ 连续可微, 那么

(1) 对任意的 $x, y \in \mathbb{R}^n$, 有

$$\|F(x)-F(y)\| \le \sup_{0\le t\le 1} \|F'(y+t(x-y))\|\,\|x-y\|.$$

(2) 对任意的 $x, y, z \in \mathbb{R}^n$, 有

$$\begin{aligned}&\|F(y)-F(z)-F'(x)(y-z)\|\\ \le&\sup_{0\le t\le 1} \|F'(z+t(y-z))-F'(x)\|\,\|x-y\|.\end{aligned}$$

由上述定理的结论 (2) 可推得下面的结论.

推论 1.1 设向量值函数 $F:\mathbb{R}^n \to \mathbb{R}^m$ 是连续可微的, 且其 Jacobi 矩阵映射是 Lipschitz 连续的, 即存在常数 $L>0$ 使得

$$\|F'(u)-F'(v)\| \le L\|u-v\|, \quad \forall\, u, v \in \mathbb{R}^n. \tag{1.12}$$

则对任意的 $x, h \in \mathbb{R}^n$, 有

$$\|F(x+h)-F(x)-F'(x)h\| \le \frac{1}{2}L\|h\|^2. \tag{1.13}$$

1.4 凸集与凸函数

本节介绍凸集、锥和多面体的有关概念.

定义 1.3 设集合 $D \subset \mathbb{R}^n$. 称集合 D 为凸集, 是指对任意的 $x, y \in D$ 及任意的实数 $\lambda \in [0,1]$, 都有 $\lambda x + (1-\lambda)y \in D$.

由上述定义不难知道凸集的几何意义. 即对非空集合 $D \subset \mathbb{R}^n$, 若连接其中任意两点的线段仍属于该集合，则称该集合 D 为凸集.

不难证明凸集的下列基本性质.

引理 1.1 设 D, D_1, D_2 是凸集，α 是一实数, 那么

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

(2) 交集 $D_1 \cap D_2$ 是凸集.

(3) 和集 $D_1 + D_2 = \{z | z = x + y,\ x \in D_1,\ y \in D_2\}$ 也是凸集.

例 1.3 n 维欧氏空间中的 l 个点的凸组合是一个凸集. 即集合

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

是凸集.

例 1.4 n 维欧氏空间中的超平面 $H=\{x|c^Tx=\alpha\}$ 是一个凸集, 其中 $\alpha\in\mathbb{R}$, $c\in\mathbb{R}^n\backslash\{0\}$ 是超平面的法向量. 此外, 下面的四个半空间

(1) 正的闭半空间 $H^+=\{x|c^Tx\geq\alpha\}$,

(2) 负的闭半空间 $H^+=\{x|c^Tx\leq\alpha\}$,

(3) 正的开半空间 $H^+=\{x|c^Tx>\alpha\}$,

(4) 负的开半空间 $H^+=\{x|c^Tx<\alpha\}$.

都是凸集

例 1.5 以 $x^0\in\mathbb{R}^n$ 为起点, $d\in\mathbb{R}^n\backslash\{0\}$ 为方向的射线

$$r(x^0;d):=\{x\in\mathbb{R}^n\,|\,x=x^0+\alpha d,\ \alpha\geq 0\}$$

是凸集.

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

$$\operatorname{conv}(D) := \bigcap_{C \supseteq D} C,$$

其中 C 为凸集.

下面我们给出锥和凸锥的定义.

定义 1.5 设非空集合 $C \subset \mathbb{R}^n$. 若对任意的 $x \in C$ 和任意的实数 $\lambda > 0$，有 $\lambda x \in C$, 则称 C 为一个锥 (cone). 若 C 同时也是凸集, 则称 C 为一个凸锥 (convex cone). 此外, 对于锥 C, 若 $0 \in C$, 则称 C 是一个尖锥 (pointed cone). 相应地, 包含 0 的凸锥称为尖凸锥.

例 1.6 多面体 $\{x \in \mathbb{R}^n | Ax \geq 0\}$ 是一个尖凸锥, 通常称之为多面锥 (polyhedral cone).

例 1.7 集合

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n | x_i \geq 0,\ i = 1, 2, \cdots, n\}$$

是一个尖凸锥, 通常称之为非负锥 (nonnegative orthant). 相应地, 凸锥

$$\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n | x_i > 0,\ i = 1, 2, \cdots, n\}$$

称为正锥 (positive orthant).

有了凸集的概念之后，就可以定义凸集上的所谓凸函数.

定义 1.6 设函数 $f : D \subset \mathbb{R}^n \to \mathbb{R}$, 其中 D 为凸集.

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

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

(2) 称 f 是 D 上的严格凸函数, 是指对任意的 $x, y \in D$, $x \neq y$ 及任意的实数 $\lambda \in [0, 1]$, 都有

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y).$$

(3) 称 f 是 D 上的一致凸函数, 是指存在常数 $\gamma > 0$, 使对任意的 $x, y \in D$ 及任意的实数 $\lambda \in [0,1]$, 都有

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

凸函数具有下列基本性质.

命题 1.1 设 f, f_1, f_2 都是凸集 D 上的凸函数, $c_1, c_2 \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, 则有

(1) $cf_1(x) + c_2 f_2(x)$ 也是 D 上的凸函数.

(2) 水平集

$$\mathcal{L}(f,\alpha) = \{x | x \in D,\ f(x) \leq \alpha\}$$

是凸集.

凸集和凸函数在优化理论中起着举足轻重的作用, 但是利用凸函数的定义来判断一个函数是否具有凸性并非一件容易的事情. 如果函数是一阶或二阶连续可微的, 则可利用函数的梯度或 Hesse 阵来判别或验证函数的凸性要相对容易一些. 下面给出几个判别定理.

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

(1) f 在 D 上为凸函数的充要条件是

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*), \quad \forall\, x^*, x \in D. \tag{1.14}$$

(2) f 在 D 上严格凸的充要条件是，当 $x \neq y$ 时，成立

$$f(x) > f(x^*) + \nabla f(x^*)^T (x - x^*), \quad \forall\, x^*, x \in D. \tag{1.15}$$

(3) f 在 D 上一致凸的充要条件是，存在常数 $c > 0$，使对任意的 $x^*, x \in D$，成立

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) + c\|x - x^*\|^2. \tag{1.16}$$

我们知道，在一元函数中，若 $f(x)$ 在区间 (a, b) 上二阶可微且 $f''(x) \geq 0\ (> 0)$，则 $f(x)$ 在 (a, b) 内凸（严格凸）. 对于二阶连续可微的多元函数 $f : D \subset \mathbb{R}^n \to \mathbb{R}$, 也可以由其二阶导数 (Hesse 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 D 上是二阶连续可微的. 若对一切 $h \in R^n$, 有 $h^T\nabla^2 f(x)h \geq 0$，则称 $\nabla^2 f$ 在点 x 处是半正定的. 若对一切 $0 \neq h \in \mathbb{R}^n$, 有 $h^T\nabla^2 f(x)h > 0$，则称 $\nabla^2 f$ 在点 x 处是正定的. 进一步，若存在常数 $c > 0$, 使得对任意的 $h \in \mathbb{R}^n$, $x \in D$, 有 $h^T\nabla^2 f(x)h \geq c\|h\|^2$，则称 $\nabla^2 f$ 在 D 上是一致正定的.

有了上述定义，我们可以把一元函数关于用二阶导数表述凸性的结果推广到多元函数上.

定理 1.5　设 n 元实函数 f 在凸集 $D \subset \mathbb{R}^n$ 上二阶连续可微, 则

(1) f 在 D 上是凸的充要条件是 $\nabla^2 f(x)$ 对一切 $x \in D$ 为半正定；

(2) f 在 D 上是严格凸的充分条件是 $\nabla^2 f(x)$ 对一切 $x \in D$ 为正定；

(3) f 在 D 上是一致凸的充要条件是 $\nabla^2 f(x)$ 对一切 $x \in D$ 为一致正定.

注意，$\nabla^2 f$ 正定是 f 严格凸的充分条件而非必要条件.

1.5 无约束问题的最优性条件

本节讨论无约束优化问题

$$\min f(x) \tag{1.17}$$

的最优性条件, 它包含一阶条件和二阶条件. 首先给出极小点的定义, 它分为全局极小点和局部极小点.

定义 1.8 若对于任意的 $x \in \mathbb{R}^n$, 都有

$$f(x^*) \leq f(x),$$

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

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

$$f(x^*) \leq f(x),$$

则称 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 $x \neq x^*$, 则称 x^* 为 f 的一个严格局部极小点.

由上述定义可知, 全局极小点一定是局部极小点, 反之不然. 一般来说, 求全局极小点是相当困难的, 因此通常只求局部极小点 (在实际应用中, 有时求局部极小点已满足了问题的要求). 故本书所讨论的求极小点的方法都是指局部极小点.

为了讨论和叙述的方便, 通篇引入下列记号:

$$g(x) = \nabla f(x), \quad g_k = \nabla f(x_k), \quad G(x) = \nabla^2 f(x), \quad G_k = \nabla^2 f(x_k).$$

定理 1.6 (**一阶必要条件**) 设 $f(x)$ 在开集 D 上一阶连续可微. 若 $x^* \in D$ 是 (1.17) 的一个局部极小点, 则必有 $g(x^*) = 0$.

证 取 $x = x^* - \alpha g(x^*) \in D$, 其中 $\alpha > 0$ 为某个常数. 则有

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

注意到 $f(x) \geq f(x^*)$ 及 $\alpha > 0$, 我们有

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

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

定理 1.7 (**二阶必要条件**) 设 $f(x)$ 在开集 D 上二阶连续可微. 若 $x^* \in D$ 是 (1.17) 的一个局部极小点, 则必有 $g(x^*) = 0$ 且 $G(x^*)$ 是半正定矩阵.

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

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

即

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

对上式令 $\alpha \to 0$, 即得 $d^T G(x^*) d \geq 0$, 从而定理成立. $\square$

定理 1.8 (**二阶充分条件**) 设 $f(x)$ 在开集 D 上二阶连续可微. 若 $x^* \in D$ 满足条件 $g(x^*) = 0$ 及 $G(x^*)$ 是正定矩阵, 则 x^* 是 (1.17) 的一个局部极小点.

证 任取 $x = x^* + \alpha d \in D$, 其中 $\alpha > 0$ 且 $d \in \mathbb{R}^n$. 由泰勒公式得

$$f(x^* + \alpha d) = f(x^*) + g(x^*)^T d + \frac{1}{2}\alpha^2 d^T G(x^* + \theta\alpha d)d,$$

其中 $\theta \in (0,1)$. 注意到 $g(x^*) = 0$, $G(x^*)$ 正定和 f 二阶连续可微, 故存在 $\delta > 0$, 使得 $G(x^* + \theta\alpha d)$ 在$\|\theta\alpha d\| < \delta$ 范围内正定. 因此由上式即得

$$f(x^* + \alpha d) > f(x^*),$$

从而定理成立. □

一般来说, 目标函数的稳定点不一定是极小点. 但对于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三者是等价的.

定理 1.9 设 $f(x)$ 在 $\mathbb{R}^n$ 上是凸函数并且是一阶连续可微的. 则 $x^* \in \mathbb{R}^n$ 是 (1.17) 的全局极小点的充要条件是 $g(x^*) = 0$.

证 只需证明充分性, 必要性是显然的. 设 $g(x^*) = 0$. 由凸函数的判别定理 1.4(1), 可得

$$f(x) \geq f(x^*) + g(x^*)^T(x - x^*) = f(x^*), \ \ \forall x \in \mathbb{R}^n,$$

这表明 x^* 是全局极小点. □

1.6 无约束优化问题的算法框架

在数值优化中, 一般采用迭代法求解无约束优化问题

$$\min f(x) \tag{1.18}$$

的极小点. 迭代法的基本思想是: 给定一个初始点 x_0, 按照某一迭代规则产生一个迭代序列 $\{x_k\}$. 使得若该序列是有限的, 则最后一个点就是问题 (1.18) 的极小点; 否则, 若序列 $\{x_k\}$ 是无穷点列时, 它有极限点且这个极限点即为问题 (1.18) 的极小点.

设 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, 则第 k 次迭代完成后可得到新一轮 (第 $k+1$ 次) 的迭代点

$$x_{k+1} = x_k + \alpha_k d_k. \tag{1.19}$$

因此, 我们可以写出求解无约束优化问题 (1.18) 的一般算法框架如下.

算法 1.1 (**无约束问题的一般算法框架**)

步 0　给定初始化参数及初始迭代点 x_0. 置 $k := 0$.

步 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 2　通过求解 x_k 处的某个子问题确定下降方向 x_k.

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

为了方便, 通常称上述算法中的 $s_k = \alpha_k d_k$ 为第 k 次迭代的位移. 从算法 1.1 可以看出, 不同的位移 (不同的搜索方向及步长因子) 即产生了不同的迭代算法. 为了保证算法的收敛性, 一般要求搜索方向为所谓的下降方向:

定义 1.10 若存在 $\bar{\alpha} > 0$, 使得对任意的 $\alpha \in (0, \bar{\alpha}]$ 和 $d_k \neq 0$, 有

$$f(x_k + \alpha d_k) < f(x_k).$$

则称 d_k 为 $f(x)$ 在 x_k 处的一个下降方向.

若目标函数 f 是一阶连续可微的, 则判别 d_k 是否为下降方向将有更为方便的判别条件.

引理 1.2 设函数 $f: D \subset \mathbb{R}^n \to \mathbb{R}$ 在开集 D 上一阶连续可微, 则 d_k 为 $f(x)$ 在 x_k 处一个下降方向的充要条件是

$$\nabla f(x_k)^T d_k < 0. \tag{1.20}$$

证 由泰勒展开式得

$$f(x_k + \alpha d_k) = f(x_k) + \alpha \nabla f(x_k)^T d_k + o(\alpha).$$

若 (1.20) 成立, 则对于充分小的 $\alpha > 0$, 显然有 $f(x_k + \alpha d_k) < f(x_k$, 即 d_k 为 $f(x)$ 在 x_k 处的一个下降方向. 反之, 则有

$$\alpha \nabla f(x_k)^T d_k + o(\alpha) < 0,$$

即

$$\nabla f(x_k)^T d_k + \frac{o(\alpha)}{\alpha} < 0.$$

由于 $\lim\limits_{\alpha\to 0} \dfrac{o(\alpha)}{\alpha} = 0$, 故可推得 (1.20) 成立. □

算法的收敛性是设计一个迭代算法的最起码要求. 关于收敛性, 有所谓的“局部收敛性”和“全局收敛性”概念:

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, 我们给出有关收敛速度的概念如下.

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

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = \theta.$$

(1) 若 $p=1$ 且 $0<\theta<1$, 则称该算法具有线性收敛速度 (或线性收敛的).

(2) 若 $p=1$ 且 $\theta=0$, 则称该算法具有超线性收敛速度 (或超线性收敛的).

(3) 若 $p=2$ 且 $0<\theta<\infty$, 则称该算法具有平方收敛速度 (或平方收敛的).

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

在计算机上实现一个迭代法, 通常需要一个迭代终止条件. 常用的基本终止条件有下列三种:

(1) 位移的绝对误差或相对误差充分小, 即

$$\|x_{k+1}-x_k\|<\varepsilon, \quad \text{或} \quad \frac{\|x_{k+1}-x_k\|}{\|x_k\|}<\varepsilon,$$

其中 ε 是充分小的正数.

(2) 目标函数的绝对误差或相对误差充分小, 即

$$|f(x_{k+1}) - f(x_k)| < \varepsilon, \quad \text{或} \quad \frac{|f(x_{k+1}) - f(x_k)|}{|f(x_k)|} < \varepsilon,$$

其中 ε 是充分小的正数.

(3) 目标函数的梯度的范数充分小, 即

$$\|\nabla f(x_k)\| < \varepsilon,$$

其中 ε 是充分小的正数.