科学专著丛书

非线性共轭梯度法

戴彧虹　袁亚湘　著

上 海 科 学 技 术 出 版 社

图书在版编目(CIP)数据

非线性共轭梯度法/戴彧虹,袁亚湘著.—上海:上海科学技术出版社,2000.10

(科学专著丛书)

ISBN 7-5323-5586-1

Ⅰ.非... Ⅱ.①戴...②袁... Ⅲ.非线性-共轭梯度法 Ⅳ.0241.6

中国版本图书馆 CIP 数据核字(2000)第 43706 号

上海科学技术出版社出版发行

(上海瑞金二路 450 号 邮政编码 200020)

上海新华印刷厂印刷 新华书店上海发行所经销

2000 年 10 月第 1 版 2000 年 10 月第 1 次印刷

开本 787×1092 小 1/16 印张 10.75 插页 4 字数 157 000

印数 1—1 200 定价:26.50 元

前　　言

共轭梯度法已有四十多年的历史. 它最早是由 Hestenes 和 Stiefel 于 1952 年在求解线性方程组时提出的,并由 Fletcher 和 Reeves 于 1964 年推广到非线性优化领域. 随后,Beale、Fletcher、Powell 等著名优化专家对非线性优化共轭梯度法进行了深入研究,取得了十分优秀的成果. 但几乎同时问世的拟牛顿方法由于其良好的计算表现以及丰富的收敛性分析很快受到了青睐,从而在很长一段时间里共轭梯度法被研究者所忽视. 近年来,随着计算机的飞速发展以及实际问题的需要,大规模优化问题越来越受到重视. 而共轭梯度法正是求解大规模问题的一种主要方法,这是因为共轭梯度法具有算法简单,易于编程,以及需要存储空间小等优点. 于是,共轭梯度法的理论研究又受到人们的关注.

本书系统地给出了无约束优化问题的共轭梯度法的收敛性理论,讨论了几个著名共轭梯度法在不同线搜索下的收敛性质. 在证明这些方法的收敛性结果的同时,还给出了一些例子说明我们所给出的结果在一定意义下是不能改进的. 书中的大部分内容都是作者近年来的研究结果.

作者非常感谢中国科学院应用数学所的韩继业教授和上海大学的王德人教授,他们对本书进行了认真细致的审阅,提出了许多宝贵的意见和建议;作者也非常感谢徐大川、张洪超和罗新龙等研究生,他们分别对本书的部分章节仔细地进行了阅读,提出了一些修改意见;此外,作者感谢本书第一作者戴彧虹的妻子王骏在本书的打字中提供的帮助.

由于水平有限,书中定有不少不妥之处,敬请广大读者批评指正.

袁亚湘、戴彧虹

1999 年 12 月于北京

第1章 绪 论

§1.1 引 言

共轭梯度法是最优化中最常用的方法之一.它具有算法简便、存储需求小等优点,十分适合于大规模优化问题.在石油勘探、大气模拟、航天航空等领域出现的特大规模的优化问题常常是利用共轭梯度法求解的.

考虑如下气象学中的一个例子[69],它是讲述如何用四维变分同化方法从大量已知数据中有效地决定大气的初始状态,从而进行天气预测的.假设在不同的时刻 T_i $(1\leqslant i\leqslant n)$,通过遥感卫星等设备观测到了某大气场的一组数据 Z_i,如湿度、风力、风向、涡度和温度等等.利用 Navier-Stokes 方程数值模型,并通过对时间和空间的离散,我们可得到许多形如

$$X_{i+1}=X_i+\Delta t G_i(X_i) \tag{1.1.1}$$

的方程,其中 X_i 是描述模型在 T_i 时刻状态的一个 N 维向量,而算子 G_i 依赖于所用的离散格式.(1.1.1)式表明了对 $i\geqslant 1$, X_i 由 T_0 时刻的初始状态向量 X_0 唯一确定.

由于各种原因,将预测结果和观测数据直接进行比较往往很困难,譬如观察到的温度可能包含辐射能,观察时采用的点不一定正好是数值模型中的网格点等等.因此为了比较预测结果和观测数据,我们还需要对变量 X_i 作变换:

$$\hat{X}_i=H_i(X_i), \tag{1.1.2}$$

其中 H_i 与观测向量 Z_i 有关.实际计算中常常使用谱方法,它使得变换 H_i 同时也是从谱空间到状态空间的变换.

通过变换 H_i,我们就可以定义预测结果 X_i 和观测结果 Z_i 之间的

距离：

$$J = \frac{1}{2}\sum_{i=1}^{n}(H_i(X_i) - Z_i)^T W(H_i(X_i) - Z_i), \qquad (1.1.3)$$

其中 W 是权矩阵，实际计算时常取为对角矩阵，以摄取主要的信息．于是我们的问题是：寻找初始状态向量 X_0，使得由(1.1.1)决定的状态向量 $X_1, \cdots, X_n$ 最小化距离函数(1.1.3)．

由于初始状态向量 X_0 的阶 N 一般为 10^6，时间方向的步长数 n 一般为 10^2，因此我们所要求解的优化问题其规模相当大．如果一天中要多次预测天气，或进行中长期天气预测，如何快速求解问题(1.1.3)就显得尤为重要．

在所有需要计算导数的优化方法中，最速下降法是最简单的，但它速度太慢．拟牛顿方法收敛速度很快，被广泛认为是非线性规划的最有效的方法．但拟牛顿法需要存储矩阵以及通过求解线性方程组来计算搜索方向，这对于求解诸如上述问题等一些大规模问题几乎是不大可能办到的．共轭梯度法在算法的简便性、所需存储量等方面均与最速下降法差别不大，而收敛速度比最速下降法要快．正因为这些原因，许多实际部门的工程师十分喜欢用共轭梯度法．

共轭梯度法最早是由计算数学家 Hestenes[46]和几何学家 Stiefel[85]在 20 世纪 50 年代初为求解线性方程组

$$Ax = b, \quad x \in R^n \qquad (1.1.4)$$

而独立提出的．他们合作的著名文章[48]被认为是共轭梯度法的奠基性工作．该文详细讨论了求解线性方程组的共轭梯度法的性质以及它和其他方法的关系．在 A 为对称正定时，线性方程组(1.1.4)等价于最优化问题

$$\min_{x \in R^n} \frac{1}{2}x^T Ax - b^T x. \qquad (1.1.5)$$

基于此，Hestenes 和 Stiefel 的方法也可视为求二次函数极小值的共轭梯度法．1964 年 Fletcher 和 Reeves[40]将此方法推广到非线性优化，得到了求一般函数极小值的共轭梯度法．

非线性优化的共轭梯度法的收敛性分析的早期工作主要由 Fletcher、Powell、Beale 等学者给出的．近年来，Nocedal、Gilbert、Nazareth、 Al-Baali、Storey 等学者在收敛性方面得到不少新结果，使得共轭梯度法这一冷清了多年的研究方向又热了起来. 可喜的是，我国学者也在共轭梯度法的理论研究中取得一定的成绩. 例如，中国科学院应用数学研究所韩继业教授课题组的工作就十分突出.

本书全面、系统地介绍共轭梯度法，向读者介绍共轭梯度法的基本思想和主要方法；给出主要的收敛性结果以及理论分析的基本技巧. 书中的大部分内容都是作者近年来的研究结果.

§1.2 线性共轭梯度法

共轭梯度法最本质的是共轭性质. 共轭性可追溯至 1936 年，Birkhoff 等学者在研究变分中自然等周条件时引入了共轭的概念. 在矩阵计算中，将 Gram-Schmidt 正交化方法推广为共轭化方法，就是寻找椭球的两两共轭的轴. 共轭是正交的一个直接推广.

定义 1.2.1 设 $A\in R^{n\times n}$对称正定，$d_1, d_2, \cdots, d_m$ 是 R^n 中的一组非零向量，如果

$$d_i^T A d_j = 0 \ (i \neq j), \tag{1.2.1}$$

则称 $d_1, d_2, \cdots, d_m$ 是相互 A-共轭.

显然可见，如果 $d_1, \cdots, d_m$ 相互 A-共轭，则它们是线性无关的. 设 I 为单位矩阵，则知 I-共轭就是正交.

共轭方向的优点之一是，它能将一个 n 元线性方程组化为 n 个单元线性方程.

定理 1.2.2 设 $A\in R^{n\times n}$对称正定，$d_1, \cdots, d_n$ 是 n 个相互 A-共轭的向量，对任给的 $b\in R^n$，

$$x^* = \sum_{i=1}^{n} \frac{b^T d_i}{d_i^T A d_i} d_i \tag{1.2.2}$$

是 $Ax = b$ 的解.

证明 从(1.2.2)和 A-共轭的定义知

$$d_j^T[Ax^* - b] = \sum_{i=1}^{n} \frac{b^T d_i d_j^T A d_i}{d_i^T A d_i} - d_j^T b = b^T d_j - d_j^T b = 0 \tag{1.2.3}$$

对 $j = 1, 2, \cdots, n$ 都成立. 因为 $d_1, \cdots, d_n$ 线性无关,故知 $Ax^* = b$. ▌

由(1.1.4)与(1.1.5)的等价性,有如下推论:

推论 1.2.3 设 $A \in R^{n\times n}$ 对称正定,$b \in R^n$, $d_1, \cdots, d_n$ 是 n 个相互 A-共轭向量,则 n-维空间中二次函数极小值问题(1.1.5)等价于 n 个 1-维二次函数极小值问题

$$\min_{\alpha\in R^1} \frac{1}{2}\alpha^2 d_i^T A d_i - \alpha d_i^T b \ (i = 1, 2, \cdots, n). \tag{1.2.4}$$

(1.2.4)的解为 $\alpha_i = b^T d_i / d_i^T A d_i$.

共轭向量可以类似于 Gram-Schmidt 正交化产生. 给出一组线性无关的向量 $u_1, u_2, \cdots, u_n$ 和对称正定阵 A,如下算法可产生一组两两 A-共轭的向量.

算法 1.2.4(共轭 Gram-Schmidt 过程)

步 1 $p_1 = u_1$, $k := 1$, 如果 $n = 1$, 则停.

步 2 对 $j = 1, \cdots, k$, 计算

$$\delta_j = p_j^T A p_j, \tag{1.2.5}$$

$$b_{k+1,j} = p_j^T A u_{k+1} / \delta_j. \tag{1.2.6}$$

步 3

$$p_{k+1} = u_{k+1} - \sum_{j=1}^{k} b_{k+1,j} p_j \tag{1.2.7}$$

如果 $k + 1 = n$, 则停;

$k := k + 1$, 转步 2.

共轭 Gram-Schmidt 过程的一个基本性质是前 k 个共轭方向用前 k 个线性无关向量的线性组合所得到.

引理 1.2.5 设 A 是对称正定阵,u_i $(i = 1, \cdots, n)$ 是 n 个线性无关的向量,p_i $(i = 1, \cdots, n)$ 由算法 1.2.4 产生,则对任何的 $1 \leqslant k \leqslant n$, 有

$$\mathrm{Span}\{p_1, \cdots, p_k\} = \mathrm{Span}\{u_1, \cdots, u_k\}, \tag{1.2.8}$$

这里符号 Span 表示由向量张成的子空间.

在求解线性方程组(1.1.4)时,如果特殊地选取 u_i 为方程组在 x_i 点的余量

$$r_i = Ax_i - b, \tag{1.2.9}$$

则只需依次利用 p_i 和 u_i,就可构造出 n 个两两共轭的向量,从而最多在 n 步便可求出(1.1.4)的解. 这就导出了大家熟知的线性共轭梯度法.

算法 1.2.6(线性共轭梯度法)

步 1　给出初值 $x_1 \in R^n$; $r_1 = Ax_1 - b$;

$$p_1 = -r_1,\ k := 1.$$

步 2　如果 $r_k = 0$, 则停;

计算步长因子

$$\alpha_k = -r_k^T r_k / p_k^T A p_k; \tag{1.2.10}$$

$$x_{k+1} = x_k + \alpha_k p_k,\ r_{k+1} = Ax_{k+1} - b.$$

步 3　计算参数

$$\beta_{k+1} = \| r_{k+1} \|^2 / \| r_k \|^2; \tag{1.2.11}$$

$$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k;$$

$k := k + 1$; 转步 2.

(1.2.11)式以及下文中,模 $\| \cdot \|$ 指 Euclidean 模. 下述定理概括了线性共轭梯度法的一些基本性质.

定理 1.2.7　算法 1.2.6 经过不超过 n 次迭代就会终止,即存在 $m \leqslant n$, 使得

$$r_{m+1} = 0, \tag{1.2.12}$$

而且对一切 $1 \leqslant k \leqslant m$ 和 $1 \leqslant j \leqslant k-1$, 都有

$$r_k^T p_k = -\| r_k \|^2, \tag{1.2.13}$$

$$r_k^T A r_j = 0, \tag{1.2.14}$$

$$r_k^T p_j = 0, \tag{1.2.15}$$

$$r_k^T r_j = 0. \tag{1.2.16}$$

证明 步长因子 α_k 的选取使得

$$r_{k+1}^T p_k = 0. \tag{1.2.17}$$

利用 $p_1 = -r_1$，(1.2.17)及 p_{k+1}的定义即知(1.2.13)对 $k \geqslant 1$ 成立. 为证(1.2.14)～(1.2.16)，我们用归纳法：显然，(1.2.14)～(1.2.16)对 $k=2$ 成立. 设(1.2.14)～(1.2.16)对某 k 和所有的 $1 \leqslant j \leqslant k-1$ 成立. 注意到

$$p_k \in \mathrm{Span}\{r_1, \cdots, r_k\}, \tag{1.2.18}$$

对 $k+1$ 和任意的 $1 \leqslant j \leqslant k-1$，有

$$r_{k+1}^T r_j = r_k^T r_j + (r_{k+1} - r_k)^T r_j = r_k^T r_j + \alpha_k p_k^T A r_j = 0. \tag{1.2.19}$$

此外，由 α_k 的定义(1.2.10)及(1.2.13)有

$$\begin{aligned} r_{k+1}^T r_k &= \| r_k \|^2 + (r_{k+1} - r_k)^T r_k = \| r_k \|^2 + \alpha_k p_k^T A r_k \\ &= \| r_k \|^2 + \alpha_k r_k^T A r_k = 0. \end{aligned} \tag{1.2.20}$$

故(1.2.16)对 $k+1$ 和所有的 $1 \leqslant j \leqslant k$ 成立. 再次利用(1.2.18)知

$$r_{k+1}^T p_j = 0,\ 1 \leqslant j \leqslant k-1. \tag{1.2.21}$$

上式和(1.2.17)表明(1.2.15)对 $k+1$ 和所有的 $1 \leqslant j \leqslant k$ 也成立. 由 d_{k+1}的定义及(1.2.16)知，对 $1 \leqslant j \leqslant k-1$，有

$$\begin{aligned} p_{k+1}^T A p_j &= (-r_{k+1} + \beta_k p_k)^T A p_j = - r_{k+1}^T A p_j \\ &= - \alpha_j^{-1} r_{k+1}^T r_{j+1} - r_j = 0. \end{aligned} \tag{1.2.22}$$

此外，利用(1.2.8)及直接计算知 $p_{k+1}^T A p_k = 0$，故(1.2.14)对 $k+1$ 和所有的 $1 \leqslant j \leqslant k$ 亦成立. 因此，(1.2.14)～(1.2.16)对所有 $1 \leqslant k \leqslant m$ 和 $1 \leqslant j \leqslant k-1$ 成立.

由(1.2.14)知，$p_1, \cdots, p_k$ 相互 A-共轭，从而由(1.2.15)知必存在 $m \leqslant n$，使得(1.2.12)成立. ▌

线性共轭梯度法除对病态问题尚有待深入研究外，其理论已相当完

善.目前,结合预条件技巧的线性共轭梯度法被广泛用于线性方程组的求解.关于线性共轭梯度法的较全面的介绍,读者可参阅文献[4]和[41].

§1.3 非线性共轭梯度法

对于无约束优化问题

$$\min_{x\in R^n} f(x), \tag{1.3.1}$$

给出一初始值 x_1,算法迭代产生 x_2, x_3, ….希望某一 x_k 是(1.3.1)的解或点列收敛于解.在第 k 次迭代,当前迭代点为 x_k,线搜索型的方法将产生一搜索方向 $d_k\in R^n$.然后下一个迭代点

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.3.2}$$

其中 $\alpha_k > 0$ 是步长因子,它满足某些线搜索条件.信赖域型的方法在当前迭代点周围的一邻域中求得一试探点 $x_k + s_k$,然后通过某种判断,在 x_k 与 $x_k + s_k$ 之间挑选其一为下一个迭代点 x_{k+1}.本书仅考虑线搜索型的共轭梯度法.

一个线搜索型的方法的每步迭代主要由两部分组成:第一是搜索方向 d_k 的计算;第二是步长因子 α_k 的计算.

搜索方向 d_k 通常要求是下降方向,即

$$d_k^T \nabla f(x_k) < 0. \tag{1.3.3}$$

上式保证一定存在 $\alpha > 0$,使得

$$f(x_k + \alpha d_k) < f(x_k). \tag{1.3.4}$$

也就是说,沿着 d_k 方向搜索,一定可找到比 x_k 更好的点.记 $g(x) = \nabla f(x)$ 以及 $g_k = g(x_k)$.最简单的下降方向是负梯度方向:

$$d_k = - g_k. \tag{1.3.5}$$

在 $f(x)$连续可微时,只要 $g_k \neq 0$,则对任何向量 $d\in R^n$ 且 $\| d \|_2 = 1$,有

$$\lim_{t\to 0_+}\frac{f(x_k)-f(x_k+td)}{f(x_k)-f(x_k-tg_k/\|g_k\|_2)}=-\frac{d^Tg_k}{\|g_k\|_2}\leqslant 1. \quad (1.3.6)$$

正因为如此，我们称负梯度方向为最速下降方向，利用最速下降方向作为搜索方向的方法称为最速下降法. 牛顿法的搜索方向为

$$d_k=-(\nabla^2 f(x_k))^{-1}g_k. \quad (1.3.7)$$

它需要计算 $n\times n$ 海色阵$\nabla^2 f(x_k)$. 如果$\nabla^2 f(x_k)$非正定，则牛顿法的搜索方向(1.3.7)可能不是下降方向. 拟牛顿法用 H_k 近似$(\nabla^2 f(x_k))^{-1}$，计算搜索方向

$$d_k=-H_kg_k, \quad (1.3.8)$$

并在每次迭代时对 H_k 进行修正. 无论是牛顿法还是拟牛顿法都需要存储 $n\times n$ 的矩阵，因此很难应用于大规模问题.

求解问题(1.3.1)的共轭梯度法是从求解线性方程组(1.1.4)的线性共轭梯度法推广而来的，其搜索方向是负梯度方向与上一次迭代的搜索方向的线性组合，它表示为

$$d_k=-g_k+\beta_kd_{k-1}, \quad (1.3.9)$$

由于(1.2.9)式给出的余量 r_i 恰好是问题(1.1.5)中目标函数在 x_i 的梯度，因此对一般的无约束优化问题，可用目标函数的梯度 g_i 代替 r_i. 这时，由算法 1.2.6 知，关于参数 β_{k+1}的一个计算公式为

$$\beta_{k+1}=\|g_{k+1}\|^2/\|g_k\|^2. \quad (1.3.10)$$

上述公式即著名的 Fletcher-Reeves 公式. 根据定理 1.2.7，当 $f(x)$为二次凸函数时，参数 β_{k+1}还有许多等价的其他计算公式(如见第 2 章～第 5 章). 这些计算公式的性质当 $f(x)$为一般非线性函数时可能大相径庭.

步长因子 α_k 的计算对一个线搜索型方法也至关重要. 简单说来，它是从 x_k 沿 d_k 方向寻找一个“好”的点作为下一个迭代点. 显然，最好的点是在这个方向函数值达到最小的点，即要求

$$f(x_k+\alpha_kd_k)=\min_{\alpha>0}f(x_k+\alpha d_k). \quad (1.3.11)$$

这时，我们称线搜索为**精确线搜索**. 对于线性共轭梯度法，由(1.2.10)给

出的 α_k 满足(1.2.17),因此其步长因子可看成由精确线搜索得到的.

精确线搜索需要求一个单变量函数的极小值,计算量较大.故在实际计算中常常进行非精确线搜索.典型的非精确线搜索是 Wolfe 非精确线搜索,简称 Wolfe 线搜索.它要求 α_k 满足 Wolfe 条件:

$$f(x_k + \alpha_k d_k) \leqslant f(x_k) + \delta \alpha_k d_k^T g_k, \tag{1.3.12}$$

$$d_k^T g(x_k + \alpha_k d_k) \geqslant \sigma d_k^T g_k, \tag{1.3.13}$$

其中 δ 和 σ 是满足 $0 < \delta < \sigma < 1$ 的常数.第一个不等式要求函数下降至少要与切线的下降成比例.第二个不等式防止步长过小,因而保证目标函数的足够下降.只要目标函数 $f(x)$ 在 $\{x_k + \alpha d_k,\ \alpha > 0\}$ 上可微且有下界,即知(1.3.12)、(1.3.13)存在解 α_k.

强 Wolfe 非精确线搜索是要求 α_k 满足(1.3.12)和

$$|d_k^T g(x_k + \alpha_k d_k)| \leqslant \sigma |d_k^T g_k|. \tag{1.3.14}$$

当 $\sigma = 0$ 时,必有 $g_{k+1}^T d_k = 0$. 因此强 Wolfe 非精确线搜索可看成是精确线搜索的一种推广.在非线性共轭梯度法的许多理论分析和数值实现中,常常使用强 Wolfe 非精确线搜索,本文简称强 Wolfe 非精确线搜索为强 Wolfe 线搜索,而条件(1.3.12)和(1.3.14)称为强 Wolfe 条件.

不等式(1.3.14)实质上就是

$$\sigma d_k^T g_k \leqslant d_k^T g(x_k + \alpha_k d_k) \leqslant - \sigma d_k^T g_k. \tag{1.3.15}$$

推广的 Wolfe 线搜索要求 α_k 满足(1.3.12)和

$$\sigma_1 d_k^T g_k \leqslant d_k^T g(x_k + \alpha_k d_k) \leqslant - \sigma_2 d_k^T g_k, \tag{1.3.16}$$

其中 $\sigma_1 \in (\delta,\ 1)$, $\sigma_2 \geqslant 0$. 当 $\sigma_1 = \sigma_2$ 时,推广的 Wolfe 线搜索就是强 Wolfe 线搜索;当 $\sigma_2 = +\infty$ 时,推广的 Wolfe 线搜索就是 Wolfe 线搜索.

线搜索型的共轭梯度算法可描述如下:

算法 1.3.1(非线性共轭梯度法)

步 1 给出初值 $x_1 \in R^n$, $\varepsilon \geqslant 0$;

$d_1 = - g_1$, $k := 1$.

步 2 如果 $\| g_k \| \leqslant \varepsilon$, 则停;

利用某种非精确线搜索求 α_k;

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

步 3 利用某种公式计算参数 β_{k+1}；

$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k;$$

$$k := k+1;\ \text{转步 2.}$$

想了解共轭梯度法计算表现的读者可参阅文献[47]中的算例. 本书主要考虑算法的理论性质，故不介绍各个方法的具体数值算例. 关于 Wolfe 线搜索实现的基本思想，读者也可参阅文献[102]. 读者如需要详细了解共轭梯度法的算法实现或需要共轭梯度法的计算程序，可直接与作者联系.

§1.4 Zoutendijk 条件

优化计算方法产生点列 x_k $(k=1, 2, \cdots)$，其目的是要得到一迭代点为问题(1.3.1)的解，或充分靠近解的点. 问题(1.3.1)的解 x^* 的定义为

$$f(x) \geqslant f(x^*),\ \forall x \in R^n. \tag{1.4.1}$$

由于条件(1.4.1)在计算机上不能实现，在实际计算中，常常用某些必要条件来替代. (1.4.1)的一阶必要条件是

$$\nabla f(x^*) = g(x^*) = 0. \tag{1.4.2}$$

我们称满足(1.4.2)的点为一阶稳定点，也简称为稳定点. 通常，给出一允许误差 $\varepsilon > 0$，只要

$$\| g_k \| \leqslant \varepsilon, \tag{1.4.3}$$

则可接受 x_k 为一近似解. 要证明算法对任何 $\varepsilon > 0$ 都能找到满意的近似解，应当有

$$\liminf_{k\to\infty} \| g_k \| = 0. \tag{1.4.4}$$

上式正是通常我们分析算法收敛性所需要证明的. 显然，它比强收敛

$$\lim_{k\to\infty} \| g_k \| = 0 \tag{1.4.5}$$

和点列收敛

$$\lim_{k\to\infty}\| x_k - x^* \| = 0 \tag{1.4.6}$$

都弱，但它足以保证对任何 $\varepsilon > 0$，都有 k 使得 x_k 为满足(1.4.3)的近似解.

以下引理在分析一般的下降算法的收敛性时经常要用到. 它由 Wolfe[89]、[90]和 Zoutendijk[94]所获得，读者可以证明这一结果对精确线搜索(1.3.11)也对.

引理 1.4.1　设目标函数 $f(x)$下方有界，导数$\nabla f(x)$满足 Lipschitz 条件

$$\| \nabla f(x) - \nabla f(\tilde{x}) \| \leqslant L \| x - \tilde{x} \|, \ \forall x, \tilde{x} \in R^n. \tag{1.4.7}$$

考虑一般方法 $x_{k+1} = x_k + \alpha_k d_k$，其中 d_k 满足 $d_k^T g_k < 0$，步长因子 α_k 满足(1.3.12)和(1.3.13)，则有

$$\sum_{k\geqslant 1} \frac{(g_k^T d_k)^2}{\| d_k \|^2} < +\infty. \tag{1.4.8}$$

关系式(1.4.8)常被称为 Zoutendijk 条件.

证明　由(1.3.13)知

$$[g_{k+1} - g_k]^T d_k \geqslant (\sigma - 1) g_k^T d_k. \tag{1.4.9}$$

另一方面，由 Lipschitz 条件(1.4.7)有

$$[g_{k+1} - g_k]^T d_k \leqslant \alpha_k L \| d_k \|^2. \tag{1.4.10}$$

利用以上两式得

$$\alpha_k \geqslant \frac{\sigma - 1}{L} \cdot \frac{g_k^T d_k}{\| d_k \|^2}. \tag{1.4.11}$$

(1.3.12)和(1.4.11)表明

$$f_k - f_{k+1} \geqslant c \frac{(g_k^T d_k)^2}{\| d_k \|^2}, \tag{1.4.12}$$

其中 $c = \delta(1-\sigma)/L$. 对上式从 $k = 1, 2, \cdots$ 求和，并注意 $f(x)$下方有界，即知(1.4.8)成立.　▎

本书将讨论各种共轭梯度法在不同线搜索下的收敛性质. 几乎所有的证明都是用反证法. 其基本思想是：假定(1.4.4)不成立，即存在常数

$\gamma > 0$，使得

$$\| g_k \| \geqslant \gamma,\ \forall k \geqslant 1, \tag{1.4.13}$$

然后通过证明

$$\sum_{k \geqslant 1} \frac{(g_k^T d_k)^2}{\| d_k \|^2} = +\infty, \tag{1.4.14}$$

而导出了与 Zoutendijk 条件相矛盾. 在精确线搜索下，

$$d_k^T g_k = -\| g_k \|^2, \tag{1.4.15}$$

因此如果能够证明 $\| d_k \|^2$ 最多线性增长，即

$$\| d_k \|^2 \leqslant c_1 + c_2 k, \tag{1.4.16}$$

则从(1.4.13)、(1.4.15)和(1.4.16)而知，(1.4.14)成立，从而得出矛盾. 在非精确线搜索时，许多收敛性证明都致力于估计 $(g_k^T d_k)^2$ 的下界以及 $\| d_k \|^2$的上界，而前者往往与证明搜索方向 d_k 的下降性相关联.

对优化算法证明(1.4.4)是给出算法的全局收敛性质，也称总体收敛性质，即分析算法从任何初始点开始都可以找到满足任意精度的近似解. 对于一个算法，我们不仅关心它所产生的点列是否收敛，也关心它收敛的快慢. 分析算法在解附近的收敛速度称为局部收敛性分析.

假定算法产生的点列 x_k 收敛于解 x^*. 局部收敛性分析就是讨论 $\| x_k - x^* \|$趋于零的速度. 对于二次凸函数，采取精确线搜索的共轭梯度法具有有限终止性，即在第一个搜索方向是最速下降方向时，可证明不超过 n 次迭代就可找到最优点. 但如果第一个搜索方向不是最速下降方向，共轭梯度法一般不会有限终止，且收敛速度仅是线性的.

如果共轭梯度算法在每 n 步沿负梯度方向作一重新开始，则从 Zoutendijk 条件可立即推出收敛关系式(1.4.4). 进一步，如果线搜索精确或渐进精确，则重新开始的共轭梯度算法的收敛速度可从原来的线性提高到 n 步超线性收敛[13]、[59]：

$$\lim_{k \to \infty} \frac{\| x_{k+n} - x^* \|}{\| x_k - x^* \|} = 0. \tag{1.4.17}$$

可见重新开始的共轭梯度算法无论是在全局收敛性还是在收敛速度方

面，都比原来的共轭梯度算法具有更好的性质. 然而，共轭梯度算法一般用于大规模问题的求解，而对于大规模问题，共轭梯度算法往往在远少于 n 步时即终止计算. 基于此，本书将着重研究不使用重新开始技术的共轭梯度法的全局收敛性.

第 2 章　Fletcher-Reeves 方法

§2.1 引　言

Fletcher-Reeves 共轭梯度法是最早的非线性共轭梯度法，它是由 Fletcher 和 Reeves[40]在 1964 年将求解线性方程组的共轭梯度法推广用于求解优化问题(1.1.4)而得来的. 下面简称 Fletcher-Reeves 共轭梯度法为 FR 方法. FR 方法形式如下：

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.1.1}$$

$$d_k = \begin{cases} -g_k, & 若\ k = 1; \\ -g_k + \beta_k d_{k-1}, & 若\ k \geqslant 2, \end{cases} \tag{2.1.2}$$

其中参数 β_k 按以下公式计算：

$$\beta_k^{FR} = \frac{\| g_k \|^2}{\| g_{k-1} \|^2}. \tag{2.1.3}$$

早期对 FR 方法的分析是基于精确线搜索. Powell[76]在精确线搜索下分析了 FR 方法可能连续产生小步长的性质，即如果 FR 方法在某一步产生一个很小的步长，则相继的许多步长也可能非常小. Powell[76]也给出了 FR 方法最简单的全局有效性分析. 他发现，当采取精确线搜索的 FR 方法进入到一个 $f(x)$为二维二次函数

$$f(x) = \frac{1}{2} x^T x, \quad x \in R^2 \tag{2.1.4}$$

的区域时，搜索方向 d_k 和负梯度方向 $-g_k$ 的夹角始终保持不变. 由于这个角度可任意接近 $\pi/2$，因此 FR 方法可能收敛非常慢. 这些分析在很大