数值优化

引论

董云达 主编

黄河水利出版社

数值优化引论

董云达　主编

黄河水利出版社

内 容 提 要

本书较为系统和全面地阐述了数值优化的基本理论、方法和应用。它主要包括无约束优化方法、约束优化的基本理论和方法、线性规划及其对偶理论和方法。可以作为计算数学专业和运筹学专业的高年级本科生、研究生的教科书，也可作为其他相关专业的科研工作者的参考书。

图书在版编目(CIP)数据

数值优化引论／董云达主编. —郑州：黄河水利出版社，2007.9
ISBN 978-7-80734-258-8

Ⅰ.数… Ⅱ.董… Ⅲ.最优化算法 Ⅳ.0242.23

中国版本图书馆 CIP 数据核字(2007)第 133471 号

出 版 社：黄河水利出版社
地址：河南省郑州市金水路 11 号　　邮政编码：450003
发行单位：黄河水利出版社
发行部电话：0371-66026940　　传真：0371-66022620
E-mail：hhslcbs@126.com
承印单位：黄河水利委员会印刷厂
开本：787 mm × 1 092 mm　1 / 16
印张：12
字数：274 千字　　印数：1—1 000
版次：2007 年 9 月第 1 版　　印次：2007 年 9 月第 1 次印刷

书号：ISBN 978-7-80734-258-8 / O · 22　　定价：22.00 元

前 言

本书较为系统和全面地阐述了数值优化的基本理论、方法和应用。它主要包括无约束优化方法、约束优化的基本理论和方法、线性规划及其对偶理论和方法。

在编写本书的过程中，一些优秀的专著、教材使我们受益不浅。尤其是袁亚湘和孙文瑜的《最优化理论和方法》(1997)，Nocedal 和 Wright 的 Numerical Optimization(1999)以及黄红选和韩继业的《数学规划》(2006)。

本书的特点是内容精炼，深入浅出。在选材上，它涉及了在优化中那些被普遍看好的数值方法以及密切相关的理论基础；在叙述上，它往往从最基本的概念出发，然后用通俗易懂的语言逐步展开，对当前的这些可能最佳的结果予以论述。

本书是由编者集中讨论、分头编写的。董云达编写第 1~3 章，第 9~11 章；左明艳编写第 4 章；黄玲玲编写第 5 章；李成芳编写第 6 章；刘甲玉编写第 7 章；李玉华编写第 8 章；任勤编写第 12~13 章；仝伟编写附录。最后由董云达对全书进行了统一润色和定稿。

由于编者水平有限，加之时间仓促，本书可能存在不妥甚至谬误之处，敬请专家批评指正。

编 者

2007 年 4 月

目 录

第1章 引 言

在人类社会中，优化现象层出不穷。比如，投资方常常设计资金运作方案，以便在可以接受的风险范围内获取最大的投资回报；在生产、销售环节中，管理方往往整合有限的人力、物力和财力资源，以便换来最大的经济效益。

在自然界，优化现象也无处不在。在一个物理系统中，各部分不断演化，使整个系统趋于最小能量状态；在一个化学系统中，分子相互作用，直至它们的电子的总体势能最小。

对于这些现象，如何用数学的方式加以描绘与研究，则是摆在人们面前的课题。

一个可行的解决办法是：适当引入自变量，将它们纳入某个数学模型中，该数学模型应包括两部分：一是用来描述自变量取值范围的，即自变量应满足什么样的约束条件；二是用来描述目标函数的。

如果目标函数以及对应于约束条件的约束函数都是线性的，则称该模型为线性规划；否则，称该模型为非线性规划。

一般来讲，自变量的引入是建立数学模型的关键步骤。自变量的个数不宜太多，否则，它会增加问题的维数；其个数也不宜太少，如果太少的话，则表明一些不宜忽略的因素被漏掉了，所得到的数学模型就会显得过于粗糙。

建立合适的数学模型后，根据目标函数与约束条件的特点，人们就可以用某种优化方法在计算机上给出一个近似的数值解。

如果是线性规划，人们用单纯形法或原始-对偶内点法来解。对于非线性规划，虽然存在不少的通用方法，但它们实际的计算性能往往并不理想，在合理的时间内计算不出结果来，这个时候，需要人们根据经验来判断应选用什么样的专门方法比较好。

一个较好的方法应至少具备下列的标准：①它首先应该是一个收敛的算法，即该方法从某个合适的初始点启动，终止于问题的一个近似解；②它应该具有较快的收敛速度。这要求从迭代开始直到一个满足精度要求的近似解被找到的过程中，需要的迭代次数较少且每次迭代的运算量不大；③它应该具备较好的鲁棒性。这里所说的鲁棒性，是指它解决同类问题中的大多数问题的能力。

值得注意的是，这些标准在实际上可能有所冲突。比如，当一个算法的迭代次数较少时，它每次迭代的平均运算量可能要大些，除了目标函数的估值外，甚至还要求一阶、二阶导数的信息。再如，一个具有较高鲁棒性的算法往往收敛得很慢，如何折中这些冲突，则是数值优化工作者必须面对的一个核心问题。

1.1 优化问题的一般模型

在本书中，优化问题常被写成以下形式：

$$\min \quad f(x) \tag{1.1}$$

$$\text{s.t.} \quad g_i(x)\geqslant 0, \quad i=1,2,\cdots,l \tag{1.2}$$

$$h_j(x)=0, \quad j=1,2,\cdots,m \tag{1.3}$$

式中：自变量 $x\in R^n$ 是 n 维向量；f：$R^n\to\mathbf{R}$ 是目标函数；式(1.2)中的函数 g_i：$R^n\to\mathbf{R}$ 称为不等式约束函数；式(1.3)中的函数 h_j：$R^n\to\mathbf{R}$ 称为等式约束函数。顺便说一下，这里的 min 以及 s.t.分别为 minimize (极小化) 和 subject to(受约束于)的英文缩写。

同时满足式(1.2)与式(1.3)的自变量集合称为上述优化问题的可行域，可行域如果非空的话，其中的任一点称为可行点；否则，称为不可行点。

1.2 优化问题的分类

为了以后叙述的方便，我们现根据目标函数、约束函数的特点将优化问题予以分类。

1.2.1 连续 / 离散优化

若优化问题式(1.1) ~ 式(1.3)的可行域为 R^n 中有限个不可数的连通集的并集，则称为连续优化。对比之下，离散优化要求自变量的某个分量的可能取值仅为有限个。

本书不打算涉及离散优化，然而，需要指出的是，连续优化方法在离散优化中起着至关重要的作用。比如，离散优化中的分枝定界方法常常将原问题分解成一系列子问题，而其每一个子问题均为连续优化问题。

1.2.2 无约束优化和约束优化

若优化问题式(1.1) ~ 式(1.3)的可行域为 R^n，则称为无约束优化问题；若其可行域不是空集也不是 R^n，则称为约束优化问题。有趣的是，一个约束优化问题常常可以通过一些技巧将它转化为无约束优化问题。因此，无约束优化算法始终在优化领域中发挥着基本的作用。

1.2.3 全局优化和局部优化

对于优化问题式(1.1) ~ 式(1.3)，我们常常可以找到一个可行点 x^*，使得它的某个充分小的邻域内的所有可行点 x 处的目标函数值都不比 $f(x^*)$ 小，则称 x^* 为优化问题式(1.1) ~ 式(1.3)的一个局部极小点。如果对于任意的可行点 x，总有

$$f(x^*)\leqslant f(x)$$

则称 x^* 为优化问题式(1.1) ~ 式(1.3)的一个全局极小点。

一般来说，对于一个优化问题，找一个全局极小点要比找一个局部极小点困难得多。但一个例外是：对于凸规划，即目标函数、约束函数均为凸函数的情形，局部极小点必为全局极小点。比如，线性规划可以看作凸规划的一个特殊情形，从而它的任一局部极小点必为全局极小点。

另外，也可以将优化问题式(1.1) ~ 式(1.3)分为光滑优化和非光滑优化、随机性优化和确定性优化等。

第2章　基本知识

2.1　关于极小点的一些定理

考察下面的无约束优化问题

$$\min_{x\in R^n} f(x) \tag{2.1}$$

式中：f：$R^n \to \mathbf{R}$是目标函数。

对于这样的问题，我们的目标是找出它的一个解。为了以后叙述的方便，现给出优化问题的解的定义。

设$x^* \in R^n$，对于任意的$x \in R^n$，有

$$f(x^*) \leqslant f(x)$$

我们则称x^*是优化问题式(2.1)的一个全局极小点。

与之相对应，我们给出局部极小点的定义。

设$x^* \in R^n$，$\varepsilon > 0$，$N_\varepsilon(x^*) = \{x : \| x - x^* \| < \varepsilon\}$为$x^*$的某个邻域，若对于任意的$x \in N_\varepsilon(x^*)$，总有

$$f(x^*) \leqslant f(x)$$

我们则称x^*是优化问题式(2.1)的一个局部极小点。

设$x^* \in R^n$，$\varepsilon > 0$，$N_\varepsilon(x^*)$为x的某个邻域，若对于任意的$x \in N_\varepsilon(x^*)$，$x \neq x^*$，总有

$$f(x^*) < f(x)$$

我们则称x^*是优化问题式(2.1)的一个严格局部极小点。

注意：在上述的定义中，$N_\varepsilon(x^*)$是一个开集。

根据上述定义，当目标函数$f(x)$=1时，每一个x都是一个局部极小点而不是严格局部极小点；当目标函数$f(x) = x^2$时，$x = 0$既是一个局部极小点，也是严格局部极小点。

下面，我们讨论$x^* \in R^n$为局部极小点时的一些性质。为此，考察下面的Taylor定理。

定理 2.1　设f：$R^n \to \mathbf{R}$是连续可微的，$d \in R^n$。那么，必然存在$t \in (0,1)$，使得

$$f(x+d) = f(x) + \nabla f(x+td)^{\mathrm{T}} d$$

而且，若f是二次连续可微的，必然存在$t \in (0,1)$，使得

$$\nabla f(x+d) = \nabla f(x) + \int_0^1 \nabla^2 f(x+td) d \ \mathrm{d}t$$

以及

$$f(x+d) = f(x) + \nabla f(x)^{\mathrm{T}} d + \frac{1}{2} d^{\mathrm{T}} \nabla^2 f(x+td) d$$

其中，$\nabla f(x)$, $\nabla^2 f(x)$ 分别称为 $f(x)$ 在点 x 处的梯度、Hesse 阵，且

$$\nabla f(x)=\begin{pmatrix}\dfrac{\partial f}{\partial x_1}\\ \vdots \\ \dfrac{\partial f}{\partial x_n}\end{pmatrix}, \qquad \nabla^2 f(x)=\begin{bmatrix}\dfrac{\partial^2 f}{\partial x_1 \partial x_1}\cdots\dfrac{\partial^2 f}{\partial x_1 \partial x_n}\\ \vdots \qquad\quad \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}\cdots\dfrac{\partial^2 f}{\partial x_n \partial x_n}\end{bmatrix}$$

定理 2.2 设 x^* 是 $f(x)$ 的一个局部极小点，且 $f(x)$ 在 x^* 的某个开邻域内连续，则

$$\nabla f(x^*)=0$$

证明：设 $d\in R^n$，由 Taylor 定理，对于任意的 $t>0$，总有

$$f(x^*+td)=f(x^*)+t\nabla f(x^*+\theta td)^{\mathrm{T}}d \quad (0<\theta<1)$$

由于 x^* 为局部极小点，$t>0$，则

$$\nabla f(x^*+\theta td)^{\mathrm{T}}d\geqslant 0$$

取 $d=-\nabla f(x^*)$，则上面所述之不等式仍应成立，即

$$-\nabla f(x^*-\theta t\nabla f(x^*))^{\mathrm{T}}\nabla f(x^*)\geqslant 0$$

令 $t\to 0^+$，则有

$$-\nabla f(x^*)^{\mathrm{T}}\nabla f(x^*)\geqslant 0$$

从而必有

$$\nabla f(x^*)=0$$

若 $\nabla f(x^*)=0$，我们则称 x^* 为一个驻点。由定理 2.2 可知，对于连续可微函数 $f(x)$ 来说，它的任意一个局部极小点必为驻点。

下面，我们来讨论二阶必要条件。先回顾以下相关的概念：一个 $n\times n$ 矩阵 $\boldsymbol{A}$ 被称为正定的(正半定的)，如果对于任意的 n 维向量，$x\neq 0$，则总有 $x^{\mathrm{T}}\boldsymbol{A}x>0(x^{\mathrm{T}}\boldsymbol{A}x\geqslant 0)$。

定理 2.3 设 x^* 是 $f(x)$ 的一个局部极小点，$\nabla^2 f(x)$ 在 x^* 的某个开邻域内连续，则 $\nabla f(x^*)=0$ 且 $\nabla^2 f(x^*)$ 是正半定的。

证明：由定理 2.2 可知，$\nabla f(x^*)=0$。现在，我们证明 $\nabla^2 f(x^*)$ 是正半定的。事实上，根据 Taylor 定理可知，必然存在 $\delta>0$，使得对于任意的 $\lambda\in(0,\ \delta)$，总有

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

既然 $\nabla f(x^*)=0$，$f(x^*+\lambda d)\geqslant f(x^*)$，则我们有

$$\frac{1}{2}\lambda^2 d^{\mathrm{T}}\nabla^2 f(x^*)d+o(\|\lambda d\|^2)\geqslant 0$$

两边同时除以 $\frac{1}{2}\lambda^2$，并令 $\lambda\to 0^+$，可得

$$d^{\mathrm{T}}\nabla^2 f(x^*)d\geqslant 0$$

即 $\nabla^2 f(x^*)$ 必是正半定的。

上面，我们讨论了如果可行点 x^* 是 $f(x)$ 的一个局部极小点，那么它应该具备什么样的性质。下面，我们将讨论可行点 x^* 具备哪些条件时，其一定是一个严格局部极小点。

定理 2.4　设 $\nabla^2 f(x)$ 在 x^* 的某个开邻域内连续，$\nabla f(x^*)=0$ 且 $\nabla^2 f(x^*)$ 是正定的，那么 x^* 必是 $f(x)$ 的一个严格局部极小点。

证明：由于 $\nabla^2 f(x)$ 是连续的且在 x^* 处正定，那么必然存在 $r>0$，使得 $\nabla^2 f(x)$ 在 $D=\{z:\|z-x^*\|<r\}$ 内的每一点处仍是正定的。取 $d\in R^n$，$0<\|d\|<r$，则 $x^*+d\in D$。于是

$$f(x^*+d)=f(x^*)+d^{\mathrm{T}}\nabla f(x^*)+\frac{1}{2}d^{\mathrm{T}}\nabla^2 f(x^*+t\,d)d,\qquad 0<t<1$$

既然，$x^*+td\in D$ 意味着 $\nabla^2 f(x^*+t\,d)$ 必为正定的，从而

$$f(x^*+d)>f(x^*)+d^{\mathrm{T}}\nabla f(x^*)=f(x^*)。$$

注意：定理 2.4 的逆命题不成立。一个简单的例子是：对于 $f(x)=x^4$，显然，$x=0$ 是它的一个严格局部极小点，然而 $\nabla^2 f(0)=(0)$ 却不是正定的。

下面，我们要引入一类重要的函数，它的局部极小点也是全局极小点。若 f：$R^n\to R$ 满足下列不等式：

$$f(\lambda x+(1-\lambda)\ y)\leqslant \lambda f(x)+(1-\lambda)f(y),\quad \forall\lambda\in(0,1),\forall x\ 、\ y\in R^n$$

则称 $f(x)$ 是 R^n 上的一个凸函数。

需要指出的是，一个凸函数的定义域必须是一个凸集 C，即对于任意的 $x,\ y\in R^n$，任意的 $\lambda\in(0,1)$，总有

$$x'\in C,\ y\in C\quad\Rightarrow\quad \lambda x+(1-\lambda)y\in C$$

现在，我们来证：一个凸函数的局部极小点是全局极小点。

定理 2.5　设 $f(x)$ 是 R^n 上的凸函数，则它的任一局部极小点必是全局极小点。若 $f(x)$ 还是可微的，那么，它的任一驻点必为全局极小点。

证明：设 x^* 是一个局部极小点而不是全局极小点，那么，必然存在 $y\in R^n$，使得 $f(x^*)>f(y)$，于是由 $f(x)$ 的凸性可知，

$$f(\lambda y+(1-\lambda)x^*)\leqslant \lambda f(y)+(1-\lambda)f(x^*)<f(x^*),\quad 0<\lambda<1$$

当 λ 充分小时，$\lambda y+(1-\lambda)x^*$ 充分接近于 x^*。这表明，我们总可以在 x^* 的任一邻域内找到一点 $\lambda y+(1-\lambda)x^*$，使得它的函数值严格小于 $f(x^*)$，这与 x^* 是局部极小点是矛盾的。

接下来，我们证明该定理的第二部分。若 x^* 不是全局极小点，且 y 如上所取，那么，由 $f(x)$ 的凸性可知，

$$\nabla f(x^*)^{\mathrm{T}}(y-x^*)=\frac{\mathrm{d}\, f(x^*+\lambda(y-x^*))}{\mathrm{d}\lambda}\Big|_{\lambda=0}$$

$$=\lim_{\lambda\downarrow 0}\frac{f(x^*+\lambda(y-x^*))-f(x^*)}{\lambda}$$

$$=\lim_{\lambda\downarrow 0}\frac{\lambda f(y)+(1-\lambda)f(x^*)-f(x^*)}{\lambda}$$

$$=f(y)-f(x^*)<0$$

因而，$\nabla f(x^*)\neq 0$。于是，x^* 不可能是驻点。

2.2 算法的一般性描述

对于优化问题式(2.1)，如何找出它的一个近似迭代解呢？

我们可以采取逐步逼近的策略。首先，我们选取一个适当的初始点 x^*，然后，找出下一个迭代点 x^1，x^1 应具有 $x^1=x^0+\lambda_0 d^0$ 的形式。其中 $\lambda_0>0$，$d^0\in R^n$，并且目标函数值

$$f(x^0+\lambda_0 d^0)<f(x^0)$$

有足够的下降量。不断重复上边的过程，直到当前的迭代点 x^k 满足某个精度准则，在此情况下我们停止迭代，当前迭代点即为满足精度要求的近似解。

下面，介绍几个相关的概念。

设 f：$R^n\to\mathbf{R}$，$d\in R^n$，$\lambda'>0$。若对于任意的 $\lambda\in(0,\lambda')$，总有

$$f(x+\lambda d)<f(x)$$

则称 d 为 $f(x)$ 在 x 点处的一个下降方向，相关系数 $\lambda>0$ 称为一个步长。

基于这些概念，我们来正式描述一般的算法框架。

(1)选取适当初始点 $x^0\in R^n$，$\varepsilon>0$，$k:=0$；

(2)对于当前迭代点 x^k，给出 $f(x)$ 在 x^k 处的一个下降方向，并给出一个适当的步长 λ_k。

(3)检验 x^k 是否满足停止准则

$$a(x^k)\leqslant\varepsilon$$

若满足，停止迭代，x^k 即为所要找的一个近似解。否则，令 $k:=k+1$，转入第(2)步。

注意：适当选取步长需要一定的技巧。从直觉上讲，我们可以通过解

$$\min_{\lambda\geqslant 0} f(x^k+\lambda d^k) \tag{2.2}$$

来确定步长 λ_k。这样做的好处是目标函数值在迭代后达到了最大下降量，这种步长技巧为精确线搜索法，它的主要缺点是求解优化问题式(2.2)本身往往很难。作为另一种选择，我们可以考虑用不精确线搜索法，比如 Armijo 步长技巧，这一点将在后面作较详细的讨论。

对于停止准则，选择方式应根据目标函数的特点而定。比如，当它是可微的时候，我们可以取 $a(x)=\|\nabla f(x)\|$。

下面，我们重点讨论一下如何确定 $f(x)$ 在当前迭代点 x^k 处的一个下降方向，这是算法设计的一个重要环节。它在很大程度上决定了一个算法的优劣。

如果 $f(x)$ 是可微的，且它在当前迭代点 x^k 处的梯度 $\nabla f(x^k)$ 不为零向量，则负梯度 $-\nabla f(x^k)$ 可以作为一个下降方向。为了说明其中的道理，我们考虑 $f(x)$ 在 x^k 处的一阶 Taylor 展开式，即

$$f(x^k+\lambda d)=f(x^k)+\lambda\nabla f(x^k)^{\mathrm{T}}d+o(\|\lambda d\|)$$

$$=f(x^k)+\lambda\|\nabla f(x^k)\|\|d\|\cos\theta+o(\|\lambda d\|),\lambda>0$$

其中，θ 为 $\nabla f(x^k)$ 与 d 之间的夹角。显然，

$$d=-\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

在所有的单位向量中，使 $\nabla f(x^k)^{\mathrm{T}}d$ 达到了极值。因此，负梯度 $-\nabla f(x^k)$ 称为最速下降方向。

利用 Taylor 一阶展开式，我们构造了 $f(x)$ 在 x^k 处的一个下降方向——最速下降方向。现在，我们利用 Taylor 二阶展开式来构造它在 x^k 处的另一类重要的下降方向。为此，考虑

$$f(x^k+d)=f(x^k)+\nabla f(x^k)^{\mathrm{T}}d+\frac{1}{2}d^{\mathrm{T}}\nabla^2 f(x^k)d+o(\|d\|^2)$$

这里，我们假设 $\nabla^2 f(x^k)$ 存在并且是正定的，将上式右边前 3 项极小化，便可得到

$$d^k=-\left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k)。$$

由定义，可以验证 d^k 为 $f(x)$ 在 x^k 处的一个下降方向，它往往称作 $f(x)$ 在 x^k 处的 Newton 方向。

与最速下降方向相比，Newton 方向有一个潜在的优势：取 x^k+d^k 作为下一个迭代点，仅需较少的迭代次数便可高精度地逼近问题的最优解，这是因为它利用了 Hesse 阵所提供的曲率信息的缘故。而 Newton 方向的主要缺点在于：它要求 Hesse 阵的估值。一般来说，高精度地计算这样一个 Hesse 阵需要大量的运算，过程往往相当麻烦。

既然如此，人们期望找到一个不利用 Hesse 阵的下降方向，且仍能使算法保持类似的较快收敛性的方法。

下面将要提到的拟 Newton 方向便是这样一类理想的下降方向。它不直接使用包含二阶导数信息的 Hesse 阵，而是利用一阶导数的信息来巧妙逼近它。

事实上，由 Taylor 定理可知

$$\nabla f(x+d)=\nabla f(x)+\nabla^2 f(x+d)d+\int_0^1[\nabla^2 f(x+td)-\nabla^2 f(x+d)]d\,\mathrm{d}t \tag{2.3}$$

显然，如果 $\nabla^2 f(x)$ 是在 R^n 上 Lipschitz 连续的，即存在一个 $L>0$，使得对于任意的 x、$y\in R^n$，总有

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \leqslant L \| x - y \|$$

那么，由式(2.3)可以得到

$$\nabla f(x+d) = \nabla f(x) + \nabla^2 f(x+d)d + o(\| d \|)$$

取 $x := x^k$，$d := x^{k+1} - x^k$，我们有

$$\nabla f(x^{k+1}) = \nabla f(x^k) + \nabla^2 f(x^{k+1})(x^{k+1} - x^k) + o(\| x^{k+1} - x^k \|)$$

当 x^k、x^{k+1}都充分接近解 x^*时，便可得到

$$\nabla^2 f(x^{k+1})(x^{k+1} - x^k) \approx \nabla f(x^{k+1}) - \nabla f(x^k) \tag{2.4}$$

现在，我们可以考虑，相应的拟 Newton 方程(也称作割线方程)，

$$\boldsymbol{B}_{k+1} s_k = y_k$$

其中，$\boldsymbol{B}_{k+1}$为对称矩阵；$s_k = x^{k+1} - x^k$；$y_k = \nabla f(x^{k+1}) - \nabla f(x^k)$。 (2.5)

显而易见，$\boldsymbol{B}_{k+1}$模拟了 $\nabla^2 f(x^{k+1})$的性质，可以看作它的一个近似表示。考虑到理论上的 Hesse 阵 $\nabla^2 f(x^{k+1})$是对称的，我们也要求 $\boldsymbol{B}_{k+1}$对称。另外，我们还要求 $\boldsymbol{B}_{k+1} - \boldsymbol{B}_k$具有较低的秩。至于 $\boldsymbol{B}_0$，可以取为单位阵。这样，我们便得到了拟 Newton 方向的一种选择

$$d^k = -\boldsymbol{B}_k^{-1} \nabla f(x^k)$$

有点令人不满意的是，它出现了 $\boldsymbol{B}_k$的求逆运算。事实上，这一点是可以避免的，详细的讨论可参见第 8 章。

在讨论算法时，我们常常提到收敛率的概念，它是反映算法收敛快慢的一个指标。

收敛率的类型有若干种，下面我们给出几种常见的形式。

设$\{x^k\}$是 R^n 中收敛于 $x^* \in R^n$ 的一个序列。若存在一个常数 $\theta \in (0,1)$，使得

$$\frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|} \leqslant \theta$$

对于所有充分大的 k 都成立，我们称序列$\{x^k\}$比式线性收敛于 x^*。

比如，取 $x^k = 0.5^k$，$k = 0,1,\cdots$，则序列$\{x^k\}$比式线性收敛于 0。

如果进一步有

$$\lim_{k \to +\infty} \frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|} = 0$$

我们称序列$\{x^k\}$比式超线性收敛于 x^*。

比如，取 $x^k = k^{-k}$，则

$$\frac{| x^{k+1} - 0 |}{| x^k - 0 |} = \frac{| (k+1)^{-k-1} - 0 |}{| k^{-k} - 0 |} = \frac{1}{k+1} \cdot \frac{1}{(1+\frac{1}{k})^k}$$

两边取极限，可得

$$\lim_{k\to+\infty}\frac{|x^{k+1}-0|}{|x^k-0|}=\lim\frac{1}{k+1}\cdot\frac{1}{\lim(1+\frac{1}{k})^k}=0\cdot e^{-1}=0$$

这表明，序列$\{x^k\}$比式超线性收敛于0。

一个更快的收敛率为比式二次收敛：序列$\{x^k\}$收敛于x^*，且存在一个正数 M(未必小于 1)，使得对于充分大的k，总有

$$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^2}\leqslant M$$

从本质上讲，对于一个序列而言，超线性收敛比线性收敛快而比二次收敛慢。

下面，我们讨论根式收敛的概念，若存在一个比式线性收敛于0的序列$\{v_k\}$，使得

$$\|x^k-x^*\|\leqslant v_k,\quad k=0,1,\cdots$$

我们则称序列$\{x^k\}$根式线性收敛于x^*。

比如，序列

$$x^k=\begin{cases}0.5^k, & k\text{为偶数}\\ 0, & k\text{为奇数}\end{cases}$$

根式线性收敛于0。

类似地，我们还可以定义序列的根式超线性收敛等。

第3章　线搜索方法

顾名思义，线搜索方法是指确定了搜索方向 d 后，沿其寻找下一个迭代点的方法。

下面，我们讨论这样一个问题：若目标函数 $f(x)$是可微的，那么在什么样的条件下，d 是 $f(x)$在 x 处的一个下降方向？

定理 3.1　**设** $f: R^n \to \mathbf{R}$ **在** x **处可微，若存在** $d \in R^n$，**使得**

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

则 d **必为** $f(x)$**在** x **处的一个下降方向。**

证明：由 Taylor 定理，对于任意的 $\lambda>0$，我们有

$$f(x+\lambda d) = f(x) + \lambda \nabla f(x)^{\mathrm{T}} d + o(\| \lambda d \|)$$

既然 $\lambda \nabla f(x)^{\mathrm{T}} d < 0$，从而必然存在 $\delta>0$，使得当 $\lambda \in (0, \delta)$时，

$$\nabla f(x)^{\mathrm{T}} d + \frac{o(\| \lambda d \|)}{\lambda} < 0$$

于是

$$f(x+\lambda d) = f(x) + \lambda (\nabla f(x)^{\mathrm{T}} d + \frac{o(\| \lambda d \|)}{\lambda}) < f(x)$$

根据定义，d 为 $f(x)$在 x 处的一个下降方向。

显然，在一个算法中，若 x^k 不是可微函数 $f(x)$ 的一个驻点，即 $\nabla f(x^k) \neq 0$，则对于任意的正定阵 $\boldsymbol{B}_k$ 来说，

$$d^k = -\boldsymbol{B}_k^{-1} \nabla f(x^k)$$

均可作为 $f(x)$ 在 x^k 处的一个下降方向。原因是

$$\nabla f(x^k)^{\mathrm{T}} d^k = -\nabla f(x^k)^{\mathrm{T}} \boldsymbol{B}_k^{-1} \nabla f(x^k) < 0 \text{。}$$

对于一个算法而言，一旦下降方向确定后，我们便可以沿其搜索下一个迭代点 x^{k+1}。从直觉上讲，我们期望

$$x^{k+1} := x^k + \alpha_k d^k = \arg\min_{\alpha>0} f(x^k + \alpha d^k)$$

其中，argmin 表示极小点。

从理论上讲，这样做的好处是：它可以使目标函数值得到最大限度的下降量。然而，从计算角度分析的话，它需要目标函数，甚至其梯度的多次估值。由此看来，用这样精确线搜索的方式确定步长 α_k 并非一种理想的选择。

实际上，在一个算法中，人们常常利用不精确线搜索的方式来确定步长 α_k，这样的