MATH 3011 运筹学 **秋季 2023**

Lecture 9: 无约束优化 线搜索方法

Lecturer: 陈士祥 *Scribes:* 陈士祥

1 问题形式

无约束最优化问题

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{9.1}$$

其目标函数 f 是定义在 $\mathbb{R}^n$ 上的实值函数，决策变量 x 的可取值之集合是全空间 $\mathbb{R}^n$.

2 梯度类算法

梯度向量 $\nabla f(x)$ 是函数 f 在点 x 处增加最快的方向，故它成为最优化时的重要工具。实际上针对无约束最优化问题，大部分求解算法属于下面的梯度方法类。

梯度类算法：

(0) 初始化：选取适当的初始点 $x^0 \in \mathbb{R}^n$, 令 $k := 0$.

(1) 计算搜索方向：利用适当的正定对称阵 H_k 计算搜索方向向量 $d^k := -H_k \nabla f(x^k)$.（如果 $\nabla f(x^k) = 0$, 则结束计算）

(2) 确定**步长**因子：解一维最优化问题 $\min_{\alpha \geq 0} f(x^k + \alpha d^k)$, 求出**步长**$\alpha = \alpha_k$, 令 $x^{k+1} = x^k + \alpha_k d^k$, $k := k + 1$, 回到第 (1) 步。

注：在机器学习领域，步长通常被称为学习率（learning rate）。

例 9.1 若 $f(x)$ 二阶可导，我们有

$$f(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + O(\|x - x^k\|^2). \tag{9.2}$$

取负梯度方向

$$d^k = -\nabla f(x^k),$$

则当 α_k 足够小时，总能使

$$f(x^k + \alpha_k d^k) < f(x^k).$$

例 9.2 若 $f(x)$ 三阶可导，我们有

$$f(x) = \begin{aligned} & f(x^k) + \nabla f(x^k)^T (x - x^k) \\ & + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k) + O(\|x - x^k\|^3) \end{aligned} \tag{9.3}$$

假设函数 f 在 x^k 点处的 *Hesse* 矩阵 $\nabla^2 f(x^k)$ 正定，取搜索方向

$$d^k = -G_k^{-1} \nabla f(x^k),$$

其中 $G_k = \nabla^2 f(x^k)$。这样的取法叫做牛顿方向，我们后面会进一步讨论。若 α_k 充分小，那么也可以得到

$$f(x^k + \alpha_k d^k) < f(x^k).$$

3 确定步长因子：一维搜索

在迭代格式中，沿着下降方向 d^k, 通过解一维最优化问题

$$\min_{\alpha \geq 0} \varphi(\alpha) = f(x^k + \alpha d^k) \tag{9.4}$$

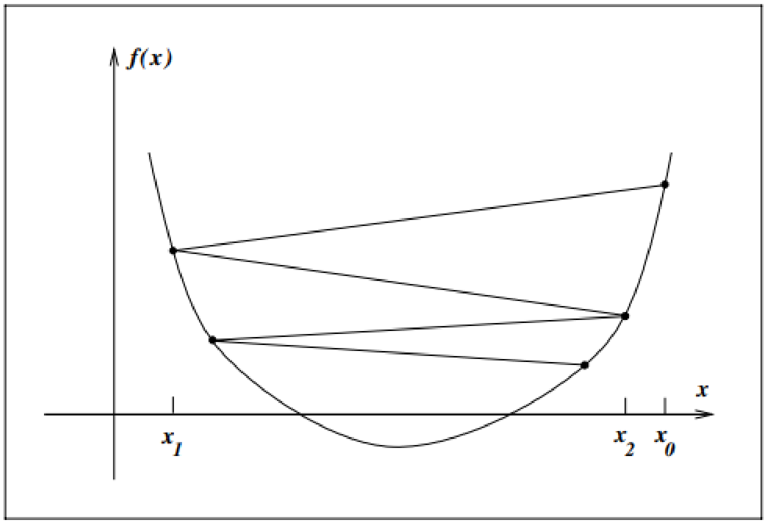
确定步长因子的方法称为**一维搜索** (Line Search).

若以问题(9.4)的最优解为步长，此时称为**精确一维搜索** (Exact Line Search).

经常用到的精确一维搜索有黄金分割法和插值迭代法。即使说是精确一维搜索，通过有限次计算求出问题(9.4)的严密解一般也是不可能的，实际上在得到有足够精度的近似解时，就采用它作为步长。

在实际计算中，往往不是求解一维最优化问题(9.4), 而是找出满足某些适当条件的粗略近似解作为步长，此时称为**非精确一维搜索** (Inexact Line Search). 与精确一维搜索相比，在很多情况下采用非精确一维搜索可以提高整体计算效率。

3.1 线索搜的重要性



上图中，由于步长 α 选择较大，迭代产生了左右震荡。反之，若是步长太小，那么算法收敛速度非常缓慢。线搜索的目标就是，使得沿着下降方向 d，每次的函数值满足充分下降（sufficient decrease）条件。

3.2 回溯线搜索法

最简单、常见的线搜索条件为回溯线搜索法 (Backtracking linesearch):

1. 选取 $\gamma \in (0,1)$ 和 $c \in (0,1)$
2. 选择最小的整数 $t \geq 0$, 使得
$$f(x^k + \gamma^t d^k) \leq f(x^k) + c\gamma^t \nabla f(x^k)^T d^k \tag{9.5}$$
3. 令 $\alpha_k = \gamma^t$, 更新 $x^{k+1} = x^k + \alpha_k d^k$

(9.5)被称为 Armijo-Goldstein 不等式。故 backtracking 也被称为 Armijo-backtracking linesearch. γ 通常选取 0.9 或者 0.5. c 通常选取 $10^{-2}, 10^{-3}$ 等较小的数。若 c 较大，则要求每次函数下降量足够大，但是需要更多的搜索步数。反之，则下降量较小，造成算法总体下降速度过慢。实际问题中，需要调试参数 γ, c 的选择对算法进行加速。
正整数 t 的存在性：

$$\lim_{\alpha \downarrow 0} \frac{f(x^k + \alpha d^k) - f(x^k)}{\alpha} = f'(x^k; d^k) < cf'(x^k; d^k) < 0.$$

故，存在 $\bar{\alpha} > 0$，使得

$$\frac{f(x^k + \alpha d^k) - f(x^k)}{\alpha} \leq cf'(x^k; d^k), \quad \forall \alpha \in (0, \bar{\alpha}) \tag{9.6}$$

回溯线搜索法和下面的 Armijo-Goldstein 线搜索非常相似。令

$$\varphi(\alpha) = f(x + \alpha d).$$

我们有 $\varphi'(\alpha) = \nabla f(x + \alpha d)^T d$.

Armijo-Goldstein 条件为:

$$\varphi(\alpha) \leq \varphi(0) + \rho\alpha\varphi^{'}(0) \tag{9.7}$$

$$\varphi(\alpha) \geq \varphi(0) + (1-\rho)\alpha\varphi^{'}(0) \tag{9.8}$$

其中 $\rho \in (0, 1/2)$ 是一个固定参数。(9.7)要求函数值满足充分下降条件，其对应于条件(9.5)。(9.8)要求函数值下降量不是很小，对应于我们在回溯法中取最小的 t 即最大的步长 $\alpha_k = \gamma^t$.

Armijo-Goldstein 条件(9.7)和(9.8)虽然可以使得函数值下降，但是如图 9.2, 其排除了局部最小值点。

3.3 Wolfe-Powell 条件

Wolfe(1968)-Powell(1976) 条件是另外一种常见的非精确线搜索方法。

Wolfe-Powell 条件如下:

$$\varphi(\alpha) \leq \varphi(0) + c_1\alpha\varphi^{'}(0) \tag{9.9}$$

$$\varphi^{'}(\alpha) \geq c_2\varphi^{'}(0) \tag{9.10}$$

其中 $0 < c_1 < c_2 < 1$ 是固定参数。**通常来说 c_1 比较小，例如 $c_1 = 10^{-3}$. $c_2 = 0.9$.**
考虑问题

$$\min_{\alpha \in \mathbb{R}} \varphi(\alpha),$$

其最优条件为 $\varphi'(\alpha) = 0$. (9.10) 使得 $\varphi'(\alpha)$ 更接近 0，也被称为弱 Wolfe-Powell 条件。在很多实际算法中，式(9.10)常被强化的双边条件(9.11)所取代（也被称为强 Wolfe-Powell 条件）

$$|\varphi^{'}(\alpha)| \leq c_2|\varphi^{'}(0)|, 0 < c_1 < c_2 < 1. \tag{9.11}$$

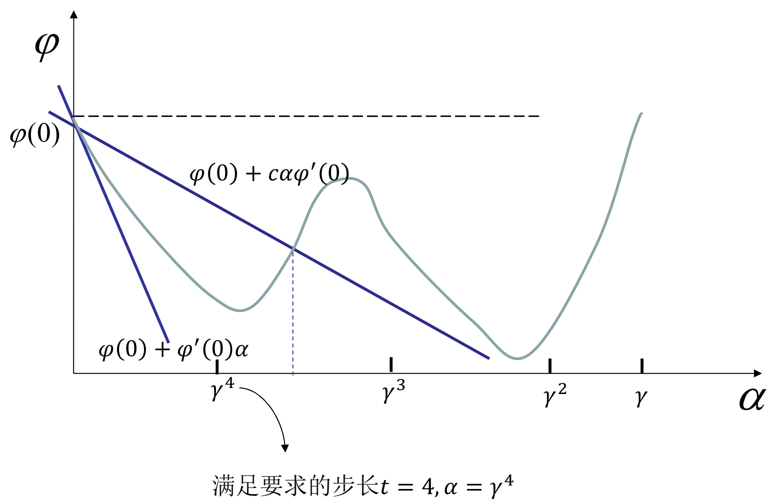


图 9.1: 满足回溯条件(9.5)的步长图例

此条件排除了 $\varphi^{'}(\alpha)$ 为非常大的正数情况。

Wolfe-Powell 条件存在性：若问题(9.1)中 f 连续可微。令 d^k 是下降方向，并且假设 f 沿着射线方向 $\{x^k+\alpha d^k \mid \alpha>0\}$ 有下界。那么一定存在 $0<c_1<c_2<1$ 使得 (9.9)、(9.10) 或(9.11)成立。

证明： 因为 f 沿着射线方向 $\{x^k+\alpha d^k \mid \alpha>0\}$ 有下界，$l(\alpha):=f(x^k)+\alpha c_1\nabla f(x^k)^Td^k$ 单调减小至 $-\infty$, 可知 $\varphi(\alpha)$ 与 $l(\alpha)$ 至少有一个交点。设 $\bar{\alpha}$ 是最小的交点。有下述不等式成立

$$f(x^k+\alpha_1 d^k)\le f(x^k)+c_1\alpha_1\nabla f(x^k)^Td^k,\quad \forall\alpha_1\in(0,\bar{\alpha}).$$

由中值定理可知，存在 $\alpha_2\in(0,\bar{\alpha})$ 使得 $f(x^k+\bar{\alpha}d^k)-f(x^k)=\bar{\alpha}\nabla f(x^k+\alpha_2 d^k)^Td^k$ 因为 $l(\bar{\alpha})=\varphi(\bar{\alpha})$, 即 $f(x^k+\bar{\alpha}d^k)=f(x^k)+c_1\bar{\alpha}\nabla f(x^k)^Td^k$, 我们有

$$\nabla f(x^k+\alpha_2 d^k)^Td^k=c_1\nabla f(x^k)^Td^k>c_2\nabla f(x^k)^Td^k, 1>c_2>c_1>0.$$

因此 $0<c_1<c_2<1, \alpha_1\in(0,\bar{\alpha}), \alpha_2\in(0,\bar{\alpha})$ 满足(9.9)、(9.10).
注意到 $\nabla f(x^k+\alpha_2 d^k)^Td^k<0$, 因此(9.11)也成立。

实际中，为了确定满足 Wolfe-Powell 条件的步长，一种方法是采用插值法。其步骤如下：

(1) 给定初始一维搜索区间 $[0,\alpha_0]$, 以及 $c_1\in(0,1/2)$, $c_2\in(c_1,1)$. 记 $a_1=0, a_2=\alpha_0$.

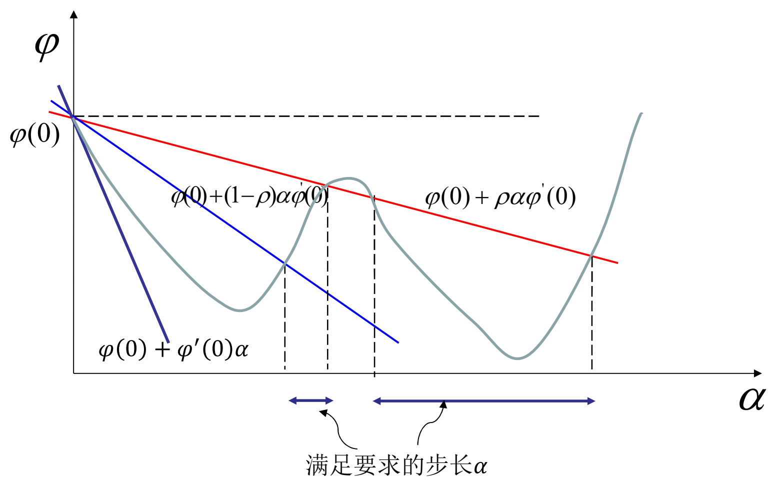


图 9.2: 满足 Armijo-Goldstein 条件(9.7)和(9.8)的步长图例

计算 $\varphi(0) = f(x^k)$, $\varphi^{'}(0) = \nabla f(x^k)^T d^k$. 并令 $a_1 = 0, a_2 = \alpha_0, \varphi_1 = \varphi(0), \varphi_1^{'} = \varphi^{'}(0)$.

选取适当的 $\alpha \in (a_1, a_2)$.

(2) 计算 $\varphi = \varphi(\alpha) = f(x^k + \alpha d^k)$. 若 $\varphi(\alpha) \le \varphi(0) + c_1 \alpha \varphi^{'}(0)$, 则转到第 (3) 步。否则, 由 $\varphi_1, \varphi_1^{'}, \varphi$ 构造两点二次插值 $p_1(\alpha) = A_1\alpha^2 + B_1\alpha + C_1$, 逼近区间 $[a_1, \alpha]$ 上的 $\varphi(\alpha)$. 使得 $p_1(a_1) = \varpi_1, p_1'(a_1) = \varphi_1'$, $p_1(\alpha) = \varphi$. 并得 $p_1(a_1)$ 极小点

$$\hat{\alpha} = a_1 + \frac{1}{2}\frac{(a_1 - \alpha)^2 \varphi_1^{'}}{(\varphi_1 - \varphi) - (a_1 - \alpha)\varphi_1^{'}}.$$

于是置 $a_2 = \alpha, \alpha = \hat{\alpha}$, 重复第 (2) 步。

(3) 计算 $\varphi^{'} = \varphi^{'}(\alpha) = \nabla f(x^k + \alpha d^k)^T d^k$. 若 $\varphi^{'}(\alpha) \ge c_2 \varphi^{'}(0)$, 则输出 $\alpha_k = \alpha$, 并停止搜索。否则，由 $\varphi, \varphi^{'}, \varphi_1^{'}$ 构造两点二次插值多项式

$$p_2(\alpha) = A_2\alpha^2 + B_2\alpha + C_2,$$

使得 $p_2'(a_1) = \varphi_1', p_2'(\alpha) = \varphi'$, $p_2(\alpha) = \varphi$. 并得其极小点

$$\hat{\alpha} = \alpha - \frac{(a_1 - \alpha)\varphi^{'}}{\varphi_1^{'} - \varphi^{'}}.$$

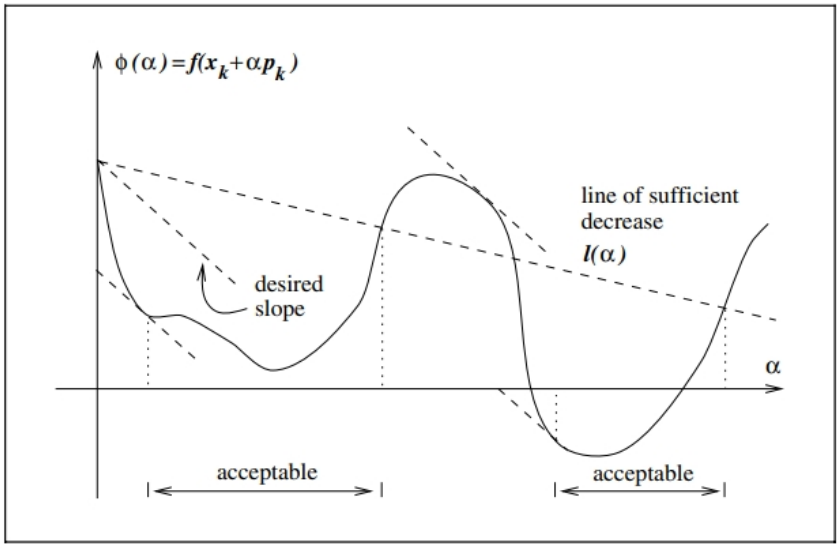


图 9.3: Wolfe-Powell 条件示例。**图片来源：Numerical optimization. By Jorge Nocedal and Stephen J. Wright.**

于是置 $a_1=\alpha, \alpha=\hat{\alpha}, \varphi_1=\varphi, \varphi_1^{'}=\varphi^{'}$, 返回第 (3) 步。

3.4 线搜索的全局收敛性

从任意初始点出发，如果某迭代算法产生的点列的极限（聚点），在适当假定下可保证恒为问题的最优解（或者稳定点），则称该迭代法具有全局收敛性 (Global Convergence).

为了证明迭代法的下降性，我们应尽量避免搜索方向与负梯度方向几乎正交的情形，即要求 d^k 偏离 $g^k=\nabla f(x^k)$ 的正交方向远一些。否则，${g^k}^T d^k$ 接近于零，d^k 几乎不是下降方向。

为此，我们假设 d^k 与 $-g^k$ 的夹角 θ_k 满足

$$\theta_k \le \frac{\pi}{2}-\mu,\ \forall k \tag{9.12}$$

其中 $\mu>0$（与 k 无关）。
显然 $\theta_k\in[0,\pi/2)$, 其定义为

$$\cos\theta_k=\frac{-{g^k}^T d^k}{\|g^k\|\|d^k\|}=\frac{-{g^k}^T s^k}{\|g^k\|\|s^k\|} \tag{9.13}$$

这里 $s^k=\alpha_k d^k=x^{k+1}-x^k$.

下面给出各种步长准则下的下降算法的全局收敛性结论。

Theorem 9.1 假设 $f(x)$ 有下界，即 $f(x)>-\infty, \forall x\in\mathbb{R}^n$. 设 $f(x)$ 在包含水平集 $L(x^0)=\{x\mid$

$f(x) \le f(x^0)\} \subset \mathcal{N}$ 的开集 $\mathcal{N}$ 上连续可微。同时，梯度 $\nabla f(x)$ 在 $\mathcal{N}$ 上是李氏（*Lipschitz continuous*）连续的，即存在 $L > 0$，使得

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathcal{N}.$$

下降算法的搜索方向 d^k 与 $-\nabla f(x^k)$ 之间的夹角 θ_k 满足式(9.12)，其中步长 α_k 由 *Wolfe-Powell* (9.9),(9.10)确定。那么，$\nabla f(x^k) \to 0$ *as* $k \to \infty$.

Proof: 全局收敛性证明： 为了记号简洁，我们记所有的 k, $g_k = \nabla f(x^k)$，$f_k = f(x^k)$. 由梯度李氏连续性可知：

$$\|g_{k+1} - g_k\| \le L\|x_{k+1} - x_k\| = L\alpha_k\|d^k\|$$

由(9.10)可知，

$$(g_{k+1} - g_k)^T d^k \ge (c_2 - 1) g_k^T d^k.$$

结合上述两个不等式，我们有

$$\alpha_k \ge \frac{c_2 - 1}{L}\frac{g_k^T d^k}{\|d^k\|^2}. \qquad \text{(步长有下界)}$$

带入(9.9), 我们有

$$f_{k+1} \le f_k - c_1\frac{1 - c_2}{L}\frac{(g_k^T d^k)^2}{\|d^k\|^2}.$$

根据(9.13), 我们有

$$f_{k+1} \le f_k - c_1\frac{1 - c_2}{L}\cos^2\theta_k\|g_k\|^2.$$

将上述不等式，对 $k = 0, ..., T$ 相加，可得

$$f_{T+1} \le f_0 - c_1\frac{1 - c_2}{L}\sum_{k=0}^{T}\cos^2\theta_k\|g_k\|^2.$$

因为 f 有下界，故

$$\sum_{k=0}^{T}\cos^2\theta_k\|g_k\|^2 < +\infty.$$

上式对任意 T 成立，故

$$\sum_{k=0}^{\infty}\cos^2\theta_k\|g_k\|^2 < +\infty.$$

又因为(9.12), 存在 $\delta > 0$, 使得

$$\cos\theta_k \ge \delta.$$

所以

$$\lim_{k\to\infty}\|\nabla f(x^k)\| = 0.$$

■