3.4 共轭梯度法

由Taylor公式可知：一个函数在某一极小值点附近的性态与一个二次函数很接近. 因此，为了给出有效的算法，往往采用二次模型来尝试，再推广到一般的非线性函数. 本节首先介绍求解凸二次函数

$$f(x) = \frac{1}{2}x^\top Gx + q^\top x \tag{3.4.1}$$

极小化问题的共轭方向法，其中$G \in \mathbb{R}^{n\times n}$为对称正定矩阵且$q \in \mathbb{R}^n$, 然后介绍共轭方向法中非常有效的共轭梯度法. 算法的计算效果介于最速下降法和牛顿法之间，既能克服最速下降法慢的收敛性，又避免了牛顿法计算量大的缺陷.

3.4 共轭梯度法 — 二次函数极小化的共轭方向法

§3.4.1 二次函数极小化的共轭方向法

首先介绍共轭方向的概念及其性质.

定义 3.4.1 设G为n阶对称正定矩阵, $d^0, d^1, \ldots, d^{k-1}$ $(k \le n)$为n维向量组，如果

$$(d^i)^\top G d^j = 0, \quad \forall i, j \in \{0, 1, \ldots, k-1\},\ i \ne j,$$

那么，称向量组$d^0, d^1, \ldots, d^{k-1}$关于$G$共轭.

若$G = I$为单位矩阵，则由共轭向量组的定义, 可以看出：它们是正交的. 因此，向量组的共轭性是其正交性的一种推广形式. 下面的定理描述了共轭向量组的性质.

3.4共轭梯度法 — 二次函数极小化的共轭方向法

§3.4.1 二次函数极小化的共轭方向法

首先介绍共轭方向的概念及其性质.

定义 3.4.1 设G为n阶对称正定矩阵, $d^0, d^1, \ldots, d^{k-1}$ $(k \le n)$为n维向量组，如果

$$(d^i)^\top G d^j = 0, \quad \forall i, j \in \{0, 1, \ldots, k-1\},\ i \neq j,$$

那么，称向量组$d^0, d^1, \ldots, d^{k-1}$关于$G$共轭.

若$G = I$为单位矩阵，则由共轭向量组的定义, 可以看出：它们是正交的. 因此，向量组的共轭性是其正交性的一种推广形式. 下面的定理描述了共轭向量组的性质.

3.4共轭梯度法 — 二次函数极小化的共轭方向法

定理 3.4.1 设G为n阶对称正定矩阵. 非零n维向量组d^0, d^1, ..., d^{k-1} $(k \le n)$关于G共轭，则

- 向量组$d^0, d^1, \ldots, d^{k-1}$线性无关;
- 若$k = n$, 则向量组$d^0, d^1, \ldots, d^{n-1}$构成$\mathbb{R}^n$的一组基;
- 若$k = n$，且向量v与向量组$d^0, d^1, \ldots, d^{n-1}$关于$G$共轭，则$v = 0$.

定义 3.4.2 设n维向量组$d^0, d^1, \ldots, d^{k-1}$ $(k \le n)$线性无关，给定$x^0 \in \mathbb{R}^n$. 称集合

$$H_k := \left\{ x^0 + \sum_{i=0}^{k-1} \lambda_i d^i \;\middle|\; \lambda_i \in \mathbb{R}, i \in \{0, 1, \ldots, k-1\} \right\}$$

为由点x^0与$d^0, d^1, \ldots, d^{k-1}$生成的$k$维超平面.

3.4共轭梯度法 — 二次函数极小化的共轭方向法

引理 3.4.1 设f为连续可微的严格凸函数，$d^0, d^1, \ldots, d^{k-1}$ $(k \leq n)$为一组线性无关的n维向量，$x^0 \in \mathbb{R}^n$, 则

$$x^k = x^0 + \sum_{i=0}^{k-1} \hat{\lambda}_i d^i$$

是函数f在x^0与$d^0, d^1, \ldots, d^{k-1}$所生成$k$维超平面$H_k$上的唯一极小值点的充要条件是

$$g_k^\top d^i = 0, \quad \forall i \in \{0, 1, \ldots, k-1\}. \tag{3.4.2}$$

3.4共轭梯度法 — 二次函数极小化的共轭方向法

证明 定义函数

$$h(\lambda_0, \lambda_1, \ldots, \lambda_{k-1}) := f\left(x^0 + \sum_{i=0}^{k-1} \lambda_i d^i\right).$$

那么，

(i) $\nabla h(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1}) = (g_k^\top d^0, g_k^\top d^1, \ldots, g_k^\top d^{k-1})^\top$;

(ii) h是k维空间$\mathbb{R}^k$上的严格凸函数;

(iii) x^k是函数f在H_k上的极小值点当且仅当$(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1})^\top$是函数$h(\lambda_0, \lambda_1, \ldots, \lambda_{k-1})$在$\mathbb{R}^k$上的极小值点.

一方面，若x^k是f在H_k上的极小值点，则由定理3.1.1中的(i)可知，$\nabla h(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1}) = 0$. 这与上述结论(i)相结合可得: (3.4.2)成立.

另一方面，若(3.4.2)成立，则由上述结论(i)得$\nabla h(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1}) = 0$. 又由上述结论(ii)知：$h$是$k$维空间$\mathbb{R}^k$上的严格凸函数. 根据定理3.1.3, 可得$(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1})^\top$是$h$在$\mathbb{R}^k$上的极小值点. 这与上述结论(iii)相结合可得: x^k是f在H_k上的唯一极小值点. 引理得证. □

3.4共轭梯度法 — 二次函数极小化的共轭方向法

定理 3.4.2 *(子空间扩展定理)* 设G为n阶对称正定矩阵, 非零n维向量组$d^0, d^1, \ldots, d^{k-1}$ $(k \le n)$关于G共轭, 且二次函数f由*(3.4.1)*式定义. 由任意初始点x^0开始, 依次进行k次精确一维线搜索, 得

$$x^{i+1} = x^i + \lambda_i d^i, \quad \forall i \in \{0, 1, \ldots, k-1\}.$$

那么, x^k是由*(3.4.1)*式所定义二次函数f在H_k上的极小值点.

证明 由引理3.4.1可知：只需证明(3.4.2) 式成立. 因为

$$\begin{aligned} g_k^\top d^i &= g_k^\top d^i - g_{k-1}^\top d^i + g_{k-1}^\top d^i - g_{k-2}^\top d^i + \cdots + g_{i+2}^\top d^i - g_{i+1}^\top d^i + g_{i+1}^\top d^i \\ &= g_{i+1}^\top d^i + \textstyle\sum_{j=i+1}^{k-1} (g_{j+1} - g_j)^\top d^i, \quad \forall i \in \{0, 1, \ldots, k-1\}, \end{aligned}$$

且$g_{j+1} - g_j = G(x^{j+1} - x^j) = \lambda_j G d^j$, 所以,

$$g_k^\top d^i = g_{i+1}^\top d^i + \sum_{j=i+1}^{k-1} \lambda_j (d^j)^\top G d^i, \quad \forall i \in \{0, 1, \ldots, k-1\}. \qquad (3.4.3)$$

一方面，利用精确一维线搜索可得：$g_{i+1}^\top d^i = 0$；另一方面，由向量组的共轭性可得：对任意的$j \in \{i+1, \ldots, k-1\}$都有$(d^j)^\top G d^i = 0$.
因此, 由(3.4.3)式可得(3.4.2)式成立. 定理得证. □

3.4共轭梯度法 — 二次函数极小化的共轭方向法

算法 3.4.1 *(二次函数极小化的共轭方向法)*

设二次函数f由*(3.4.1)*式定义. 给定初始点$x^0 \in \mathbb{R}^n$及初始下降方向d^0. 计算g_0. 如果$g_0 = 0$, 算法终止. 否则, 置$k := 0$.

步*1* 由精确一维线搜索$f(x^k + \lambda_k d^k) = \min_{\lambda \geq 0} f(x^k + \lambda d^k)$, 计算步长$\lambda_k$.

步*2* 置$x^{k+1} := x^k + \lambda_k d^k$.

步*3* 计算g_{k+1}. 如果$g_{k+1} = 0$, 算法终止. 否则, 转步*4*.

步*4* 取共轭方向d^{k+1}, 使对任意的$i \in \{0, 1, \ldots, k\}$有$(d^{k+1})^\top G d^i = 0$.

步*5* 置$k := k + 1$, 转步*1*.

注. (i) 由定理3.4.2知：使用算法3.4.1求解由(3.4.1)式定义的严格凸二次函数极小化问题时，至多n步迭代可达其极小值点. 因此，算法3.4.1具有二次终止性. (ii) 算法3.4.1是一个概念性算法，实现它的关键在于如何选择一组合适的共轭方向. 下一小节介绍一个基于目标函数梯度构造共轭方向的方法，称之为共轭梯度法.

3.4 共轭梯度法 — 二次函数极小化的共轭梯度法

§3.4.2 二次函数极小化的共轭梯度法

假设二次函数f由(3.4.1)定义. 给定初始点x^0, 取初始下降方向:

$$d^0 = -g_0.$$

从x^0出发, 沿d^0进行精确一维线搜索得迭代点x^1, 完成第一次迭代. 不妨设$g_1 \neq 0$, 考虑如何选取共轭方向d^1. 定义一维子空间

$$L_1 := \left\{\beta_0 d^0 \mid \beta_0 \in \mathbb{R}\right\},$$

将负梯度方向分解成$-g_1 = u^1 + v^1$, 其中$u^1 \in L_1$且v^1与d^0关于G共轭(由线性代数的知识，这种分解是存在的). 令$d^1 = v^1$, 关键在于如何求解d^1. 显然，存在$\beta_0 \in \mathbb{R}$使得$-u^1 = \beta_0 d^0$. 因此，

$$d^1 = -g_1 + \beta_0 d^0.$$

用$(d^0)^\top G$左乘上式两边，并且利用共轭性, 可得$0 = (d^0)^\top G d^1 = -(d^0)^\top G g_1 + \beta_0 (d^0)^\top G d^0$. 于是有

$$\beta_0 = \frac{g_1^\top G d^0}{(d^0)^\top G d^0}, \quad 进而\ d^1 = -g_1 + \frac{g_1^\top G d^0}{(d^0)^\top G d^0} d^0.$$

3.4 共轭梯度法 — 二次函数极小化的共轭梯度法

类似递推，假设已沿k个共轭方向$d^0, d^1, \ldots, d^{k-1}$进行了$k$次精确一维线搜索，依次得到迭代点$x^1, x^2, \ldots, x^k$. 不失一般性，假设$g_k \neq 0$, 下面关键在于如何选取共轭方向$d^k$. 定义$k$ $(k > 1)$维子空间

$$L_k := \left\{ \sum_{i=0}^{k-1} \beta_i d^i \;\middle|\; \beta_i \in \mathbb{R},\ i \in \{0, 1, \ldots, k-1\} \right\},$$

将负梯度方向分解成

$$-g_k = u^k + v^k,$$

其中$u^k \in L_k$且v^k与L_k的基$d^0, d^1, \ldots, d^{k-1}$关于$G$共轭. 令$d^k = v^k$, 关键在于如何求解$d^k$. 类似于前面的推导, 可得

$$d^k = -g_k + \beta_{k-1} d^{k-1}, \quad \text{其中 } \beta_{k-1} = \frac{g_k^\top G d^{k-1}}{(d^{k-1})^\top G d^{k-1}},$$

并且可以得到下面的结论.

3.4 共轭梯度法 — 二次函数极小化的共轭梯度法

定理 3.4.3 设G为n阶对称正定矩阵且二次函数f由(3.4.1)式定义. 共轭梯度法的迭代方向为由以下公式产生的共轭方向$d^0, d^1, \ldots, d^{k-1}$ $(k \leq n)$:

$$\begin{cases} d^0 = -g_0, \\ d^k = -g_k + \beta_{k-1} d^{k-1}, \quad \text{其中 } \beta_{k-1} = \frac{g_k^\top G d^{k-1}}{(d^{k-1})^\top G d^{k-1}}, \ k > 0, \end{cases} \tag{3.4.4}$$

且迭代步长由精确一维线搜索得到. 则共轭梯度法$k-1$ $(k \leq n)$次迭代可求得二次函数f的极小点, 且对所有的$i \in \{0, 1, \ldots, k-1\}$, 有

- $(d^i)^\top G d^j = 0, g_i^\top g_j = 0, g_i^\top d^j = 0, \ \forall j \in \{0, 1, \ldots, i-1\};$
- $g_i^\top d^i = -g_i^\top g_i.$

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

§3.4.3　一般函数极小化的共轭梯度法

利用迭代方向的校正公式(3.4.4)式和定理3.4.3的结论，容易得到β_{k-1} $(k > 0)$的一些等价公式. 以Flecher-Reeves公式为例，简单推导如下：由于

$$Gd^{k-1} = G\left[\frac{1}{\lambda_{k-1}}(x^k - x^{k-1})\right] = \frac{1}{\lambda_{k-1}}(g_k - g_{k-1}),$$

所以，

$$\begin{aligned} g_k^\top Gd^{k-1} &= \frac{1}{\lambda_{k-1}} g_k^\top (g_k - g_{k-1}) = \frac{1}{\lambda_{k-1}} g_k^\top g_k, \\ (d^{k-1})^\top Gd^{k-1} &= \frac{1}{\lambda_{k-1}}(-g_{k-1} + \beta_{k-2} d^{k-2})^\top (g_k - g_{k-1}) = \frac{1}{\lambda_{k-1}} g_{k-1}^\top g_{k-1}, \end{aligned}$$

故有

$$\beta_{k-1} = \frac{g_k^\top g_k}{g_{k-1}^\top g_{k-1}}.$$

以上公式是Flecher和Reeves在1964年提出的，所以称之为Flecher-Reeves公式，简称为FR公式.

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

类似的推导，可得到以下三个常用的等价公式：

$$\beta_{k-1} = \frac{g_k^\top (g_k - g_{k-1})}{g_{k-1}^\top g_{k-1}}, \quad (Polak-Ribiere-Polyak\text{公式, 简称为}PRP\text{公式});$$

$$\beta_{k-1} = -\frac{g_k^\top g_k}{g_{k-1}^\top d^{k-1}}, \quad (Dixon\text{公式});$$

$$\beta_{k-1} = \frac{g_k^\top g_k}{(g_k - g_{k-1})^\top d^{k-1}}. \quad (Dai-Yuan\text{公式, 简称为}DY\text{公式}).$$

虽然FR公式、PRP公式、Dixon公式和DY公式是依据二次函数极小化来推导得出的, 但是公式中不显含二次函数的特有特征(如：不含矩阵G), 因此，这些公式常被应用于求解一般非线性函数的极小化问题(3.0.1). 另外，这些公式中主要的计算量只是计算一些向量的乘法，故计算量小，适合于求解大规模无约束优化问题.

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

以求解无约束优化(3.0.1)的FR共轭梯度法为例，具体算法如下：

算法 3.4.2 *(FR共轭梯度法)* 设函数f由问题*(3.0.1)*给出. 给定初始点$x^0 \in \mathbb{R}^n$. 置$k := 0$.

步*1* 计算g_k. 如果$g_k = 0$, 算法终止. 否则, 转步*2*.

步*2* 取

$$d^k = \begin{cases} -g_k, & \text{若} k = 0, \\ -g_k + \frac{g_k^\top g_k}{g_{k-1}^\top g_{k-1}} d^{k-1}, & \text{否则}. \end{cases}$$

步*3* 由线搜索计算步长λ_k.

步*4* 置$x^{k+1} := x^k + \lambda_k d^k$ 且$k := k + 1$, 转步*1*.

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

注. (i) 在算法3.4.2中，步长λ_k可由精确一维线搜索或非精确线搜索得到. 以下只考虑步长λ_k由精确一维线搜索得到的情况. (ii) 如果算法3.4.2中的FR公式被PRP公式、Dixon公式、DY公式来代替，则分别得到PRP共轭梯度法、Dixon共轭梯度法、DY共轭梯度法. (iii) 采用精确一维线搜索的FR共轭梯度法、PRP共轭梯度法、Dixon共轭梯度法、DY共轭梯度法均为下降算法，因为它们产生的搜索方向若非零，则满足

$$\begin{cases} g_k^\top d^k = -g_0^\top g_0 < 0, & \text{若} k = 0, \\ g_k^\top d^k = g_k^\top(-g_k + \beta_{k-1} d^{k-1}) = -g_k^\top g_k < 0, & \text{否则}. \end{cases}$$

以FR共轭梯度法为例，其算法的全局收敛性结论如下：

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

定理 3.4.4 设问题(3.0.1)中函数f在有界水平集$\mathscr{L}(f) := \{x \in \mathbb{R}^n \mid f(x) \le f(x^0)\}$上连续可微且有下界, 序列$\{x^k\}$是由采用精确一维线搜索的$FR$共轭梯度法产生的序列. 如果$\{x^k\}$有限, 那么$\{x^k\}$有限地终止于$f$的一个稳定点; 否则, $\{x^k\}$的每个聚点都是f的一个稳定点.

证明 不妨假设序列$\{x^k\}$是无限点列. 由于迭代方向d^k是下降方向, 结合函数f的函数值在水平集上有下界, 因而$\{f(x^k)\}$是单调下降且有下界的序列. 所以, $\lim_{k\to\infty} f(x^k)$存在. 又由于$\{x^k\} \subseteq \mathscr{L}$且$\mathscr{L}$有界, 所以$\{x^k\}$为有界点列, 因而存在收敛子列. 不失一般性, 不妨设$\lim_{k\to\infty} x^k = x^*$. 那么由$f$的连续性可知: $\lim_{k\to\infty} f(x^k) = f(\lim_{k\to\infty} x^k) = f(x^*)$.

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

下面采用反证法证明$g_* = 0$. 假设$g_* \neq 0$. 那么必存在一下降方向d^*, 使得对任意充分小的$\lambda_* > 0$, 有

$$f(x^* + \lambda_* d^*) < f(x^*). \tag{3.4.5}$$

而利用精确一维线搜索可知：$f(x^{k+1}) = f(x^k + \lambda_k d^k) \leq f(x^k + \lambda_* d^k)$, 两边取极限可得

$$f(x^*) \leq f(x^* + \lambda_* d^*).$$

这与(3.4.5)式相矛盾. 定理得证. □

进一步, 可以证明：在定理3.4.4的条件下，采用精确一维线搜索的FR共轭梯度法所产生的迭代序列线性收敛于目标函数f的极小值点.

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

例 3.4.1 采用精确一维线搜索的FR共轭梯度法求解优化问题

$$\min\ f(x) = \frac{1}{2}x_1^2 + x_2^2,$$

其中初始点取为$x^0 = (2,1)^\top$.

解 简单计算可得$g(x) = (x_1, 2x_2)^\top$. 由于$g_0 = (2,2)^\top \neq (0,0)^\top$, 因而取$d^0 = (-2,-2)^\top$.

从x^0出发沿d^0进行精确一维线搜索，即求解

$$\min\ f(x^0 + \lambda d^0) = 6\lambda^2 - 8\lambda + 3,$$

得步长$\lambda_0 = \frac{2}{3}$. 进而，$x^1 = x^0 + \lambda_0 d^0 = (\frac{2}{3}, -\frac{1}{3})^\top$, 且$g_1 = (\frac{2}{3}, -\frac{2}{3})^\top$. 由FR公式得$\beta_0 = g_1^\top g_1 / g_0^\top g_0 = \frac{1}{9}$. 从而，$d^1 = -g_1 + \beta_0 d^0 = (-\frac{8}{9}, \frac{4}{9})^\top$.

从x^1出发沿d^1进行精确一维线搜索，即求解

$$\min\ f(x^1 + \lambda d^1) = \frac{1}{162}(96\lambda^2 - 144\lambda + 54),$$

得步长$\lambda_1 = \frac{3}{4}$. 进而，$x^2 = x^1 + \lambda_1 d^1 = (0,0)^\top$, 且$g_2 = (0,0)^\top$.

因此，所求问题的最优解为$x^* = (0,0)^\top$, 最优值为$f^* = 0$. □

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

对于求解严格凸二次函数的极小化问题，FR共轭梯度法、PRP共轭梯度法、Dixon共轭梯度法、DY共轭梯度法都是等价的，但是，对于求解一般非线性函数的极小化问题，这些算法是不同的，并且目标函数的Hesse阵不是常数矩阵，迭代过程中所产生的迭代方向不再是共轭方向了. 众所周知：在最优解附近，目标函数与一个严格凸二次函数很接近. 因此，当迭代点进入目标函数逼近严格凸二次函数的区域后，如果能及时产生接近于共轭方向的搜索方向，那么算法就能迅速地收敛到最优解. 所以，当迭代点进入目标函数逼近严格凸二次函数的区域时，重新取负梯度方向作为搜索方向，那么后面的迭代将产生近似的共轭方向，从而提高算法的效率. 基于这一想法，共轭梯度法可修改如下：每迭代n次，就重新取负梯度方向作为搜索方向，得到的算法称为n步重新开始的共轭梯度法. 以采用精确一维线搜索的FR共轭梯度法为例，具体算法如下：

3.4 共轭梯度法 — 一般函数极小化的共轭梯度法

算法 3.4.3 *(n步重新开始的FR共轭梯度法)*

设函数f由问题(3.0.1)给出. 给定初始点$x^0 \in \mathbb{R}^n$. 置$k := 0$.

步1 计算g_k. 如果$g_k = 0$, 算法终止. 否则, 转步2.

步2 若k是n的倍数，则取$d^k = -g_k$, 否则，取

$$d^k = -g_k + \frac{g_k^\top g_k}{g_{k-1}^\top g_{k-1}} d^{k-1}.$$

步3 由精确一维线搜索$f(x^k + \lambda_k d^k) = \min_{\lambda \geq 0} f(x^k + \lambda d^k)$, 计算步长$\lambda_k$.

步4 置$x^{k+1} := x^k + \lambda_k d^k$ 且$k := k + 1$, 转步1.

与共轭梯度法一样，n步重新开始的共轭梯度法具有全局收敛性和线性收敛速度.

3.4 共轭梯度法 — 作业

3.12 (Gram-Schmidt共轭化) 假设$G \in \mathbb{R}^{n\times n}$为对称正定矩阵. $u^1, u^2, \ldots, u^n \in \mathbb{R}^n$ 线性无关. 假设$d^1, d^2, \ldots, d^n$由以下方式产生：

$$d^1 = u^1; \quad d^{k+1} = u^{k+1} - \sum_{i=1}^{k} \frac{(u^{k+1})^\top G d^i}{(d^i)^\top G d^i} d^i, \ \forall k \in \{1, 2, \ldots, n-1\}.$$

试证明: $d^1, d^2, \ldots, d^n$关于G共轭.

3.13 考虑函数$f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$, 设初始点取为$x^0 = (1,1)^\top$, $d^0 = (-1,0)^\top$. 沿方向d^0进行精确一维线搜索可得λ_0, 令$x^1 = x^0 + \lambda_0 d^0$, 用FR公式求得方向$d^1$. 试证明: 方向$d^0$和$d^1$不是关于单位矩阵$I$共轭的.

3.14 用精确一维线搜索的FR共轭梯度法(及PRP共轭梯度法)求解下列无约束优化问题.

(1) $\min \ f(x_1, x_2) = x_1^2 - x_1x_2 + x_2^2 + 2x_1 - 4$, 其中初始点取为$x^0 = (2,2)^\top$.

(2) $\min \ f(x_1, x_2) = x_1^2 + 2x_2^2$, 其中初始点取为$x^0 = (1,1)^\top$.