科学版

研究生教学丛书

非线性最优化理论与方法

王宜举 修乃华 编著

科学出版社

科学版研究生教学丛书

非线性最优化理论与方法

王宜举　修乃华　编著

科学出版社

北　京

内 容 简 介

本书系统地介绍了非线性最优化问题的有关理论与方法，主要包括一些传统理论与经典方法，如非线性最优化问题的最优性理论，无约束优化问题的线搜索方法、共轭梯度法、拟牛顿方法，约束优化问题的可行方法、罚函数方法和 SQP 方法等，同时也吸收了新近发展成熟并得到广泛应用的成果，如信赖域方法、投影方法等.

本书在编写过程中既注重基础理论的严谨性和方法的实用性，又保持内容的新颖性. 该书内容丰富、系统性强，可作为运筹学专业的研究生和数学专业高年级本科生从事非线性最优化研究的入门教材或参考书，也可作为相关专业科研人员的工具参考书.

图书在版编目(CIP)数据

非线性最优化理论与方法/王宜举，修乃华编著.—北京：科学出版社，2012
（科学版研究生教学丛书）
ISBN 978-7-03-033028-4
Ⅰ.①非… Ⅱ.①王… ②修… Ⅲ.①非线性-最优化算法-高等学校-教材
Ⅳ.①O224
中国版本图书馆 CIP 数据核字(2011) 第 260027 号

责任编辑：李 欣 赵彦超 / 责任校对：陈玉凤
责任印制：钱玉芬 / 封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

中国科学院印刷厂印刷
科学出版社发行 各地新华书店经销
*
2012 年 1 月第 一 版 开本：B5 (720×1000)
2012 年 1 月第一次印刷 印张：16
字数：310 000

定价：45.00 元
(如有印装质量问题，我社负责调换)

前　　言

第二次世界大战期间, 运筹学伴随军事上的需要而产生. 战后, 运筹学开始转向民用工业的运用, 并不断取得进展. 20 世纪 60 年代, 最优化方法发展成为运筹学的一门新兴学科. 而后, 近代科学技术的发展, 特别是电子计算机技术的飞速发展促进了最优化方法的迅速发展. 很快, 这门新兴的基础学科便渗透到各个技术领域, 形成了最优化方法与技术这门应用学科, 并发展出了新的更细的研究分支.

作为运筹学的一个重要研究分支, 非线性最优化问题的研究在近三十年得到快速发展, 新的理论和方法不断出现. 为及时吸收新近发展成熟并在实际中得到广泛应用的成果以应用于教学科研中, 作者参阅国内外关于最优化理论与方法的许多专著、教材和研究文献, 并结合自己的教学实践编写了这本书.

非线性最优化问题的研究内容十分丰富. 但由于篇幅所限, 本书主要对这类问题的传统理论与经典梯度型数值方法及其新的研究进展做了比较详尽的论述, 使读者不但能掌握非线性最优化问题的基础理论和梯度型数值方法的操作流程, 而且能清楚这些方法的设计思想和性能, 从而为设计更有效的优化方法提供理论依据和帮助.

法国数学家拉格朗日说过, 一个数学家, 只有当他走出去, 对他在大街上遇到的第一个人清楚地解释自己的工作时, 他才算完全理解了自己的工作. 实际上, 他定义的这种境界也是每一个数学工作者, 特别是数学教育工作者追求的目标. 因此, 作者在编写本书时力求把对先修课程的要求放得最低, 使读者只需具备多元微积分和线性代数的基础知识即可阅读此书. 同时, 为增强可读性, 作者尽可能多地介绍一些方法和技术的引入背景、思想及其发展历程, 并对书中结论给出了比较详细的证明过程.

在本书十余年的锤炼过程中, 作者得到曲阜师范大学的王长钰教授、中国农业大学的邓乃扬教授、大连理工大学的夏尊铨教授和张立卫教授、香港理工大学的祁力群教授、澳大利亚科廷大学的张国礼教授、南京师范大学的孙文瑜教授、南京航空航天大学的倪勤教授、重庆师范大学的杨新民教授和重庆大学的李声杰教授的悉心指导和帮助, 国内的很多同行也提出了许多宝贵的指导性建议, 在此一并向他们表示诚挚的谢意! 最后, 感谢国家 973 项目 (2010CB732501) 和国家自然科学基金 (11171180) 经费资助.

恳请读者不吝赐教, 来信请发至:

wang-yiju@163.com 或 nhxiu@bjtu. edu.cn.

王宜举　修乃华

2012 年 1 月

目　　录

符　号　表

R^n	n 维欧氏空间
$\langle x, y\rangle$	向量 x, y 的内积
$\|x\|$	向量 x 的 2-范数
$\mathbf{0}$	零向量
e_i	第 i 个分量为 1 其余分量为 0 的向量
$(x_1; x_2; \cdots; x_n)$	向量 $(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$
$\|A\|$	矩阵 A 的谱范数 (2-范数)
$\|A\|_{\mathrm{F}}$	矩阵 A 的 Frobenius 范数
A^{T}	矩阵 A 的转置
$\lambda_{\max}(A)$	矩阵 A 的最大实特征根
$\lambda_{\min}(A)$	矩阵 A 的最小实特征根
$\mathrm{tr}(A)$	矩阵 A 的迹
$\mathrm{rank}(A)$	矩阵 A 的秩
$\det(A)$	矩阵 A 的行列式
A^+	矩阵 A 的广义逆 (伪逆)
$\kappa(A)$	矩阵 A 的条件数
$\mathcal{R}(A)$	矩阵 A 的值空间
$\mathcal{N}(A)$	矩阵 A 的核空间
$[a_i, i \in \mathcal{E}]$	下标属于 $\mathcal{E}$的列向量 a_i 构成的矩阵
$\mathrm{span}[a_1, a_2, \cdots, a_s]$	列向量 $a_1, a_2, \cdots, a_s$ 所生成的线性子空间
$\mathrm{diag}(d_1, d_2, \cdots, d_n)$	以 $d_1, d_2, \cdots, d_n$ 为对角元的对角阵
$\nabla f(x)$	函数 $f: R^n \to R$ 在 x 点的梯度
$\nabla^2 f(x), \nabla_{xx} f(x)$	函数 $f: R^n \to R$ 在 x 点的 Hesse 矩阵
$\mathrm{D}F(x), \mathrm{D}_x F(x)$	向量函数 $F: R^n \to R^m$ 在 x 点的 Jacobi 矩阵
$\|\mathcal{E}\|$	指标集 $\mathcal{E}$ 中元素的个数
$\mathrm{bd}(S)$	集合 S 的边界集
$\mathrm{Aff}(S)$	集合 S 的仿射包
$\mathrm{int}(S)$	集合 S 的内点集
$\mathrm{cl}(S)$	集合 S 的闭包
$\mathrm{ri}(S)$	集合 S 的相对内点集
$N(x, \delta)$	以 x 为中心以 δ 为半径的邻域
$\mathcal{K}^\circ$	锥 $\mathcal{K}$ 的极锥
$\mathcal{L}(x_0)$	水平集 $\{x \in R^n \mid f(x) \leqslant f(x_0)\}$

第 1 章　引　　论

本章首先介绍了非线性最优化问题的理论与方法中常用的一些基本概念和基础知识, 然后介绍了一些常见的求解方法及性能分析, 给出了凸集和凸函数的概念和有关性质, 最后给出了无约束优化问题的最优性条件.

1.1　最优化问题

在现实生活中, 经常会遇到这样一类实际问题, 要求在众多的方案中选择一个最优方案. 例如, 在工程设计中, 如何选择参数使设计方案既满足设计要求, 又能降低成本; 资源分配时, 怎样分配现有资源才能使得分配方案既满足要求, 又能获得好的经济效益; 加工产品时, 如何搭配各种原料的比例才能既降低成本, 又能提高产品的质量; 农田规划中, 如何安排农作物的布局, 才能使农田高产稳产, 发挥地区优势. 这类基于现有资源使效益极大化或为实现某目标使成本最低化的问题称为最优化问题.

上述问题在数学上可写成如下形式的最优化问题

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & x \in \Omega. \end{aligned} \tag{1.1.1}$$

对于极大化目标函数的情形, 可通过在目标函数前添加负号等价地转化为极小化目标函数. 因此, 这里只考虑极小化目标函数的情形.

在 (1.1.1) 式中, s.t. 是英文 subject to 的缩写; 数值函数 $f : R^n \to R$ 称为目标函数, 又称费用函数 (或效益函数); $\Omega \subseteq R^n$ 称为可行域或决策集, 它是在极小化目标函数的过程中对决策变量 x 取值范围的界定.

可行域有多种表述形式, 一般常用等式和不等式来定义, 即

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

对 $i \in \mathcal{E}$, $c_i(x) = 0$ 称为等式约束, $\mathcal{E}$ 称为等式约束指标集; 对 $i \in \mathcal{I}$, $c_i(x) \geqslant 0$ 称为不等式约束, $\mathcal{I}$ 称为不等式约束指标集.

最优化问题形形色色, 对应的最优化模型多种多样, 人们从不同角度对其进行分类.

(1) 根据有无约束划分. 若 $\Omega = R^n$, 即 x 是自由变量, 则称 (1.1.1) 式为无约束

优化问题; 若 $\Omega \subset R^n$ 且 $\Omega \neq R^n$, 则称 (1.1.1) 式为约束优化问题.

约束优化问题和无约束优化问题在某些情形可以互相转化. 如对于 n 阶实对称正定矩阵 A, 下述两最优化问题等价:

$$\min_{x\in R^n} \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x}, \quad \min\{x^{\mathrm{T}}Ax \mid x^{\mathrm{T}}x = 1\}.$$

这两类问题在理论分析和算法设计方面有很大不同, 而且约束优化问题比无约束优化难求解. 因此, 约束优化问题的一种求解策略是将约束优化问题转化为无约束优化问题, 或通过一个近似的无约束优化问题求解.

(2) 根据约束函数和目标函数的线性与否划分. 若目标函数及约束函数都是线性的, 则称 (1.1.1) 式为线性规划问题; 若目标函数与约束函数中至少有一个函数是非线性的, 则称 (1.1.1) 式为非线性最优化问题. 特别地, 若目标函数是二次函数, 约束函数是线性的, 则称 (1.1.1) 式为二次规划问题. 线性规划和二次规划问题是最优化问题中最简单的两类优化问题. 对这两类问题, 目前已建立起比较完善的理论和有效的算法.

(3) 根据目标函数和可行域的凸性与否划分. 若目标函数为凸函数且可行域为非空闭凸集, 则称 (1.1.1) 式为凸规划问题, 否则称之为非凸优化问题. 凸规划问题的最大特点是其稳定点、局部最优值点和全局最优值点是一致的.

(4) 根据函数的可微性质划分. 若目标函数及约束函数都是连续可微的, 则称 (1.1.1) 式为光滑优化问题; 若目标函数与约束函数中至少有一个函数是不可微的, 称 (1.1.1) 式为非光滑优化问题. 对光滑优化问题, 可利用目标函数和约束函数的梯度信息来估计其邻域内点的函数值信息, 从而建立起梯度型数值方法, 而对非光滑优化问题要建立类似的求解方法, 则需要借助次梯度或光滑化等技术.

(5) 根据可行域中含有可行点的个数划分. 若可行域中含有无穷多个不可数的点且可行域中的点连续变化, 则称 (1.1.1) 式为连续优化问题; 若可行域中含有有限个或可数个点, 即该优化问题在由有限个点或可数个点组成的可行域中寻求最优解, 则称 (1.1.1) 式为离散优化问题. 在很多情况下, 离散优化问题可行域中的点是通过某些元素的排列组合产生的, 因此, 又称其为组合优化问题.

对离散优化问题, 根据变量的取值, 又分离出整数规划问题, 即变量只能取整数的规划问题. 在整数规划问题中, 若变量只能取 0 和 1, 则称其为 0-1 规划问题. 在优化问题中, 如果部分变量为整数变量, 而部分变量为连续变量, 则称其为混合整数规划问题.

对连续优化问题, 特别是光滑的连续优化问题可以利用目标函数与约束函数的连续性质建立求解方法. 而离散优化问题则不然, 这是因为可行域中邻近两点对应的目标函数值差别可能很大. 对于整数规划问题, 若用松弛技术将离散变量连续化,

即将离散优化问题中的变量取整约束放松为取实数, 而其他约束条件不变, 那么求解后者得到的最优解无论通过什么方式取整都不能保证它是原问题的最优解. 这就是说, 离散优化问题一般只能用离散优化问题的方法解决. 尽管如此, 这两类优化问题还是密切相关的, 这一方面表现在有些离散优化问题, 如 0-1 规划, 可以通过约束条件 $x(x-1)=0$ 将其化为连续优化问题; 另一方面, 连续优化问题的一些研究方法和技术, 如对偶, 可移植到离散优化问题的求解算法中.

(6) 根据变量的确定性划分. 若优化问题 (1.1.1) 中的所有系数都是确定的, 则称其为确定型规划问题; 若优化问题 (1.1.1) 中的某些系数具有某种不确定性, 则称其为不确定规划问题. 常见的不确定规划问题主要有随机规划和模糊规划.

最优化问题还有其他一些分类. 从 1947 年线性规划的产生至今, 人们对最优化问题的研究先后经历了从线性到非线性、从连续到离散、从确定到动态、随机和模糊的发展过程. 本书主要讨论目标函数和约束函数均连续可微的确定型非线性最优化问题, 并简单地称之为非线性最优化问题, 有时又称之非线性规划问题.

下面给出非线性最优化问题解的定义.

(1) 对于约束优化问题 (1.1.1), 可行域 Ω 中的点称为可行解或可行点.

(2) 设 $x^*\in\Omega$. 若对任意 $x\in\Omega$, 有 $f(x^*)\leqslant f(x)$, 则称 x^* 为 (1.1.1) 式的全局最优解或全局最优值点, 对应的目标函数值称为全局最优值或全局最小值, 并记

$$x^*=\arg\min_{x\in\Omega} f(x),$$

这里, arg min 取自英文 the argument of the minimum. 若 x^* 还满足对任意 $x\in\Omega$, $x\neq x^*$ 有 $f(x^*)<f(x)$, 则称 x^* 为 (1.1.1) 式的严格全局最优解.

有些优化问题没有最优解, 但目标函数在可行域上有下界, 那么目标函数在可行域上的下确界称为该优化问题的最优值. 如二元函数 $f(x)=x_1^2+(1-x_1x_2)^2$ 在 R^2 上的最优值为零, 但其只能在 $x_1=\dfrac{1}{x_2}$ 且 $x_2\to\infty$ 时达到. 基于上述情况, 有时把最优化问题 (1.1.1) 写成

$$\inf\{f(x)\mid x\in\Omega\}.$$

(3) 设 $x^*\in\Omega$. 若存在 x^* 点的邻域 $N(x^*,\delta)$, 使对任意 $x\in N(x^*,\delta)\cap\Omega$, 有 $f(x^*)\leqslant f(x)$, 则称 x^* 为 (1.1.1) 式的局部最优解或局部最优值点. 若 x^* 还满足对任意 $x\in N(x^*,\delta)\cap\Omega$, $x\neq x^*$, 有 $f(x^*)<f(x)$, 则称 x^* 是 (1.1.1) 式的严格局部最优解.

非线性最优化问题的研究核心是最优解的存在性及其结构性质、求解算法及其性能分析. 对于一般的非线性最优化问题, 求解和验证其全局最优解是一件非常棘手甚至是不可能的事情. 因此, 人们寄希望于求得问题的局部最优解. 即便如此, 由于计算误差等因素, 几乎所有的数值算法只能给出近似解.

1.2　方 法 概 述

如同一元二次方程的求根公式, 对于非线性最优化问题, 一个直接的想法是借助微分学和变分法等数学工具, 通过逻辑推理和分析运算给出问题的最优解, 这就是所谓的解析方法. 该方法得到的解称为解析解. 解析解精确、简洁、直观, 并有助于问题的理论分析. 但它仅适用于特殊形式的非线性优化问题, 而且有时不实用. 如对下述二次规划问题

$$\min_{x\in R^n} \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x,$$

当矩阵 A 对称正定时, 其解析解为 $x = A^{-1}b$. 但该解析解在实际计算时不但计算量大而且稳定性差. 所以, 在实际中, 人们选择 Guass 消元法、三角分解法或线性共轭梯度法求解该问题.

求解非线性优化问题的第二类方法是图解法和实验法. 这类 “手工作坊” 式的方法操作简单、通俗易懂, 但效率较低, 且仅适用于变量维数很小的情况. 尽管如此, 我国运筹学的奠基人华罗庚于 20 世纪 60 年代提出的 “优选法” 在科技相对落后的时代在我国的工农业生产中发挥了巨大作用.

求解非线性优化问题的第三类方法是形式转化法. 该方法首先利用非线性最优化问题的结构性质或最优性条件将其转化成有别于原问题的另一类数学问题, 然后对后者套用现有的方法求解. 不过, 形式转化法只是找到了解决问题的一种途径, 它并非完全有效, 因为等价转化一般是需要条件的, 而且转化后的问题也并不总存在有效算法.

求解非线性最优化问题的第四类方法是迭代方法. 该方法从当前的一个近似解点, 利用目标函数和约束函数在该点的函数值或梯度信息, 通过调优产生更好的近似解点, 直到不能改进为止. 这类算法属于数值方法. 与解析解相对应, 通过这种方法得到的解称为数值解. 一般情况下, 它是近似解.

对于工程技术中的非线性最优化问题, 其求解方法以迭代形式的数值方法最为典型和常见. 这种方法大多与计算机结合, 它们不仅能够计算较复杂的非线性最优化问题, 而且计算效率较高并能进行大规模计算, 因而成为非线性最优化问题的首选方法. 根据迭代过程中下一迭代点的确定性, 它又分为随机搜索型方法和确定型方法.

随机搜索型方法是一种仿生智能优化方法. 它是人们受自然界规律的启迪, 根据其原理来模拟自然界的一些自然现象而建立的优化方法. 这类方法在计算过程中主要利用目标函数的函数值信息, 其有效性可以借助马尔可夫链的遍历理论等概率论和随机过程的知识来给它以数学上的描述, 并在概率意义下得到问题的全局最

优解. 它适用于组合优化问题和规模较小的连续优化问题. 目前应用比较广泛的这类方法主要有遗传方法、模拟退火方法、蚁群方法和神经网络方法等.

根据利用函数信息的程度, 确定型方法分为模式搜索方法和梯度型方法. 模式搜索法又称为直接搜索法. 它主要根据函数值的变化规律探测目标函数的下降方向, 并沿该方向寻求更优的点. 模式搜索法简单、直观, 它不需要计算目标函数的梯度, 主要适用于变量较少、约束简单、目标函数结构比较复杂且梯度不易计算的非线性最优化问题. 常见的主要有坐标轮换法、Hooke-Jeeves 法、Powell 共轭方向法和单纯形调优法等. 与模式搜索法不同, 梯度型方法在计算过程中主要利用函数在当前迭代点或已有迭代点的函数值信息、梯度信息甚至 Hesse 矩阵信息. 因此, 与模式搜索方法相比, 梯度型方法对目标函数和约束函数的解析性质要求较高. 它一般有快的收敛速度, 而且更容易建立算法的理论性质.

对于梯度型数值方法, 一般通过两种策略来由当前迭代点产生下一迭代点：线搜索方法和信赖域方法. 线搜索方法是最常见也是研究最多的一类方法. 在算法的每一迭代步, 首先基于目标函数在当前迭代点或已有迭代点的梯度信息, 产生一个搜索方向, 然后沿该方向寻求一个更靠近最优值点的迭代点, 使目标函数值有某种程度的下降. 当前迭代点与新迭代点之间的 “距离” 称为步长. 由于这种过程执行一次之后并不能得到目标函数的最优解, 所以要重复执行, 直到满足某种条件为止. 具体地, 对无约束优化问题, 线搜索方法的基本框架如下.

算法 1.2.1

步 1. 取初始点 x_0 及有关参数, 令 $k = 0$.

步 2. 验证停机准则.

步 3. 求 x_k 点的搜索方向 d_k.

步 4. 计算迭代步长 α_k, 使满足 $f(x_k + \alpha_k d_k) < f(x_k)$.

步 5. 产生下一迭代点, 即令 $x_{k+1} = x_k + \alpha_k d_k$, $k = k + 1$, 转步 2.

下面对该算法框架的有关问题作一说明.

对于初始点, 其选取不但会影响算法的效率, 而且当目标函数在可行域内含有多个极值点时对最终的数值结果也有较大影响. 习惯上, 取零点或分量全为 1 的点为初始点, 或随机产生初始点. 一个理想的初始点取法是通过挖掘问题的结构性质在最优值点附近取到初始点.

对于算法中的参数, 其不同取值会影响算法的计算效率. 借助理论分析可得参数合理的取值范围, 而通过大量的数值实验可得其经验值. 如果参数能根据迭代状况自动调整, 无疑会有好的数值效果.

对于该下降算法, 无论设置多么苛刻的条件, 都很难在有限迭代步内得到问题

的精确解. 因此, 一般选择在算法的迭代进程停滞不前时终止计算. 据此, 常用的停机准则主要有最优性条件准则、点距准则和函数下降量准则. 具体来讲, 就是当优化问题一旦近似满足某最优性条件, 或算法产生的迭代点进展非常缓慢 (相邻两迭代点之间的距离很小), 或目标函数值下降非常缓慢 (相邻两迭代点的目标函数值相差很小) 时算法就终止.

对于搜索方向, 其选取原则是要保证从当前迭代点沿该方向移动时目标函数值有所下降. 也就是说, 搜索方向应是下降方向.

定义 1.2.1 称 $d \in R^n$ 为函数 $f : R^n \to R$ 在 $x \in R^n$ 点的下降方向, 如果存在 $\delta > 0$, 使对任意的 $t \in (0, \delta]$,

$$f(x + td) < f(x).$$

对于连续可微函数 $f : R^n \to R$, 借助其梯度可判断一个方向是否为下降方向. 具体地, 设 $x \in R^n$, 若 $d \in R^n$ 满足 $d^{\mathrm{T}}\nabla f(x) < 0$, 则对充分小的正数 α,

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

因此, d 是目标函数 $f(x)$ 在 x 点的下降方向. 特别地, 当搜索方向取负梯度方向时, 该搜索方向为目标函数在该点函数值下降最快的方向, 故称为最速下降方向. 由于搜索方向的下降性大小与其长度无关, 所以有时将搜索方向设置成单位长度.

搜索方向确定后, 需要通过线搜索, 也就是计算函数 $f(x_k + \alpha d_k)$ 关于 $\alpha > 0$ 的 (近似) 最小值求得步长. 一般地, 该算法产生的迭代点列对应的目标函数值数列是单调下降的, 因此线搜索方法又称为下降算法.

线搜索方法的核心是搜索方向的选取和迭代步长的计算. 但就它们对算法效率的影响力而言, 搜索方向要大于迭代步长. 也就是说, 在线搜索过程中, 方向比速度重要.

与线搜索方法不同, 信赖域方法是利用目标函数 $f(x)$ 在 x_k 点的信息构造二次模型 $m_k(d)$, 使其在 x_k 点附近与 $f(x)$ 有好的近似, 然后根据该二次模型的最小值点来产生下一迭代点, 并视二次模型与目标函数的近似程度来调整信赖域半径的大小.

具体地, 先求二次模型 $m_k(d)$ 在信赖域内的最小值点 d_k, 即求解子问题

$$\min \{m_k(d) \mid d \in R^n,\ \|d\| \leqslant \Delta_k\},$$

其中, $\Delta_k > 0$ 为信赖域半径. 如果试探点 $\hat{x}_{k+1} = x_k + d_k$ 能使目标函数值有 “充

分” 的下降, 就取 $x_{k+1} = \hat{x}_{k+1}$. 如果近似效果特好, 在下一步就扩大信赖域半径; 否则, 就压缩信赖域半径, 重新求解信赖域子问题.

一般地, 二次模型 $m_k(d)$ 取如下形式

$$m_k(d) = f(x_k) + d^{\mathrm{T}}\nabla f(x_k) + \frac{1}{2}d^{\mathrm{T}}B_k d,$$

其中,B_k 取为 $\nabla^2 f(x_k)$ 或其近似.

无约束优化问题的信赖域方法最早由 Powell (1970) 提出, 而后得到广泛研究. 后来, Davidon(1980) 在二次模型的基础上提出了信赖域方法的锥模型. 信赖域方法不如线搜索那样成熟, 应用也没有线搜索那样广泛. 但由于其强的收敛性和可靠性, 信赖域方法的研究越来越受到重视. 从本质上讲, 它和线搜索方法的区别在于线搜索方法是借助搜索方向将一个多元函数的极值问题转化为一单元函数的极值问题, 而信赖域方法是在一值得 “信赖” 的区域内将复杂的目标函数用一个简单的二次函数近似.

由于梯度型数值方法在迭代过程中过多地依赖约束函数和目标函数在已产生点的函数值信息和梯度信息, 而这些信息只能反映函数值的局部变化情况, 因而梯度型方法大多只能得到问题的局部最优解. 若求全局最优解, 则需要用多个初始点分别进行计算, 然后在得到的多个局部最优解中取其最优者当做全局最优解. 另外, 也可利用隧道 (Levy et al.,1985) 和填充函数 (Ge,1990) 等技术由局部最优解向全局最优解一步步靠近. 但相对于局部优化数值算法, 全局优化算法还不成熟, 因为人们至今还没有找到一个令人满意的全局最优解的有效算法和检验准则. 正因如此, 在本书以后的叙述中, 除非特别说明, 我们对全局最优解和局部最优解不再严格区分, 而泛泛地称之为最优解.

对于非线性最优化问题, 一个数值方法要被认可, 既要有理论保障, 又要有满意的数值效果. 具体地, 一个好的数值方法应对如下指标有好的特性.

(1) *全局收敛与局部收敛*. 对梯度型数值方法, 由于很难保证在有限步内得到问题的最优解, 故人们希望算法产生的迭代点有越来越靠近最优解的趋势, 这便引出了算法收敛性的概念.

如果从任意的初始点出发, 算法产生的迭代点列都收敛到问题的最优值点, 称该算法具有全局收敛性. 若算法只有在初始点和最优值点具有某种程度的靠近时才能保证迭代点列收敛到最优值点, 则称该算法具有局部收敛性. 此外, 若算法产生的迭代点列的某一聚点为优化问题的最优值点, 则称该算法具有弱收敛性.

需要强调的是, 无论是全局收敛还是局部收敛, 这都属于理论分析, 因为在进行实际数值计算时, 算法必须在有限步内终止, 而我们也只能在计算机运行机时的许可范围内得到满足一定精度要求的近似最优解.

(2) **收敛速度与二次终止性**. 大量数值实验表明：一个算法的计算效率在很大程度上依赖于在最优值点附近迭代点靠近最优值点的速度. 也就是说, 一个数值方法高效的基本标志就是一旦迭代点进入目标函数的一个 “狭长的凹谷”, 那么以后产生的迭代点应迅速移向该 “凹谷” 的最低点. 对此, 从收敛速度的角度进行分析.

算法的收敛速度主要考虑迭代点列 $\{x_k\}$ 与最优值点 x^* 之间的距离范数所确定的数列 $\{\|x_k - x^*\|\}$ 趋于零的速度. 所以讨论迭代点列 $\{x_k\}$ 的收敛速度的前提是该点列收敛到某最优值点 x^*. 显然, 数列 $\{\|x_k - x^*\|\}$ 趋于零的速度越快, 相应算法的效率就越高. 对此, 有以下两种衡量尺度：Q-收敛和 R-收敛 (Ortega, Rheinboldt, 1970).

Q-收敛是通过前后两迭代点靠近最优值点的程度之比定义的：设点列 $\{x_k\}$ 收敛到 x^*, 且存在 $q \geqslant 0$ 满足

$$\limsup_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leqslant q.$$

若 $0 < q < 1$, 称 $\{x_k\}$ Q-线性收敛到 x^*. 若 $q = 0$, 称 $\{x_k\}$ Q-超线性收敛到 x^*.

容易证明：如果点列 $\{x_k\}$ Q-超线性收敛到 x^*, 则

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x_k\|}{\|x_k - x^*\|} = 1. \tag{1.2.1}$$

对收敛到 x^* 的点列 $\{x_k\}$, 若存在 $0 \leqslant p < \infty$ 和 $r \geqslant 1$, 使

$$\limsup_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^r} \leqslant p,$$

则称 $\{x_k\}$ Q-r 阶收敛到 x^*, 有时简单地称 $\{x_k\}$ r- 阶收敛到 x^*. 其中最常见的是 2- 阶收敛. 若 $r > 1$, r- 阶收敛必为超线性收敛.

与 Q- 收敛不同, R- 收敛是借助一个收敛于零的数列来度量 $\{\|x_k - x^*\|\}$ 趋于零的速度：设点列 $\{x_k\}$ 收敛到最优值点 x^*. 若存在 $\kappa \in (0, \infty)$, $q \in (0, 1)$, 使

$$\|x_k - x^*\| \leqslant \kappa q^k,$$

则称 $\{x_k\}$ R- 线性收敛到 x^*.

对上述点列, 若存在 $\kappa \in (0, \infty)$ 和收敛于零的正数列 $\{q_k\}$, 使

$$\|x_k - x^*\| \leqslant \kappa \prod_{i=0}^{k} q_i,$$

则称点列 $\{x_k\}$ R- 超线性收敛到 x^*.

这里, Q 和 R 分别取自英文单词 “Quotient” 和 “Root” 的第一个字母. 在这些收敛速度中, 超线性收敛比线性收敛速度快. 另外, 若一个点列 Q-(超) 线性收敛, 则它必 R-(超) 线性收敛.

收敛速度用来刻画迭代点列靠近问题最优解的快慢程度. 一般地, 具有超线性收敛性或二阶收敛速度的算法是比较快的. 但由于计算误差和算法程序本身带来的影响, 算法收敛速度的理论结果并不能保证算法在进行数值计算时有同样的数值效果.

另外, 算法的二次终止性也是判断算法优劣的一个重要指标. 它指对于任意严格凸二次函数, 从任意初始点出发, 算法都能经过有限步达到其最优值点.

由于严格凸二次函数是非线性函数中形式最简单、条件最强的函数, 所以一个好的算法理应在有限步内到达最优解. 其次, 对于一般的目标函数, 它在最优值点附近可以用一个严格凸二次函数来拟合. 因此, 可以猜想, 对于严格凸二次函数数值效果好的算法, 对于一般的目标函数也应具有好的数值效果. 无约束优化问题的共轭梯度法和拟牛顿方法之所以有这么强的吸引力, 主要原因就在于它们的二次终止性.

(3) *稳定性*. 算法的稳定性, 实际上是数值方法的可靠性. 在数值计算过程中, 初始数据的舍入误差会通过系列运算进行遗传和传播. 如果初始数据的误差对最终结果的影响较小, 即在计算过程中舍入误差增长缓慢, 称该算法是稳定的. 若输出结果的误差随初始数据的舍入误差呈恶性增长, 则称该算法是不稳定的.

一般地, 数值稳定性是对算法而言, 但有时也与问题本身有关. 对所谓的病态问题, 如果输入数据有微小扰动, 则问题的解会产生大的扰动. 这种情况是由问题本身的性质决定的, 与算法无关. 在这方面最简单的例子是线性方程组的求解. 如果系数矩阵的条件数过大, 那么在计算过程中, 数据存储的舍入误差可能会引起计算结果大的偏差. 这就是说, 对于病态问题, 用任何算法直接计算都会产生不稳定性. 对此, 人们常用调比技术对问题进行某种预处理或在算法中引入正则化技术来增强算法的稳定性.

(4) *计算复杂性和存储消耗*. 最优化问题的所有数值方法最终都要在计算机上实现. 一个算法在理论上有快的收敛速度是保证其高效的一个因素, 而算法中每一迭代步的计算量和存储量也是影响算法效率的重要因素. 因为即便一个算法有快的收敛速度, 但若其每一迭代步的计算量或存储量偏大, 则会导致算法的迭代进程变慢, 从而影响算法的整体效率.

(5) *自适应性和自我校正能力*. 算法的自适应性主要指算法的有关参数能否根据算法的运行状况自动调整, 从而使迭代点列能够快速靠近或到达问题的最优值点; 算法的自我校正能力是指当算法在迭代过程中停滞不前时, 算法无需借助外力而能通过自我调整使迭代点列快速转向问题的最优值点. 显然, 具有上述性能的算

法一般具有高的效率.

上述五个指标主要侧重算法的理论分析. 它一方面使我们清楚算法对良态问题所具有的诱人性质和对病态问题可能出现的最坏结果, 从而找到算法所适用的问题类; 另一方面使我们明白为什么 “这样” 取初始点, “那样” 选取参数, 同时它还帮助我们发现算法中的缺陷, 进而改进之. 只是人们在借助数学分析等工具对算法进行理论分析的时候, 一般要对问题或其解点做些假设, 而这些假设有时难于验证.

(6) *数值效果*. 对非线性最优化问题的数值方法, 进行数值实验是非常重要也是非常必要的. 首先, 算法本身就是为问题求解设计的. 因此, 一个方法最终能否被接受和认可关键在于其数值效果, 而不是算法设计的难度、技巧或好的理论性质. 其次, 数值实验虽不能给算法的理论分析提供什么保证, 但有时会很可靠地显露出某些可能的理论结果. 只是在进行数值分析的时候, 需要考虑到参数和初始点的选取对数值效果的影响, 同时还要考虑到算法程序中某些微小的变动对数值效果的影响.

显然, 一个理想的数值方法应具有良好的理论性质, 同时又有诱人的数值效果. 遗憾的是, 如同线性规划问题的单纯形方法和椭球算法, 非线性最优化问题的有些算法的理论性质和数值效果也不一致. 这其中的原因很复杂, 既有计算过程中数据舍入误差和参数取值的影响, 也有理论分析过程中所需条件在实际问题中得不到满足的因素. 同时, 对同一算法, 其性能指标与具体的问题有关系, 对此很难找到统一的量化指标.

从 20 世纪 50 年代至今, 人们提出了求解非线性最优化问题的各式各样的数值方法, 但目前尚未找到一个理论性质和数值效果都十分令人满意的通用算法. 一般情况下, 人们只能宣称某个方法对某类问题比较有效. 这也是在非线性最优化问题的数值方法研究中多种方法并存的主要原因.

1.3 凸集与凸函数

在非线性最优化问题的理论分析中, 常用到凸集和凸函数的概念, 下面给出定义和有关性质.

定义 1.3.1 称 $S \subset R^n$ 为凸集, 若对任意的 $x_1, x_2 \in S$ 和任意的 $\lambda \in [0,1]$, $\lambda x_1 + (1-\lambda)x_2 \in S$.

根据定义, 对凸集中的任意两点, 它们的连线都在集合中. 不但如此, 凸集中多个元素的凸组合也属于该集合.

定义 1.3.2 设 $x_1, x_2, \cdots, x_n \in R^n$, $\lambda_i \geqslant 0, i = 1, 2, \cdots, n$ 满足 $\lambda_1 + \lambda_2 + \cdots +$

$\lambda_n = 1$, 则称 $\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$ 为 $x_1, x_2, \cdots, x_n$ 的一个凸组合.

若一个凸集是由有限个元素的凸组合所生成, 则称其为多面胞. 除去凸组合中系数的非负性条件, 便得到仿射组合的定义并可以建立仿射集的概念.

定义 1.3.3 设 $x_1, x_2, \cdots, x_n \in R^n$, $\lambda_i \in R, i = 1, 2, \cdots, n$ 满足 $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$, 则称 $\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$ 为 $x_1, x_2, \cdots, x_n$ 的一个仿射组合.

定义 1.3.4 设 $S \subset R^n$, 由集合 S 中点的所有的仿射组合所组成的集合称为 S 的仿射包, 记为 $\mathrm{Aff}(S)$. 若 $S = \mathrm{Aff}(S)$, 则称 S 为仿射集.

根据定义, 仿射集中任意两点之间的连线及其延伸所张成的区域都属于该集合. 因此, 仿射集是凸集. 由于仿射组合 $\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$ 可以写成

$$x_1 + \lambda_2 (x_2 - x_1) + \cdots + \lambda_n (x_n - x_1),$$

所以, 如果 $S \subset R^n$ 为仿射集, 则对任意 $x \in S$, 集合 $S - \{x\}$ 为 R^n 的一个子空间. 也就是说, 仿射集是子空间的一个平移. 因此, 仿射集又称仿射子空间, 而子空间 $S - \{x\}$ 的维数称为仿射集 S 的维数.

容易验证, 齐次线性方程组 $Ax = 0$ 的解集是一个线性子空间, 而非齐次线性方程组 $Ax = b$ 的解集是一个仿射空间. 设 x_0 为 $Ax = b$ 的一个特解, 则将子空间 $\mathcal{N}(A)$ 平移至 x_0 点便得到 $Ax = b$ 的解集 $\{x_0\} + \mathcal{N}(A)$.

在欧氏空间中, 一个集合的内点是借助 δ- 邻域定义的, 而借助仿射包可将集合的内点进行推广.

定义 1.3.5 称 $x \in S$ 为集合 $S \subset R^n$ 的相对内点, 若存在 $\delta > 0$, 使得 $N(x, \delta) \cap \mathrm{Aff}(S) \subset S$, 即 x 在仿射子空间 $\mathrm{Aff}(S)$ 中是集合 S 的内点. 集合 S 的所有相对内点所组成的集合记为 $\mathrm{ri}(S)$.

下面看两类特殊的凸集.

定义 1.3.6 称 $\mathcal{K} \subset R^n$ 为锥, 若对于任意的 $x \in \mathcal{K}$ 和 $\lambda \geqslant 0$, 都有 $\lambda x \in \mathcal{K}$. 若锥 $\mathcal{K}$ 为凸集, 则称 $\mathcal{K}$ 为凸锥. 进一步, 若锥 $\mathcal{K}$ 中的任一元素都可以表示成固定有限个元素的非负组合, 也就是

$$\mathcal{K} = \left\{ \sum_{i=1}^{m} \mu_i b_i \mid \mu_i \geqslant 0,\ i = 1, 2, \cdots, m \right\},$$

其中, $b_1, \cdots, b_m \in R^n$, 则称该锥为有限生成锥.

设 $\mathcal{K}$ 是闭凸锥, 则对任意的 $x, y \in \mathcal{K}$ 和任意的 $\lambda \geqslant 0$, $\mu \geqslant 0$,

$$\lambda x + \mu y \in \mathcal{K}.$$

称 $\lambda x + \mu y$ 为 x, y 的锥组合, 记为 $\mathrm{cone}\{x, y\}$. 与锥密切相关的是它的极锥, 即

$$\mathcal{K}^\circ = \{y \in R^n \mid \langle x, y\rangle \leqslant 0,\ \forall\ x \in \mathcal{K}\}.$$

对矩阵 $A \in R^{m\times n}$, 其核空间 $\mathcal{N} = \{x \in R^n \mid Ax = 0\}$ 是一种特殊的锥. 容易验证其极锥为 $\mathcal{R}(A^{\mathrm{T}})$. 一个更一般的结论是子空间的极锥为其正交补空间.

定义 1.3.7 设 $A \in R^{m\times n}, b \in R^m$, 称集合 $S = \{x \in R^n \mid Ax \geqslant b\}$ 为多面体. 特别地, 集合 $\mathcal{K} = \{x \in R^n \mid Ax \geqslant 0\}$ 称为多面锥.

容易验证, 多面锥和有限生成锥等价, 且多面锥 $\mathcal{K} = \{x \in R^n \mid Ax \geqslant 0\}$ 的极锥为

$$\mathcal{K}^\circ = \{-A^{\mathrm{T}} y \mid y \in R^m_+\}.$$

由于多面体本身是凸集, 故又称其凸多面体. 有界的凸多面体就是多面胞. 如果凸多面体 S 无界, 则存在 $d \in R^n$, 使对任意的 $x \in S$ 和 $\alpha \geqslant 0$, 有 $x + \alpha d \in S$. 称这样的 d 为 S 的回收方向, S 的所有回收方向所构成的锥称为 S 的回收锥. 基于多面胞和凸多面体的回收锥可得到凸多面体的 Minkowski 分解定理.

定理 1.3.1 若多面体 S 无界, 则存在多面胞 P 和多面锥 $\mathcal{K}$, 使 $S = P + \mathcal{K}$.

证明 设 $S = \{x \in R^n \mid Ax \geqslant b\}$. 根据多面锥和有限生成锥的等价性, 多面锥

$$\mathcal{K}_1 = \left\{ \begin{pmatrix} x \\ \lambda \end{pmatrix} \in R^{n+1} \;\middle|\; Ax - \lambda b \geqslant 0 \right\}$$

可以表示成

$$\begin{pmatrix} x_1 \\ \lambda_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ \lambda_2 \end{pmatrix}, \cdots, \begin{pmatrix} x_m \\ \lambda_m \end{pmatrix}$$

的有限生成锥. 根据锥的性质, 可设 $\lambda_i = 0, 1$. 不失一般性, 设 $\lambda_i = 1,\ i = 1, 2, \cdots, m_1$; $\lambda_i = 0,\ i = m_1 + 1, m_1 + 2, \cdots, m$, 并记 P 为由 $x_1, x_2, \cdots, x_{m_1}$ 所生成的多面胞, $\mathcal{K}$ 为由 $x_{m_1+1}, x_{m_1+2}, \cdots, x_m$ 生成的多面锥.

对任意的 $x \in S$, 显然有 $\begin{pmatrix} x \\ 1 \end{pmatrix} \in \mathcal{K}_1$, 即

$$\begin{pmatrix} x \\ 1 \end{pmatrix} \in \mathrm{cone}\left\{ \begin{pmatrix} x_1 \\ \lambda_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ \lambda_2 \end{pmatrix}, \cdots, \begin{pmatrix} x_m \\ \lambda_m \end{pmatrix} \right\}.$$

从而存在非负数组 μ 满足

$$x=\sum_{i=1}^{m_1}\mu_i x_i+\sum_{i=m_1+1}^{m}\mu_i x_i,\qquad \sum_{i=1}^{m}\mu_i\lambda_i=\sum_{i=1}^{m_1}\mu_i=1.$$

根据 P 和 $\mathcal{K}$ 的定义知 $x\in P+\mathcal{K}$. 证毕

实际上, 上述结论的逆命题也成立, 即若集合 S 可以表示成一个多面胞和一个多面锥的和, 则 S 是凸多面体.

凸集在约束优化问题的理论分析中起着非常重要的作用. 在后面的章节中, 会进一步介绍其有关性质. 下面看凸函数的概念和性质.

对于非线性最优化问题, 一般很难保证其局部最优值点是全局最优值点, 主要原因在于目标函数在定义域上是 "多峰" 函数, 而凸函数可以保证目标函数在定义域上为 "单峰" 函数.

定义 1.3.8 称函数 $f:R^n\to R$ 是凸函数, 若满足

$$f\left(\lambda x+(1-\lambda)y\right)\leqslant\lambda f(x)+(1-\lambda)f(y),\quad\forall\, x,y\in R^n,\ \lambda\in[0,1].$$

根据定义, 若 $f(x)$ 连续可微, 则它为凸函数的充分必要是下述条件之一:

$$f(y)-f(x)\geqslant\nabla f(x)^{\mathrm{T}}(y-x),\quad\forall\, x,y\in R^n;$$

$$(\nabla f(y)-\nabla f(x))^{\mathrm{T}}(y-x)\geqslant 0,\quad\forall\, x,y\in R^n.$$

进一步, 若 $f(x)$ 二阶连续可微, 则它为凸函数等价于

$$h^{\mathrm{T}}\nabla^2 f(x)h\geqslant 0,\quad\forall\, x,h\in R^n.$$

若上述不等式取严格不等号, 则称函数 $f(x)$ 为严格凸的. 将上述各条件进一步加强得到一致凸函数的定义.

定义 1.3.9 称函数 $f:R^n\to R$ 为一致凸函数, 若存在 $\eta>0$, 使对任意 $x,y\in R^n$ 和 $\lambda\in[0,1]$, 成立

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

根据定义, 若 $f(x)$ 连续可微, 则它为一致凸函数的充分必要条件是

$$(y-x)^{\mathrm{T}}(\nabla f(y)-\nabla f(x))\geqslant\eta\|y-x\|^2,\quad\forall\, x,\ y\in R^n.$$

进一步, 若函数 f 二阶连续可微, 则它为一致凸函数的充分必要条件是

$$h^{\mathrm{T}}\nabla^2 f(x)h \geqslant \eta\|h\|^2, \quad \forall\ x,\ h \in R^n.$$

显然, 一致凸函数必严格凸, 而对于二次函数, 严格凸和一致凸等价. 利用凸函数可建立一些常用的不等式.

对于凸函数 $f : R \to R$, 容易推出

$$f\Big(\sum_{i=1}^{n}\alpha_i x_i\Big) \leqslant \sum_{i=1}^{n}\alpha_i f(x_i),$$

其中, $x_i \in R,\ \alpha_i \geqslant 0,\ i = 1, 2, \cdots, n,\ \sum\limits_{i=1}^{n}\alpha_i = 1$. 该不等式称为 Jensen 不等式.

若取 $f(x) = -\ln x$, $x \in (0, \infty)$, 则得到加权的算术几何不等式

$$\sum_{i=1}^{n}\alpha_i x_i \geqslant \prod_{i=1}^{n} x_i^{\alpha_i},$$

其中, $x_i \geqslant 0, \alpha_i \geqslant 0,\ i = 1, 2, \cdots, n,\ \sum\limits_{i=1}^{n}\alpha_i = 1$. 特别地, 若取 $\alpha_i = \dfrac{1}{n}$, 则得到算术几何不等式

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geqslant \Big(\prod_{i=1}^{n} x_i\Big)^{\frac{1}{n}}.$$

这里, 等号成立的充分必要条件是 $x_1 = x_2 = \cdots = x_n$.

在非线性最优化问题的理论分析中, 常用到上述公式和下面的 Cauchy-Schwarz 不等式

$$\sum_{i=1}^{n} x_i y_i \leqslant \Big(\sum_{i=1}^{n} x_i^2\Big)^{\frac{1}{2}}\Big(\sum_{i=1}^{n} y_i^2\Big)^{\frac{1}{2}}.$$

将其写成向量形式, 就是

$$x^{\mathrm{T}} y \leqslant \|x\| \cdot \|y\|.$$

1.4 无约束优化最优性条件

考虑无约束优化问题

$$\min_{x \in R^n} \quad f(x), \tag{1.4.1}$$

其中, 目标函数 $f : R^n \to R$ 连续可微. 根据定义, 要验证一个点是否为上述优化问题的局部最优值点, 就要逐一比较该点的目标函数值和附近所有点的目标函数值的大小, 其工作量不言而喻. 但如果利用目标函数连续可微的性质, 就得到一个比较实用的判断方法, 这就是无约束优化问题的最优性条件.

定理 1.4.1(一阶必要条件) 若 x^* 是无约束优化问题 (1.4.1) 的局部最优值点, 则 $\nabla f(x^*) = 0$.

证明 假若 $\nabla f(x^*) \neq 0$, 取 $d = -\nabla f(x^*)$. 则对 $\alpha > 0$ 充分小, 利用 Taylor 展式,

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

这与 x^* 是局部最优解矛盾. 命题结论得证. 证毕

目标函数梯度为零的点称为无约束优化问题的稳定点. 稳定点可能是目标函数的极大值点也可能是极小值点, 甚至二者都不是. 最后一种情况对应的稳定点称为函数的鞍点, 即在从该点出发的一个方向上是函数的极大值点, 而在另一个方向上是极小值点. 如单元函数 $f(x) = x^3$ 的稳定点就是鞍点.

定理 1.4.1 表明, 对无约束优化问题, 目标函数在最优值点的任意方向上的导数都为零, 即目标函数在最优值点的切平面是水平的. 不过, 无约束优化问题的局部最大值点和鞍点也满足上述优性条件. 因此, 要确认一个稳定点是否为最优值点, 需考虑该点的二阶最优性条件.

定理 1.4.2 (二阶必要条件) 设 $x^* \in R^n$ 是无约束优化问题 (1.4.1) 的局部最优解, 且 $f(x)$ 在 x^* 点附近二阶连续可微. 则 $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ 半正定.

证明 由定理 1.4.1 知 $\nabla f(x^*) = 0$. 若 $\nabla^2 f(x^*)$ 非半正定, 则存在单位向量 $d \in R^n$, 使 $d^{\mathrm{T}} \nabla^2 f(x^*) d < 0$. 对 $\alpha > 0$ 充分小, 利用 Taylor 展式,

$$f(x^* + \alpha d) = f(x^*) + \alpha \nabla f(x^*)^{\mathrm{T}} d + \frac{1}{2} \alpha^2 d^{\mathrm{T}} \nabla^2 f(x^*) d + o(\alpha^2) < f(x^*),$$

这与 x^* 是局部最优解矛盾. 命题结论得证. 证毕

需要指出的是, 上述结论给出的最优性条件不是充分的. 如单元函数 $f(x) = x^3$ 在 $x = 0$ 点同时满足一阶和二阶最优性必要条件, 但它并不是该函数的局部最优值点.

定理 1.4.3 (二阶充分条件) 设 $x^* \in R^n$ 满足 $\nabla f(x^*) = 0$ 且 $\nabla^2 f(x^*)$ 正定, 则 x^* 是无约束优化问题 (1.4.1) 的严格局部最优解.

证明 对任意充分靠近 x^* 的 $x \in R^n$, 存在单位向量 $d \in R^n$ 及充分小的 $\alpha > 0$

使 $x = x^* + \alpha d$. 由 Taylor 展式及 $\nabla^2 f(x^*)$ 的正定性得

$$f(x) = f(x^*) + \frac{1}{2}\alpha^2 d^{\mathrm{T}} \nabla^2 f(x^*) d + o(\alpha^2) > f(x^*).$$

从而 x^* 是问题 (1.4.1) 的严格局部最优解. 证毕

同样, 上述结论给出的二阶充分性条件也不是必要的. 如 $x = 0$ 是单元函数 $f(x) = x^4$ 的严格局部最优解, 但上述二阶充分性条件在该点并不成立. 由此推断, 无约束优化问题不存在二阶充分必要的最优性条件.

无约束优化问题的二阶最优性条件是借助目标函数在局部最优值点邻域上的凸性来刻画的. 一般地, 非线性最优化的数值方法只能保证求得的点满足一阶必要条件, 而不能保证满足二阶充分条件. 换句话讲, 这些方法只能得到稳定点, 只有在特殊情况下, 如目标函数为凸函数, 才能保证问题的稳定点为其全局最优解.

定理 1.4.4 对无约束优化问题 (1.4.1), 若目标函数 f 是连续可微的凸函数, 则全局最优解、局部最优解和稳定点等价.

证明 若 x^* 是局部最优解, 显然有 $\nabla f(x^*) = 0$, 从而 x^* 是稳定点. 反过来, 若 x^* 是稳定点, 利用凸函数的性质, 对任意 $x \in R^n$,

$$f(x) - f(x^*) \geqslant \langle \nabla f(x^*), x - x^* \rangle = 0.$$

故 x^* 是问题 (1.4.1) 的全局最优解. 显然, 全局最优解必是局部最优解. 证毕

上述结论说明, 对无约束优化问题 (1.4.1), 若目标函数为凸函数, 满足 $\nabla f(x^*) = 0$ 的点 x^* 即为其全局最优值点.

习 题

1. (1) 设 $x, y \geqslant 0$, $\alpha \in (0, 1)$, 则 $x^\alpha y^{1-\alpha} \leqslant \alpha x + (1-\alpha) y$.

(2) 设 $x_i > 0, i = 1, 2, \cdots, n$. 证明

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \frac{1}{x_i} \geqslant n^2,$$

且等号成立的充分必要条件是所有的 x_i 都相等.

2. 试用 Cauchy-Schwarz 不等式证明

$$\|x\|_1 \leqslant \sqrt{n}\|x\|, \quad \forall\, x \in R^n.$$

3. 设 $b \in R^n$, $Q \in R^{n\times n}$ 对称正定. 试求下述优化问题的最优解和最优值

$$\max\ b^{\mathrm{T}} x, \quad \text{s.t. } x^{\mathrm{T}} Q x \leqslant 1,$$

并利用该结果证明对任意的 $x, y \in R^n$ 有

$$(x^{\mathrm{T}}y)^2 \leqslant (x^{\mathrm{T}}Qx)(y^{\mathrm{T}}Q^{-1}y).$$

4. 用图解法求下述优化问题的最优解：

$$\min x_1 + x_2, \quad \text{s.t. } x_1^2 + x_2^2 \leqslant 2.$$

5. 证明函数 $f : R^n \to R$ 为凸函数的充分必要条件是函数 f 的上图为凸集, 其中上图定义为

$$\mathrm{epi}f = \{(x, y) \mid x \in R^n, y \geqslant f(x)\}.$$

6. 讨论由 $a_1, a_2, \cdots, a_m \in R^n$ 所生成的多面胞、仿射集、子空间及这些点和原点所生成的仿射包之间的关系.

7. 设 $d \in R^n$ 是连续可微函数 $f : R^n \to R$ 在点 $x \in R^n$ 的下降方向. 试建立 α 为单元函数 $f(x + \alpha d)$ 在 $\alpha \geqslant 0$ 上的最小值点的一个必要条件, 并讨论在什么条件下该条件是充分的.

8. 设 $A \in R^{m\times n}$, $b \in R^m$. 试给出无约束优化问题

$$\min_{x\in R^n} \|Ax - b\|^2$$

的一阶最优性条件, 并验证该条件是否是充分的. 它的最优解唯一吗?

9. 试在 3 维欧式空间中确定一通过点 $(3, 4, 5)$ 的平面, 使其与非负象限中的三个坐标面构成的四面体的体积最小.

10. 设 $b \in R^n$, $A \in R^{n\times n}$. 试给出下述优化问题的最优性条件

$$\min \frac{1}{2}x^{\mathrm{T}}Ax + b^{\mathrm{T}}x.$$

第 2 章　线搜索方法与信赖域方法

线搜索方法是求解无约束优化问题的一个最基本的方法, 它具有简单、可靠等优点. 与线搜索方法相比, 信赖域方法发展较晚, 但也已构成非线性最优化问题的一种基本方法. 本章主要介绍线搜索方法的几种常见步长规则及收敛性质, 信赖域方法的基本框架和收敛性质.

2.1　精确线搜索方法

假设目标函数 $f: R^n \to R$ 至少一阶连续可微. 除非特别说明, 记

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

在线搜索方法中, 令搜索方向 d_k 为目标函数在当前迭代点的下降方向. 若步长取

$$\alpha_k = \arg\min_{\alpha \geqslant 0} f(x_k + \alpha d_k),$$

则称 α_k 为精确步长, 又称最优步长. 该步长规则称为精确线搜索步长规则, 又称最优步长规则. 最优步长的一个重要性质是它满足如下正交性条件：

$$d_k^{\mathrm{T}} \nabla f(x_k + \alpha_k d_k) = 0,$$

该步长规则对应的下降算法模型如下.

算法 2.1.1

步 1. 取初始点 $x_0 \in R^n$ 和参数 $\varepsilon \geqslant 0$. 令 $k = 0$.

步 2. 若 $\|g_k\| \leqslant \varepsilon$, 算法终止; 否则, 进入下一步.

步 3. 计算下降方向 d_k, 使 $d_k^{\mathrm{T}} g_k < 0$.

步 4. 计算步长 $\alpha_k = \arg\min\{f(x_k + \alpha d_k) \mid \alpha \geqslant 0\}$.

步 5. 令 $x_{k+1} = x_k + \alpha_k d_k$; 令 $k = k + 1$, 转步 2.

尽管最优步长的计算是一单元函数的极值问题, 它也是很难求的. 常用的方法主要有进退试探法、黄金分割法和多项式插值法等, 但一般也只能得到近似最优步长.

对算法的终止规则, 若取 $\varepsilon = 0$, 则算法会产生无穷迭代点列. 因此, 在数值计

算时应当避免. 但在算法的收敛性分析中, 却是允许和必要的, 因为算法的理论分析需要问题的精确解.

定理 2.1.1 设目标函数 $f: R^n \to R$ 二阶连续可微有下界. 对算法 2.1.1, 设 d_k 与 $-g_k$ 的夹角 θ_k 满足 $\theta_k \leqslant \pi/2 - \mu$, 其中, $0 < \mu \leqslant \pi/2$. 若算法产生迭代无穷迭代点列 $\{x_k\}$, 且存在常数 $M > 0$, 使对任意的 k 及 $\alpha > 0$,

$$\|\nabla^2 f(x_k + \alpha d_k)\| \leqslant M,$$

则 $\lim\limits_{k\to\infty} \|g_k\| = 0$.

证明 对任意 $\alpha > 0$ 和 $k \geqslant 0$, 利用假设条件, 存在 $\xi_k \in (x_k, x_k + \alpha d_k)$, 使得

$$\begin{aligned} f(x_k + \alpha d_k) - f_k &= \alpha d_k^{\mathrm{T}} g_k + \frac{1}{2}\alpha^2 d_k^{\mathrm{T}} G(\xi_k) d_k \\ &\leqslant \alpha d_k^{\mathrm{T}} g_k + \frac{1}{2}\alpha^2 M \|d_k\|^2 \\ &= \frac{1}{2} M \|d_k\|^2 \Big(\alpha + \frac{d_k^{\mathrm{T}} g_k}{M\|d_k\|^2}\Big)^2 - \frac{1}{2}\frac{(d_k^{\mathrm{T}} g_k)^2}{M\|d_k\|^2} \\ &= \frac{1}{2} M \|d_k\|^2 \Big(\alpha + \frac{d_k^{\mathrm{T}} g_k}{M\|d_k\|^2}\Big)^2 - \frac{1}{2M}\|g_k\|^2 \cos^2(d_k, -g_k). \end{aligned}$$

显然, 当 $\alpha = \hat{\alpha}_k \triangleq -\dfrac{d_k^{\mathrm{T}} g_k}{M\|d_k\|^2}$ 时, 最后一式取最小值. 利用步长规则得

$$f_{k+1} - f_k \leqslant f(x_k + \hat{\alpha}_k d_k) - f_k \leqslant -\frac{1}{2M}\|g_k\|^2 \cos^2(d_k, -g_k).$$

由于数列 $\{f_k\}$ 单调下降有下界, 故必有极限. 从而将上式两边对 k 求和得

$$\sum_{k=1}^{\infty} \|g_k\|^2 \cos^2(d_k, -g_k) < \infty.$$

利用 $\cos(d_k, -g_k) \geqslant \cos(\pi/2 - \mu) = \sin\mu > 0$ 得

$$\lim_{k\to\infty} \|g_k\| = 0.$$ 证毕

上述证明过程给出了目标函数在每一迭代步的下降量估计. 进一步, 若目标函数 $f(x)$ 为一致凸函数 (常数为 η), 则目标函数值在每一迭代步的下降量有如下估计:

$$f(x_k) - f(x_k + \alpha_k d_k) = -\int_0^{\alpha_k} d_k^{\mathrm{T}} \nabla f(x_k + \tau d_k) \mathrm{d}\tau$$

$$
\begin{aligned}
&= \int_0^{\alpha_k} d_k^{\mathrm{T}} \big[\nabla f(x_k + \alpha_k d_k) - \nabla f(x_k + \tau d_k)\big] \mathrm{d}\tau \\
&\geqslant \int_0^{\alpha_k} \eta \|d_k\|^2 (\alpha_k - \tau) \mathrm{d}\tau \\
&= \frac{1}{2} \eta \|\alpha_k d_k\|^2. \qquad (2.1.1)
\end{aligned}
$$

若目标函数的梯度函数一致连续, 则有如下结论.

定理 2.1.2　设目标函数 f 在 R^n 上连续可微有下界, 梯度函数 ∇f 在包含水平集 $\mathcal{L}(x_0)$ 的某个邻域内一致连续. 对算法 2.1.1, 设搜索方向 d_k 与 $-g_k$ 的夹角 θ_k 满足 $\theta_k \leqslant \pi/2 - \mu$, 其中, $0 < \mu \leqslant \pi/2$. 若算法不有限步终止, 则 $\lim\limits_{k\to\infty} \|g_k\| = 0$.

证明　若命题结论不成立, 则存在 $\varepsilon_0 > 0$ 及自然数列 N 的一无穷子列 N_1, 使对任意 $k \in N_1$, 有 $\|g_k\| > \varepsilon_0$. 从而由

$$
\begin{aligned}
-d_k^{\mathrm{T}} g_k &= \|d_k\| \|g_k\| \cos\theta_k \\
&\geqslant \|d_k\| \|g_k\| \cos\left(\frac{\pi}{2} - \mu\right) \\
&= \|d_k\| \|g_k\| \sin\mu,
\end{aligned}
$$

得

$$
\frac{-d_k^{\mathrm{T}} g_k}{\|d_k\|} \geqslant \|g_k\| \sin\mu \geqslant \varepsilon_0 \sin\mu, \quad \forall\ k \in N_1. \qquad (2.1.2)
$$

对任意 $\alpha > 0$ 和 $k \in N_1$, 由微分中值定理, 存在 $\xi_k \in (x_k, x_k + \alpha d_k)$, 使得

$$
\begin{aligned}
f(x_k + \alpha d_k) - f_k &= \alpha d_k^{\mathrm{T}} \nabla f(\xi_k) \\
&= \alpha d_k^{\mathrm{T}} g_k + \alpha d_k^{\mathrm{T}} (\nabla f(\xi_k) - g_k) \\
&\leqslant \alpha d_k^{\mathrm{T}} g_k + \alpha \|d_k\| \|\nabla f(\xi_k) - g_k\| \\
&= \alpha \|d_k\| \Big(\frac{d_k^{\mathrm{T}} g_k}{\|d_k\|} + \|\nabla f(\xi_k) - g_k\| \Big). \qquad (2.1.3)
\end{aligned}
$$

不妨设 $\nabla f(x)$ 在包含水平集 $\mathcal{L}(x_0)$ 的 $\hat{\delta}$ 邻域上一致连续. 则对任意 $\varepsilon > 0$, 存在 $0 < \delta(\varepsilon) \leqslant \hat{\delta}$, 使当 $0 < \alpha\|d_k\| \leqslant \delta(\varepsilon)$ 时,

$$
\|\nabla f(\xi_k) - g_k\| \leqslant \varepsilon. \qquad (2.1.4)
$$

取 $\varepsilon = \dfrac{1}{2}\varepsilon_0 \sin\mu$, 则对任意 $k \in N_1$ 及 $\alpha = \dfrac{\delta(\varepsilon)}{\|d_k\|}$, 利用 (2.1.2)$\sim$(2.1.4) 及步长规则,

有

$$\begin{aligned} f_{k+1}-f_k &\leqslant f(x_k+\alpha d_k)-f_k \\ &\leqslant \alpha\|d_k\|\left(-\varepsilon_0\sin\mu+\frac{1}{2}\varepsilon_0\sin\mu\right) \\ &=-\frac{1}{2}\delta(\varepsilon)\varepsilon_0\sin\mu. \end{aligned}$$

从而,

$$\begin{aligned} \lim_{k\to\infty} f_k &=\sum_{k=0}^{\infty}(f_{k+1}-f_k)+f_0 \\ &\leqslant \sum_{k\in N_1}(f_{k+1}-f_k)+f_0 \\ &\leqslant \sum_{k\in N_1}\left(-\frac{1}{2}\varepsilon_0\delta(\varepsilon)\sin\mu\right)+f_0 \\ &=-\infty. \end{aligned}$$

这与目标函数 f 在 R^n 上有下界的假设矛盾. 结论得证. 证毕

算法 2.1.1 的上述收敛性结论是建立在目标函数的 Hesse 矩阵或梯度满足一定条件且搜索方向和负梯度方向成一定角度的基础上. 若无此假设, 则有如下结论.

定理 2.1.3 设目标函数 f 在 R^n 上连续可微, 且算法 2.1.1 产生的无穷迭代点列的某一子列 $\{x_k\}_{k\in N_0}$ 满足

$$\lim_{\substack{k\in N_0\\ k\to\infty}} x_k=x^*,\quad \lim_{\substack{k\in N_0\\ k\to\infty}} d_k=d^*,$$

则 $g(x^*)^{\mathrm{T}}d^*=0$. 进一步, 若目标函数 $f(x)$ 二阶连续可微, 则 $(d^*)^{\mathrm{T}}\nabla^2 f(x^*)d^*\geqslant 0$.

证明 若 $d^*=0$, 则命题结论自然成立. 下面考虑 $d^*\neq 0$ 的情况.

对第一个结论, 假若它不成立, 则存在 $\varepsilon_0>0$ 使 $\nabla f(x^*)^{\mathrm{T}}d^*<-\varepsilon_0<0$. 由于函数 $f(x)$ 连续可微, 故存在 x^* 点的邻域 $N(x^*,\delta)$ 及 d^* 的邻域 $N(d^*,\delta)$, 使对任意 $x\in N(x^*,\delta)$ 及 $d\in N(d^*,\delta)$ 有

$$\nabla f(x)^{\mathrm{T}}d\leqslant -\frac{\varepsilon_0}{2}<0. \tag{2.1.5}$$

由题设, 对 $k\in N_0$ 充分大, 有 $x_k\in N\big(x^*,\delta/2\big)$, $d_k\in N(d^*,\delta/2)$, 且存在 $M>0$ 使对任意 $k\in N_0$, $\|d_k\|<M$.

取 $\bar{\alpha}=\dfrac{\delta}{2M}$, 则存在 k_0, 当 $k\in N_0, k\geqslant k_0$ 时, $x_k+\bar{\alpha}d_k\in N(x^*,\delta)$. 再由 (2.1.5) 式知, 对任意 $k\in N_0, k\geqslant k_0$, 存在 $\theta_k\in(0,1)$ 使得

$$\begin{aligned} f_{k+1}-f_k &\leqslant f(x_k+\bar{\alpha}d_k)-f_k \\ &= \bar{\alpha}d_k^{\mathrm{T}}\nabla f(x_k+\theta_k\bar{\alpha}d_k) \\ &\leqslant \bar{\alpha}\left(-\frac{\varepsilon_0}{2}\right). \end{aligned}$$

由于数列 $\{f_k\}$ 单调不增, 从而将上式关于 $k\in N_0$ 求和得

$$\sum_{k=0}^{\infty}(f_{k+1}-f_k)\leqslant \sum_{k\in N_0}(f_{k+1}-f_k)\leqslant \sum_{k=0}^{\infty}\bar{\alpha}\left(-\frac{\varepsilon_0}{2}\right)=-\infty.$$

由于单调数列 $\{f_k\}$ 的子列 $\{f_k\}_{N_0}$ 存在极限 $f(x^*)$, 故 $\lim\limits_{k\to\infty} f_k=f(x^*)$. 从而由上式得

$$f(x^*)-f_0=\lim_{k\to\infty}f_k-f_0=\sum_{k=0}^{\infty}(f_{k+1}-f_k)\leqslant -\infty,$$

得到矛盾. 第一个结论得证.

对第二个结论, 若存在 $\varepsilon_0>0$ 使 $(d^*)^{\mathrm{T}}\nabla^2 f(x^*)d^*<-\varepsilon_0<0$, 则存在 x^* 点的邻域 $N(x^*,\delta)$ 及 d^* 的邻域 $N(d^*,\delta)$, 使对任意 $x\in N(x^*,\delta)$ 及 $d\in N(d^*,\delta)$ 有

$$d^{\mathrm{T}}\nabla^2 f(x)d<-\varepsilon_0/2.$$

同样取 $\bar{\alpha}=\dfrac{\delta}{2M}$, 则对充分大的 $k\in N_0$, $d_k\in N(d^*,\delta)$ 且 $x_k+\bar{\alpha}d_k\in N(x^*,\delta)$. 从而对充分大的 $k\in N_0$,

$$\begin{aligned} f_{k+1}-f_k &\leqslant f(x_k+\bar{\alpha}d_k)-f_k \\ &= \bar{\alpha}d_k^{\mathrm{T}}g_k+\frac{1}{2}\bar{\alpha}^2 d_k^{\mathrm{T}}G(\zeta_k)d_k \\ &\leqslant \frac{1}{2}\bar{\alpha}^2 d_k^{\mathrm{T}}G(\zeta_k)d_k \\ &\leqslant \frac{\bar{\alpha}^2}{2}\left(-\frac{\varepsilon_0}{2}\right), \end{aligned}$$

其中, $\zeta_k\in(x_k,x_k+\bar{\alpha}d_k)$. 类似的讨论可得矛盾, 证得第二个结论.　　证毕

为讨论精确线搜索方法的收敛速度, 先给出几个引理.

引理 2.1.1　设函数 $\varphi(\alpha)$ 在 $[0,b]$ 上二阶连续可微, $\varphi'(0)<0$, $\alpha^*\in(0,b)$ 为函数 $\varphi(\alpha)$ 在 $[0,b]$ 上的一个极小值点. 若存在 $M>0$ 使对任意 $\alpha\in[0,b]$, 有 $\varphi''(\alpha)\leqslant M$, 则 $\alpha^*\geqslant\dfrac{-\varphi'(0)}{M}$.

证明　由于 $\varphi(\alpha)$ 关于 α 在 $[0,b]$ 上二阶连续可微, $\alpha^*\in(0,b)$ 为 $\varphi(\alpha)$ 在 $[0,b]$ 上的极小值点, 所以存在 $\xi\in(0,\alpha^*)$ 使

$$\varphi'(0)=\varphi'(\alpha^*)+(0-\alpha^*)\varphi''(\xi)=-\alpha^*\varphi''(\xi),$$