第三章　无约束优化方法

本章考虑无约束优化问题

$$\min_{x\in\mathbb{R}^n} \; f(x) \tag{3.0.1}$$

其中$f:\mathbb{R}^n\to\mathbb{R}$. 在介绍问题(3.0.1)最优性条件之后，着重介绍求解该问题的各种迭代算法及其收敛性. 下面的符号将被频繁地使用：

$$g(x) := \nabla f(x), \quad g_k := \nabla f(x^k), \quad g_* := \nabla f(x^*),$$

$$G(x) := \nabla^2 f(x), \quad G_k := \nabla^2 f(x^k), \quad G_* := \nabla^2 f(x^*).$$

§3.1　算法理论基础

3.1 算法理论基础 — 最优性条件

§3.1.1 最优性条件

回忆一元函数$\varphi:\mathbb{R}\to\mathbb{R}$的极值问题, 有以下结论:

- (一阶必要条件) 若函数φ一阶连续可微且λ_*为φ的局部极小值点, 则$\varphi'(\lambda_*)=0$;
- (二阶必要条件) 若函数φ二阶连续可微且λ_*为φ的局部极小点值, 则$\varphi'(\lambda_*)=0$并且$\varphi''(\lambda_*)\geq 0$;
- (二阶充分条件) 若函数φ二阶连续可微, $\varphi'(\lambda_*)=0$且$\varphi''(\lambda_*)>0$, 则λ_*为函数φ的严格局部极小值点.

下面的引理给出了一元函数与多元实值函数极值之间的关系.

3.1 算法理论基础 —最优性条件

引理 3.1.1 假设多元实值函数$f:\mathbb{R}^n \to \mathbb{R}$是连续的. 给定$x^* \in \mathbb{R}^n$和任意的$d \in \mathbb{R}^n \setminus \{0\}$, 那么$x^*$是函数$f$的(严格)局部极小值点当且仅当$\lambda_* = 0$是一元函数$\varphi(\lambda) = f(x^* + \lambda d)$的(严格)局部极小值点.

证明 $\varphi(0) = f(x^*)$是显然的.

- 假设x^*是多元实值函数f的局部极小值点, 那么对任意靠近x^*的$x \in \mathbb{R}^n$ 有$f(x^*) \le f(x)$. 而对任意的$d \in \mathbb{R}^n \setminus \{0\}$, 当$\lambda$充分靠近0时, $x^* + \lambda d$将充分靠近x^*, 因此, $f(x^*) \le f(x^* + \lambda d)$, 所以$\lambda_* = 0$是一元函数$\varphi(\lambda) = f(x^* + \lambda d)$的局部极小值点.

3.1 算法理论基础 —最优性条件

- 假设$\lambda_* = 0$是一元函数$\varphi(\lambda) = f(x^* + \lambda d)$的局部极小值点，那么存在$\lambda_* = 0$的一个邻域

$$\mathcal{N}^0(\delta) = (-\delta, \delta) \text{ 其中}\delta\text{是一个固定的正数}$$

使得对任意的$\lambda \in \mathcal{N}^0(\delta)$，都有$\varphi(0) \le \varphi(\lambda)$. 为了证明$x^*$是多元实值函数$f$的局部极小值点，我们构造以下邻域:

$$\mathcal{N}(\delta) = \{x \in \mathbb{R}^n \mid \|x - x^*\| < \delta\}.$$

对任意的$x \in \mathcal{N}(\delta)$，存在$\mu \in \mathbb{R}$和某个单位向量$d \in \mathbb{R}^n$，使得$x = x^* + \mu d$. 由于$x \in \mathcal{N}(\delta)$，所以

$$|\mu| = \|\mu d\| = \|x - x^*\| < \delta,$$

这表明: $\mu \in \mathcal{N}^0(\delta)$. 因此,

3.1 算法理论基础 —最优性条件

$$f(x^*) = \varphi(0) \le \varphi(\mu) = f(x).$$

由x任意性知: 对于邻域$N(\delta)$的任意点x, 都有$f(x^*) \le f(x)$, 所以x^*是多元实值函数f的局部极小值点.

因此, x^*是多元实值函数f的局部极小值点当且仅当$\lambda_* = 0$是一元函数$\varphi(\lambda) = f(x^* + \lambda d)$的局部极小值点. 类似可得: x^*是多元实值函数f的严格局部极小值点当且仅当$\lambda_* = 0$是一元函数$\varphi(\lambda) = f(x^* + \lambda d)$的严格局部极小值点. □

借助于一元函数的极值条件和引理3.1.1, 下面来讨论无约束优化问题(3.0.1)的最优性条件.

3.1 算法理论基础 —最优性条件

定理 3.1.1 *(必要条件)*

(i) *(一阶必要条件)* 若函数f在$x^* \in \mathbb{R}^n$的某邻域内一阶连续可微且x^*为f的局部极小值点, 那么$g_* = 0$.

(ii) *(二阶必要条件)* 若函数f在$x^* \in \mathbb{R}^n$的某邻域内二阶连续可微且x^*为f的局部极小值点, 那么$g_* = 0$且G_*对称半正定.

证明 只证明(ii)中的结论, (i)中的结论类似可得. 任取$d \in \mathbb{R}^n \setminus \{0\}$, 定义$\varphi(\lambda) = f(x^* + \lambda d)$. 由于函数$f$在$x^*$的某邻域内二阶连续可微, 所以函数$\varphi$在$\lambda_* = 0$的某邻域内二阶连续可微且

$$\varphi'(\lambda) = g(x^* + \lambda d)^\top d, \quad \varphi''(\lambda) = d^\top G(x^* + \lambda d)d.$$

由于x^*为f的局部极小值点, 由引理3.1.1知: $\lambda_* = 0$是一元函数φ的局部极小值点. 因此, 由一元函数的极值条件$\varphi'(0) = 0$和$\varphi''(0) \geq 0$可得

$$g_*^\top d = 0, \quad d^\top G_* d \geq 0.$$

由d的任意性进一步可得: $g_* = 0$且G_*半正定. 定理得证. □

3.1 算法理论基础 —最优性条件

称满足一阶必要条件$g_* = 0$的点为稳定点（也称为驻点）. 稳定点可分为三种类型: 极小值点, 极大值点和鞍点. 所谓鞍点, 就是沿着某些方向它是极小值点; 而沿着另一些方向它是极大值点. 当$n = 3$, 即在立体空间中, 由于在该点附近函数的图象形如马鞍, 故称其为鞍点.

定理 3.1.2 (二阶充分条件) 假设函数f在$x^* \in \mathbb{R}^n$的某邻域内二阶连续可微. 如果$g_* = 0$并且G_*正定, 那么x^*为函数f的严格局部极小值点.

证明 任取$d \in \mathbb{R}^n \setminus \{0\}$, 定义$\varphi(\lambda) = f(x^* + \lambda d)$. 由于函数$f$在$x^*$的某邻域内二阶连续可微, 所以函数$\varphi$在$\lambda_* = 0$的某邻域内二阶连续可微且 $\varphi'(0) = g_*^\top d, \quad \varphi''(0) = d^\top G_* d.$

由于$g_* = 0$且G_*正定, 所以可得

$$\varphi'(0) = g_*^\top d = 0 \text{ 且 } \varphi''(0) = d^\top G_* d > 0, \quad \forall d \in \mathbb{R}^n \setminus \{0\}.$$

这表明: $\lambda_* = 0$是一元函数φ的严格局部极小值点. 所以由引理3.1.1, 可知x^*为f的严格局部极小值点. 定理得证. □

3.1 算法理论基础 —最优性条件

如果$f:\mathbb{R}^n\to\mathbb{R}$为凸函数, 那么有以下更好的结论.

定理 3.1.3 *(无约束凸规划的最优性条件)* 假设$f:\mathbb{R}^n\to\mathbb{R}$是一阶连续可微的凸函数, 则$x^*$为函数$f$的全局极小值点当且仅当$g_*=0$.

证明 必要性由定理3.1.1可得, 下证充分性. 由于f是$\mathbb{R}^n$上的一阶连续可微的凸函数, 根据凸函数的一阶判别定理(即: 定理1.3.4), 因而可得对任意的$x\in\mathbb{R}^n$, 有

$$f(x)-f(x^*)\geq g_*^\top(x-x^*).$$

结合上式和条件$g_*=0$, 于是有: 对任意的$x\in\mathbb{R}^n$有$f(x)\geq f(x^*)$, 这表明: x^*为函数f的全局极小值点. 定理得证. □

§3.1.2 线搜索迭代下降算法及其收敛性

线搜索迭代下降算法是求解无约束优化问题(3.0.1)的流行算法之一，其基本思想为：从某个初始点开始，按照一定的规则不断地寻找迭代方向和（用线搜索）迭代步长进行迭代，直到算法终止. 一般迭代框架如下：

算法 3.1.1 (线搜索迭代下降算法) 选择初始点$x^0 \in \mathbb{R}^n$. 置$k := 0$.

步1 如果$g_k = 0$, 算法终止.

步2 确定下降方向$d^k \in \mathbb{R}^n$.

步3 使用线搜索确定步长$\lambda_k > 0$, 使得$f(x^k + \lambda_k d^k) < f(x^k)$.

步4 置$x^{k+1} := x^k + \lambda_k d^k$和$k := k + 1$, 并返回到步1.

3.1 算法理论基础 — 算法及其收敛性

下面讨论算法3.1.1的收敛性. 假设θ_k表示方向d^k与$-g_k$之间的夹角，则

$$\cos\theta_k = -\frac{g_k^\top d^k}{\|g_k\|\|d^k\|}. \qquad (3.1.1)$$

假设算法3.1.1中的步长λ_k分别由精确一维线搜索、Armijo线搜索、Wolfe线搜索得到, 那么对应算法的全局收敛性分别由以下三个定理给出.

定理 3.1.4 假设函数$f:\mathbb{R}^n \to \mathbb{R}$连续可微且有下界，且其梯度函数$g$具有*Lipschitz*连续性，即存在常数$\mathcal{L} > 0$使得对任意的$x, y \in \mathbb{R}^n$, 有$\|g(x) - g(y)\| \le \mathcal{L}\|x - y\|$. 设无穷序列$\{x^k\}$由算法*3.1.1* 产生，其中步长$\lambda_k$由精确一维线搜索得到. 那么,

$$\sum_{k=0}^{\infty} \|g_k\|^2 \cos^2\theta_k < \infty. \qquad (3.1.2)$$

特别地, 若存在常数$\beta > 0$使得$\cos\theta_k \ge \beta$, 则$\lim_{k\to\infty} \|g_k\| = 0$.

3.1 算法理论基础 — 算法及其收敛性

证明 由于最后的结论可由(3.1.2)式的结论直接得到，因而下面只需证明(3.1.2)式的结论成立. 因为函数f连续可微, 所以对任意的$\lambda > 0$, 存在$\alpha_k \in (0, 1)$使得

$$
\begin{aligned}
f(x^k + \lambda d^k) &= f(x^k) + \lambda g(x^k + \alpha_k \lambda d^k)^\top d^k \\
&= f(x^k) + \lambda g_k^\top d^k + \lambda [g(x^k + \alpha_k \lambda d^k) - g_k]^\top d^k \\
&\leq f(x^k) + \lambda g_k^\top d^k + \lambda \|g(x^k + \alpha_k \lambda d^k) - g_k\| \|d^k\| \\
&\leq f(x^k) + \lambda g_k^\top d^k + \lambda^2 \mathcal{L} \|d^k\|^2,
\end{aligned}
$$

其中，第一个等式成立是由于中值定理，第一个不等式成立是由于Cauchy-Schwartz不等式. 特别地, 取$\hat{\lambda}_k := -\frac{g_k^\top d^k}{2\mathcal{L}\|d^k\|^2}$, 则由上式可得

$$
f(x^k + \hat{\lambda}_k d^k) \leq f(x^k) - \frac{(g_k^\top d^k)^2}{4\mathcal{L}\|d^k\|^2}.
$$

3.1 算法理论基础 — 算法及其收敛性

又由精确一维线搜索知: 步长λ_k满足$f(x^{k+1}) = f(x^k+\lambda_k d^k) \le f(x^k+\hat{\lambda}_k d^k)$. 因此,

$$f(x^{k+1}) \le f(x^k + \hat{\lambda}_k d^k) \le f(x^k) - \frac{(g_k^\top d^k)^2}{4\mathcal{L}\|d^k\|^2}.$$

根据(3.1.1)式, 由上式可得

$$\|g_k\|^2 \cos^2\theta_k \le 4\mathcal{L}[f(x^k) - f(x^{k+1})].$$

由假设知: 函数f有下界. 因此, 对上式求和即可得到(3.1.2)式成立. 定理得证. □

定理 3.1.5 假设无穷序列$\{x^k\}$由算法3.1.1 产生, 其中迭代步长λ_k是由Armijo线搜索得到, 并且定理3.1.4的其它条件成立. 如果存在常数$c > 0$使得

$$\|g_k\| \le c\|d^k\|. \tag{3.1.3}$$

那么定理3.1.4的结论成立.

定理 3.1.6 如果无穷序列$\{x^k\}$由算法3.1.1 产生, 其中迭代步长λ_k是由Wolfe线搜索得到, 并且定理3.1.4的其它条件成立, 那么定理3.1.4的结论成立.

下面的定理给出算法3.1.1的收敛速度的有关性质.

定理 3.1.7 假设$f:\mathbb{R}^n\to\mathbb{R}$二次连续可微，无穷序列$\{x^k\}$由算法3.1.1 产生，其中迭代步长$\lambda_k$由*Armijo*线搜索得到(其中$\sigma\in(0,1/2)$)或由*Wolfe*线搜索得到. 假定$x^k\to x^*$, $g_*=0$并且G_*对称正定. 如果

$$\lim_{k\to\infty}\frac{\|g_k+G_kd^k\|}{\|d^k\|}=0, \tag{3.1.4}$$

那么，当k充分大时有$\lambda_k=1$. 进一步，序列$\{x^k\}$局部超线性收敛于x^*.

3.1 算法理论基础 — 作业

3.1 求出函数

$$f(x) = 2x_1^2 + x_2^2 - 2x_1x_2 + 2x_1^3 + x_1^4$$

所有的稳定点, 并判断哪些稳定点是极小值点? 是否是整体极小值点?

3.2 试证明：函数

$$f(x_1, x_2) = x_1^2 - 4x_2^2 + 20x_1 + 64x_2$$

存在唯一的稳定点，但该点既不是函数f的极大值点，也不是函数f的极小值点.

3.3 试讨论参数a取何值时，点$x = (0,0,0)^\top$是函数$f(x) = ax_1^2e^{x_2} + x_2^2e^{x_3} + x_3^2e^{x_1}$ 的极大值点，极小值点，鞍点?