4-1 最速下降法

最速下降法

1874年法国科学家Cauchy提出了**最速下降法**(steepest descent method)，其主要思想是以负梯度方向作为下降方向的极小化算法，又称梯度法，最速下降法是无约束最优化中最简单的方法。

设目标函数$f(\boldsymbol{x})$在当前迭代点$\boldsymbol{x}_k$附近连续可微，且$\boldsymbol{g}_k \triangleq \nabla f(\boldsymbol{x}_k) \neq 0$, 将$f(\boldsymbol{x})$在$\boldsymbol{x}_k$处做泰勒展开：

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \alpha \boldsymbol{g}_k^T \boldsymbol{d}_k + o(\| \alpha \boldsymbol{d}_k \|). \tag{4.1.1}$$

记$\boldsymbol{x} - \boldsymbol{x}_k = \alpha \boldsymbol{d}_k$。

若搜索方向$\boldsymbol{d}_k$满足: $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$, 则$\boldsymbol{d}_k$是下降方向，即选取合适的步长$\alpha$, 有：$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$成立。

最速下降法

当α取定后，$\boldsymbol{d}_k^T\boldsymbol{g}_k$的值越小，即$-\boldsymbol{d}_k^T\boldsymbol{g}_k$的值越大，函数$f(\boldsymbol{x})$在$\boldsymbol{x}_k$处下降量越大。

由Cauchy-Schwartz不等式:

$$|\boldsymbol{g}_k^T\boldsymbol{d}_k| \leq \|\boldsymbol{g}_k\| \cdot \|\boldsymbol{d}_k\| \tag{4.1.2}$$

当且仅当:

$$\text{取}\boldsymbol{d}_k = -\boldsymbol{g}_k\text{时，}\boldsymbol{g}_k^T\boldsymbol{d}_k\text{最小，}-\boldsymbol{g}_k^T\boldsymbol{d}_k\text{最大，}$$

从而$-\boldsymbol{g}_k$是最速下降方向。

一般地，称以负梯度方向为迭代方向的方法为负梯度方法。

最速下降法步骤

特别地，采用精确线性搜索的步长，以负梯度方向为下迭代方向的方法叫最速下降法(steepest descent, SD)，迭代格式为:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$$

算法4.1.1 – 最速下降法

步1 给出$\boldsymbol{x}_0 \in \mathbb{R}^n$，$0 \leq \varepsilon \ll 1$，$k := 0$;

步2 若停机条件满足(比如$\| \boldsymbol{g}_k \| \leq \varepsilon$)，则迭代停止;

步3 计算$\boldsymbol{d}_k = -\boldsymbol{g}_k$;

步4 一维精确线性搜索求α_k;

步5 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, $k := k + 1$，转步2.

最速下降法–正定二次函数

考虑正定二次函数：

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T G\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c \tag{4.1.3}$$

其中$G \in \mathbb{R}^{n\times n}$为正定矩阵, 极小点$\boldsymbol{x}^*$满足：$G\boldsymbol{x} + \boldsymbol{b} = 0$。

计算迭代方向$\boldsymbol{d}_k = -\boldsymbol{g}_k = -G\boldsymbol{x}_k - \boldsymbol{b}$。求解一维问题：$\min_{\alpha>0} f(\boldsymbol{x}_k - \alpha\boldsymbol{g}_k)$得到最优步长因子为：

$$\alpha_k = \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k G\boldsymbol{g}_k}$$

下一个迭代点为：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k G\boldsymbol{g}_k}\boldsymbol{g}_k.$$

最速下降法–正定二次函数

例4.1.1 正定二次函数(4.1.3)中, $\boldsymbol{b} = (2,3)^T, c = 10$, 二阶海森矩阵$G$分别取:

$$G_1 = \begin{bmatrix} 21 & 4 \\ 4 & 15 \end{bmatrix} \quad G_2 = \begin{bmatrix} 21 & 4 \\ 4 & 1 \end{bmatrix}$$

用最速下降法分别求解(4.1.3)的极小点。初始点取$\boldsymbol{x}_0 = (-30, 100)^T$, 终止准则为:$\|\boldsymbol{g}_k\| < 10^{-5}$。

解: 利用最速下降法求解**问题1**(G_1)和**问题2**(G_2)，并给出迭代点的信息与相应梯度变化信息，对比针对两个问题最速下降法收敛速度的快慢。

最速下降法–正定二次函数

问题1: G_1

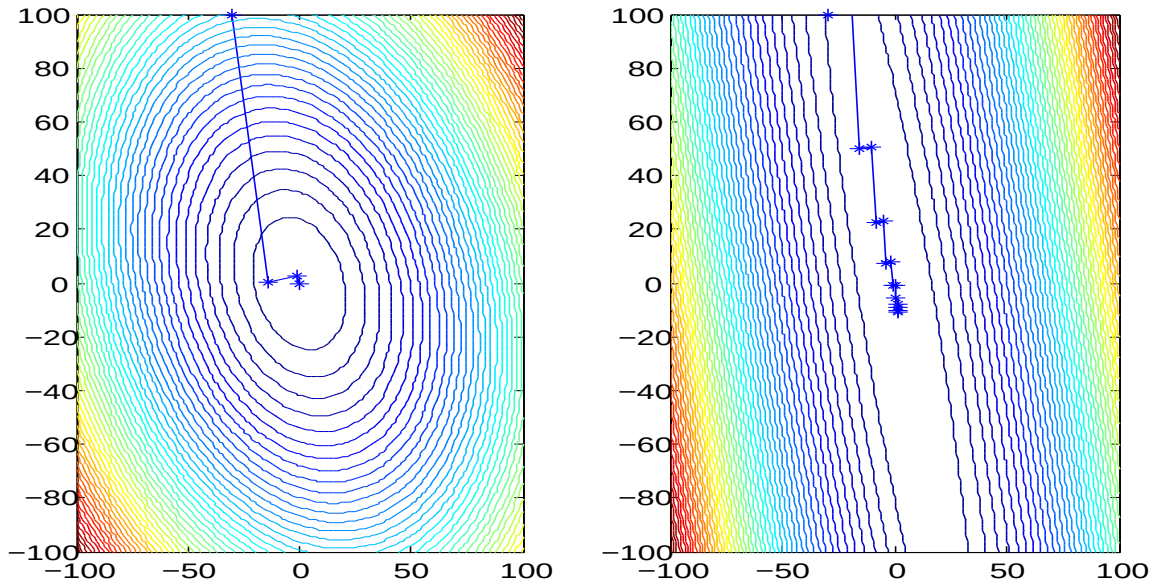
k	$\boldsymbol{x}_k^T$	$\|\boldsymbol{g}_k\|$
0	$(-30.0000, 100.0000)$	1401.6679
1	$(-13.5763, 0.3277)$	285.4239
2	$(-0.8387, 2.4212)$	36.4480
⋮	⋮	⋮
11	$(-0.0602, -0.1840)$	0.3393e−005
12	$(-0.0602, -0.1840)$	0.4333e−006

问题2: G_2

k	$\boldsymbol{x}_k^T$	$\|\boldsymbol{g}_k\|$
0	$(-30.0000, 100.0000)$	228.6329
1	$(-19.3868, 1000.7913)$	26.3171
2	$(-15.6406, 50.0660)$	125.7811
⋮	⋮	⋮
58	$(2.0000, -11.0000)$	0.6807e−005
59	$(2.0000, -11.0000)$	0.7835e−006

最速下降法在问题1上的收敛速度明显快于在问题2的收敛速度。

最速下降法–正定二次函数



数值试验表明：

- 当目标函数的等值线接近于一个圆(球)时，下降较快
- 当目标函数的等值线是一个扁长的椭球时，开始几步下降较快，后来就出现锯齿现象，下降十分缓慢。

最速下降法–锯齿现象

由于精确线性搜索满足$\boldsymbol{g}_{k+1}^T\boldsymbol{d}_k = 0$,则

$$\boldsymbol{g}_{k+1}^T\boldsymbol{g}_k = \boldsymbol{d}_{k+1}^T\boldsymbol{d}_k = 0, \tag{4.1.4}$$

这表明最速下降法:

- 相邻两次的搜索方向是相互直交的，这就产生了锯齿形状。
- 越接近极小点，步长越小，前进越慢。

课堂练习 试用最速下降法求$f(\boldsymbol{x}) = (x_1 - 1)^2 + (x_2 - 1)^2$的极小点。已知初始点$\boldsymbol{x}_0 = (0, 0)$, 梯度误差精度$\varepsilon = 0.1$。

最速下降法收敛速度

定义4.1.1 设G是$\mathbb{R}^{n\times n}$对称正定, $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n$, 则$\boldsymbol{u}$与$\boldsymbol{v}$在$G$度量意义下的内积$(\boldsymbol{u}^T\boldsymbol{v})_G$定义为:

$$(\boldsymbol{u}^T\boldsymbol{v})_G=\boldsymbol{u}^TG\boldsymbol{v}$$

$\boldsymbol{u}$在G度量意义下的范数定义为:

$$\|\boldsymbol{u}\|_G=\boldsymbol{u}^TG\boldsymbol{u}$$

对**正定二次函数情形** $f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^TG\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{x}$, 推导可得:

$$\frac{1}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|_G^2=f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*) \tag{4.1.5}$$

在G度量意义下, $\boldsymbol{x}_k$的误差等价于它们目标函数值$f(\boldsymbol{x}_k)$的误差。

最速下降法收敛速度

定理4.1.1 最速下降法收敛速度.

对正定二次函数，最速下降法的收敛速度为：

$$\frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|_G^2}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|_G^2} \leq \left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2 \tag{4.1.6}$$

其中：$\lambda_{\max},\lambda_{\min}$分别为$G$的最大与最小特征值。

证明： 针对正定二次函数的最速下降法满足：

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k) - \frac{1}{2}\frac{(\boldsymbol{g}_k^T\boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T G\boldsymbol{g}_k}$$

由于$G\boldsymbol{x}^* = -\boldsymbol{b}$得:

$$f(\boldsymbol{x}^*) = -\frac{1}{2}\boldsymbol{b}^T G^{-1}\boldsymbol{b}$$

定理4.1.1证明续

证明续： 从而：

$$\frac{f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)}=\frac{f(\boldsymbol{x}_k)-\frac{1}{2}\frac{(\boldsymbol{g}_k^T\boldsymbol{g}_k)^2}{\boldsymbol{g}_k^TG\boldsymbol{g}_k}-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)}=1-\frac{(\boldsymbol{g}_k^T\boldsymbol{g}_k)^2}{(\boldsymbol{g}_k^TG\boldsymbol{g})k)(\boldsymbol{g}_k^TG^{-1}\boldsymbol{g}_k)}$$

由(4.1.5)式得：

$$\frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|_G^2}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|_G^2}=1-\frac{(\boldsymbol{g}_k^T\boldsymbol{g}_k)^2}{(\boldsymbol{g}_k^TG\boldsymbol{g})k)(\boldsymbol{g}_k^TG^{-1}\boldsymbol{g}_k)}$$

根据Kantorovich不等式：对任意的$\boldsymbol{x}\in\mathbb{R}^n\backslash\{0\}$, 均有以下不等式成立：

$$\frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{(\boldsymbol{x}^TG\boldsymbol{x})(\boldsymbol{x}^TG^{-1})\boldsymbol{x}}\geq\frac{4\lambda_{\mathsf{max}}\lambda_{\mathsf{min}}}{(\lambda_{\mathsf{max}}+\lambda_{\mathsf{min}})^2}$$

根据以上两式得(4.1.6)结论。 ■

最速下降法收敛速度

从**定理**4.1.1可以看出，最速下降法的收敛速度是线性的，这个速度依赖于G的最大、最小特征值。在二范数意义下，矩阵条件数为：

$$\mathsf{cond}(G) = \|G\|_2\|G^{-1}\|_2 = \frac{\lambda_{\mathsf{max}}}{\lambda_{\mathsf{min}}}$$

所以有：

$$\frac{\lambda_{\mathsf{max}} - \lambda_{\mathsf{min}}}{\lambda_{\mathsf{max}} + \lambda_{\mathsf{min}}} = \frac{\mathsf{cond}(G) - 1}{\mathsf{cond}(G) + 1} \triangleq \mu$$

这说明最速下降法收敛速度依赖于G的条件数：

- 当G条件数接近于1时, μ接近于零, 最速下降法的收敛速度接近超线性收敛速度。
- 当G条件数越大, μ越接近于1，该方法的收敛速度越慢。

最速下降法–优缺点

最速下降法优缺点：

- **优点：** 具有程序设计简单，计算工作量小，存储量小，对初始点没有特别要求，从不太好的初始点出发也可能接近极小点。
- **缺点：** 最速下降方向仅是函数的局部性质，对整体求解过程而言，这个方法下降非常缓慢。

最速下降法收敛性

定理4.1.2 总体收敛性.

设$\nabla f(\boldsymbol{x})$在水平集$L = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$上存在且一致连续，则**最速下降法**产生的序列满足: 或对某个k 有$\boldsymbol{g}_k = 0$，或$f(\boldsymbol{x}_k) \to -\infty$, $\boldsymbol{g}_k \to 0$.

证明： 对于最速下降法，有$\theta_k = 0$，利用精确线性搜索收敛性定理-**定理3.1.2**立即可知最速下降法是总体收敛的。 ■

定理4.1.3.

设函数$f(\boldsymbol{x})$二阶连续可微，且$\|\nabla^2 f(\boldsymbol{x})\| \leq M$，其中$M$是某正常数。对任何给定的初始点$\boldsymbol{x}_0$，**最速下降算法4.1.1**或有限终止，或$\lim\limits_{k\to\infty} f(\boldsymbol{x}_k) = -\infty$，或$\lim\limits_{k\to\infty} \boldsymbol{g}_k = 0$.

定理4.1.3证明

证明： 考虑无限迭代下去的情形，由精确线性搜索函数值下降估计**定理**3.1.1，有

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{1}{2M}\|\boldsymbol{g}_k\|^2.$$

对$k = 0, 1, \ldots,$ 进行累加得：

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k) = \sum_{i=0}^{k-1}[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{i+1})] \geq \frac{1}{2M}\sum_{i=0}^{k-1}\|\boldsymbol{g}_i\|^2.$$

两边取极限得：

$$\lim_{k\to\infty} f(\boldsymbol{x}_k) = -\infty \text{或者} \lim_{k\to\infty} \boldsymbol{g}_k = 0$$

从而定理成立。 ■

BB梯度法

最速下降法利用前一个迭代点梯度信息确定当前搜索步长。类似于拟牛顿确定搜索方向，Barzilai&Borwein提出了利用两点梯度信息确定步长的梯度法。在拟牛顿法中，迭代公式为：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - H_k \boldsymbol{g}_k$$

这里H_k为海森矩阵的近似，需满足拟牛顿条件：

$$\boldsymbol{s}_{k-1} = H_k \boldsymbol{y}_{k-1}$$

其中，H_k为海森矩阵G_k的近似，$\boldsymbol{s}_{k-1} = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1} = \boldsymbol{g}_k - \boldsymbol{g}_{k-1}$。将最速下降法迭代格式：$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$, 写成：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - D_k \boldsymbol{g}_k$$

其中：$D_k = \alpha_k I_k$.

BB梯度法

为使D_k具有类似拟牛顿性质，需要通过优化以下问题确定步长α_k:

$$\min \|\boldsymbol{s}_{k-1} - D_k \boldsymbol{y}_{k-1}\| \tag{4.1.7}$$

求解以上优化问题可得：

$$\alpha_k^{\mathsf{BB1}} = \frac{\boldsymbol{s}_{k-1}^T \boldsymbol{y}_{k-1}}{\boldsymbol{y}_{k-1}^T \boldsymbol{y}_{k-1}} \tag{4.1.8}$$

利用对称性，也可以考虑以下优化问题：

$$\min \|D_k^{-1} \boldsymbol{s}_{k-1} - \boldsymbol{y}_{k-1}\| \tag{4.1.9}$$

于是得到步长α_k为：

$$\alpha_k^{\mathsf{BB2}} = \frac{\boldsymbol{s}_{k-1}^T \boldsymbol{s}_{k-1}}{\boldsymbol{s}_{k-1}^T \boldsymbol{y}_{k-1}} \tag{4.1.10}$$

BB梯度法

算法4.1.2 BB梯度下降法

步1 给定初始点$\boldsymbol{x}_0$ $0 < \varepsilon \ll 1$, 令$k = 0$;

步2 若$\|\boldsymbol{g}_k\| \leq \varepsilon$, 停止；否则，令$\boldsymbol{d}_k = -\boldsymbol{g}_k$;

步3 若$k = 0$, 利用线性搜索确定α_0; 否则，利用(4.1.8)或者(4.1.10)计算α_k;

步4 进行迭代：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k,$$

令$k := k + 1$, 并转步1.

BB法与最速下降、极小梯度

考虑极小化正定二次函数：

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T G\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}$$

由$\boldsymbol{g}_k = G\boldsymbol{x}_k + \boldsymbol{b}$可得：

$$\boldsymbol{s}_{k-1} = -\alpha_{k-1}\boldsymbol{g}_{k-1}, \boldsymbol{y}_{k-1} = -\alpha_{k-1}G\boldsymbol{g}_{k-1}$$

BB法的两个步长公式分别化为：

$$\alpha_k^{\text{BB1}} = \frac{\boldsymbol{g}_{k-1}^T\boldsymbol{g}_{k-1}}{\boldsymbol{g}_{k-1}^T G\boldsymbol{g}_{k-1}} \tag{4.1.11}$$

$$\alpha_k^{\text{BB2}} = \frac{\boldsymbol{g}_{k-1}^T G\boldsymbol{g}_{k-1}}{\boldsymbol{g}_{k-1}^T G^2\boldsymbol{g}_{k-1}} \tag{4.1.12}$$

BB法与最速下降、极小梯度

最速下降(SD)与极小梯度(MD)步长分别为：

$$\alpha_k^{\mathsf{SD}} = \mathsf{argmin}_{\alpha>0} f(\boldsymbol{x}_k - \alpha \boldsymbol{g}_k) = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T G \boldsymbol{g}_k} \tag{4.1.13}$$

$$\alpha_k^{\mathsf{MD}} = \mathsf{argmin}_{\alpha>0} \|\boldsymbol{g}(\boldsymbol{x}_k - \alpha \boldsymbol{g}_k)\|_2^2 = \frac{\boldsymbol{g}_k^T G \boldsymbol{g}_k}{\boldsymbol{g}_k^T G^2 \boldsymbol{g}_k} \tag{4.1.14}$$

可以看出BB方法与最速下降、极小梯度方法的步长关系如下：

$$\alpha_k^{\mathsf{BB1}} = \alpha_{k-1}^{\mathsf{SD}}, \quad \alpha_k^{\mathsf{BB2}} = \alpha_{k-1}^{\mathsf{MD}}$$

BB方法的步长相较于SD方法与MD方法延后一步使用，从实际计算效果看，BB方法优于其他两种方法。

SD法、MD法、两种BB法数值比较

例4.1.2 分别利用SD方法、MD方法、两种BB法求解正定二次函数极小点，其中$G = \mathrm{diag}(1,5,10,20), \boldsymbol{b} = \boldsymbol{0}$。初始点取为$(1,1,1,1)^T$. BB方法的初始步长取为$\alpha_0^{\mathrm{BB1}} = \alpha_0^{\mathrm{BB2}} = \alpha_0^{\mathrm{SD}}$, $\|\boldsymbol{g}_k\|_2 \leq 10^{-8}$时停止迭代。

求解该问题时，SD法与MD法分别用了179次和174次迭代，而BB1与BB2分别迭代了36次和44次。

	SD 方法		MG 方法		BB1 方法		BB2 方法	
k	α_k^{SD}	$f(x_k)$	α_k^{MG}	$f(x_k)$	α_k^{BB1}	$f(x_k)$	α_k^{BB2}	$f(x_k)$
0	0.058	1.8e+1	0.054	1.8e+1	0.058	1.8e+1	0.058	1.8e+1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.079	7.9e−2	0.077	7.6e−2	0.162	5.8e−2	0.973	4.0e−4
11	0.120	6.4e−2	0.126	6.4e−2	0.050	2.9e−1	0.052	2.8e−2
12	0.079	5.2e−2	0.077	4.9e−2	0.050	5.1e−5	0.050	5.1e−4
13	0.120	4.2e−2	0.126	4.2e−2	0.095	1.3e−5	0.072	2.4e−4
14	0.079	3.4e−2	0.077	3.2e−2	0.100	1.1e−7	0.166	9.4e−5

观察以上表格发现，SD法与MD法的步长出现周期现象，BB方法目标函数值并非一直单调减小的。