# Note of numerical optimaization

PB19000252 王梓睿

2022 年 9 月 4 日

## 1 introduction

1、连续优化问题通常比离散优化问题容易，因为连续性带来在特定点附近，限制条件不会变化太大。

2、有时候约束优化问题可以表示成无约束优化问题，方法是将约束添加到目标函数的惩罚项当中，从而阻止违反约束。

3、凸优化的局部最优解就是全局最优解。

证明. 设 x 为局部最优解（最小值），y≠x 为全局最优解，连接 y 与 x，路径与 x 的任一邻域有交，从而由凸函数性质，与 x 的局部最优性矛盾. □

本书大部分情况下我们只关心局部最优。

4、随机最优化解出的是模型的期望表现。

5、凸优化的定义：

⓪是约束优化问题的一个特例；

①目标函数为凸函数；

②约束等式中右侧函数为线性函数；

③约束不等式中右侧函数为凹函数；

# 2 无约束最优化

## 2.1 关于解

1、孤立局部最小值一定严格局部最小，反之则不一定。

2、Taylor 公式（利用两种余项），由此可以推导几个结论。

①f 在局部最小值点 x* 邻域内连续可微，则 $\nabla f(x*) = 0$；

②f 在局部最小值点 x* 邻域内，二阶导存在连续，则 $\nabla f(x*) = 0$ 且 $\nabla^2 f(x*)$ 半正定；

③（充分性）在某点 x* 邻域内二阶导存在连续，且 f(x*)=0,$\nabla^2 f(x*)$ 正定，那么 x* 为 f 的局部最小值点；

3、凸函数的局部最优值为全局最优值。若还可微，则任意稳定点都是全局最优值。

## 2.2 算法综述

1、线性探测法的最速下降方向是 $-\nabla f_k$，但收敛速度过慢。容易证明在与 $-\nabla f_k$ 夹角小于 $\pi/2$ 的方向上目标函数总会递减，称这些方向为下降方向。

2、牛顿方向为 $-(\nabla^2 f_k)^{-1}\nabla f_k$ 适用于 $f_k$ 的 Hessian 矩阵正定，且函数越光滑误差越小。它收敛速度较快，但 Hessian 矩阵计算量较大。

3、准牛顿方法用近似的 $B_k$ 矩阵代替 Hessian，给出 $-(B_k)^{-1} f_k$ 并在迭代中更新 $B_k$，从而减小计算量。

4、非线性共轭方法，采用公式 $p_k = -\nabla f_k + \beta_k p_{k-1}$，保留了最速下降法的计算简便性的同时提高了其收敛速度。

5、缩放，可以通过对角缩放等方式将变量大小拉到同一数量级。牛顿法能够比最速下降法更好地处理 poorly scaled 问题。

# 3 线性探测法

线性探测法先计算行进方向 $p_k$，再计算步长 $\alpha_k$，迭代公式为 $x_{k+1} = x_k + \alpha_k p_k$。行进方向的选取通常要求 $f_k$ 递降，一般令 $p_k = -(B_k)^{-1}\nabla f_k$，如果 $B_k$ 正定，由 Taylor 公式，一次项可化为 $-\nabla f_k^T B_k^{-1}\nabla f_k < 0$，这样 $f_k$ 能够递降。

## 3.1 步长

给定初值 $x_k$ 和方向 $p_k$，我们记 $f_k(x_k + \alpha_k p_k)$ 为单变量函数 $\phi(\alpha_k)$，想让其尽可能的小。我们引入概念“充分递降”。

### 3.1.1 Wolfe 条件

1、Armijo 条件：$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1\alpha_k \nabla f_k^T p_k$。它保证了迭代过程至少线性收敛。

2、曲率条件：$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k$。事实上就是 $\phi'(\alpha_k) \geq c_2\phi'(0)$，来保证下降的梯度在某一区间内。

3、$0 < c_1 < c_2 < 1$ 时，以上两式合称 Wolfe 条件。若 2 式改写为 $|LHS| \leq |RHS|$，则称为强 Wolfe 条件（进一步缩小了梯度的可能区间）。引理：任一连续可微有下界的函数，一定有满足（强）Wolfe 条件的区间。

### 3.1.2 其它

1、Goldenstein 条件：$f(x_k) + (1-c_1)\alpha_k \nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c_1\alpha_k \nabla f_k^T p_k$。

2、回溯线性搜索：仅用 Armijo 条件进行判断，但每次不满足都执行 $\alpha = \rho\alpha$，$\rho$ 为收缩因子。$\rho$ 可以在迭代过程中发生变化。

## 3.2 收敛性

考虑 $\theta_k$ 为方向 $p_k$ 与最速下降方向 $-\nabla f_k$ 的夹角。

1、Zoutendijk 条件：对满足一系列条件的 f 成立

$$\sum_{0}^{\infty} cos^2\theta_k ||\nabla f_k||^2 < \infty$$

**Theorem 3.2.**

*Consider any iteration of the form (3.1), where $p_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions (3.6). Suppose that $f$ is bounded below in $\mathbb{R}^n$ and that $f$ is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} \stackrel{\text{def}}{=} \{x : f(x) \le f(x_0)\}$, where $x_0$ is the starting point of the iteration. Assume also that the gradient $\nabla f$ is Lipschitz continuous on $\mathcal{N}$, that is, there exists a constant $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \le L\|x - \tilde{x}\|, \quad \text{for all } x, \tilde{x} \in \mathcal{N}. \tag{3.13}$$

*Then*

$$\sum_{k \ge 0} \cos^2\theta_k \, \|\nabla f_k\|^2 < \infty. \tag{3.14}$$

2、如果上述级数收敛，说明 $\cos^2\theta_k ||\nabla f_k||^2 \to 0$。当 $\theta_k$ 有界地远离 90 度时，即存在 $\delta, cos\theta_k >= \delta > 0$，那么有 $||\nabla f_k \to 0$，我们称为全局收敛。

3、类牛顿方法中，如果矩阵 $B_k$ 的条件数 $||B_k||||B_k^{-1}|| \le M$，M 为一个常数，则可以证明全局收敛性。

## 3.3 收敛率

1、我们可以定义 Q 范数，并证明 $f(x) = \frac{1}{2}x^TQx - b^Tx$ 时，有 $\frac{1}{2}||x - x^*||_Q^2 = f(x) - f(x^*)$。

2、thm3.3，可以证明对严格凸的二次函数 f，有

$$||x_{k+1} - x^*||_Q^2 \le (\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1})^2 ||x_k - x^*||_Q^2.$$

其中 $0 < \lambda_1 \le \cdots \le \lambda_n$ 为 Q 的特征值。这意味着 $f_k$ 至少以线性收敛到 $f^*$。

3、thm3.5，如果 f 二次可微，Hessian 在解 $x^*$ 附近李普希兹连续，那么 x

(i) 若起始点 $x_0$ 离 $x^*$ 足够近，那么迭代序列收敛到 $x^*$。

(ii)$x_k$ 有二阶收敛速度，$||\nabla f_k||$ 二阶收敛到 0.

4、thm3.6，此定理说明了在 $p_k$ 满足一个极限式的时候，k 较大时可直接将 $\alpha_k$ 取 1，且此时 $x_k$ 超线性收敛到 $x^*$。thm3.7 说明了 $\alpha_k$ 恒取为 1 的情况下，3.6 的条件是充要的。而准牛顿方法满足这一条件，因此能够超线性收敛。

## 3.4 Hessian 修正

我们希望能够找到方阵 $E_k$，使得 $B_k = \nabla^2 f(x_k)$ 正定。要求这一修改尽量小，以保留 Hessian 矩阵的更多信息。1、可以考虑 $\nabla^2 f_k + \tau I$，并在迭代中不断增加 $\tau_k$ 的值。但是这样每次更新 $\tau$ 后都要重新对矩阵进行分解，计算量太大。

2、Cholesky 分解，可以在算法中引入两个新的变量列，从而边分解边进行矩阵修改。（注意并不是先分解为 $L^T DL$ 再修改 D，因为首先我们无法保证分解的存在性，其次修改后的结果不稳定）

3、更稳定的方法是进行对称不定分解 $A = L^T BL$，其中 A 须为对称矩阵，B 为准对角阵，对角块大小为 1 或 2. 这一分解必定存在。

## 3.5 步长选择算法

### 3.5.1 插值算法

1、从 $\phi(0), \phi'(0)$, 及初始步长 $\alpha_0$ 构造出 $\alpha_1$，随后构造出三次插值函数，利用 $\phi(0), \phi'(0), \alpha_{k-1}, \alpha_k$ 来计算 $\alpha_{k+1}$。这一方法并没有计算在诸 $\alpha_i$ 处的导数值。

2、如果导数值比较好算，我们可以用 $\phi(\alpha_{i-1}), \phi(\alpha_i), \phi'(\alpha_{i-1}), \phi'(\alpha_i)$ 来对 $\alpha_{i+1}$ 作三次插值。

### 3.5.2 Wolfe 条件下的线性搜索算法

先利用两个不等式确定一个目标值存在的区间，再调用 Zoom 函数不断缩小区间的范围，以得到满足 Wolfe 条件的值。